Mobility Prediction Based Proactive Dynamic Network Orchestration for Load Balancing With QoS Constraint (OPERA)

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

***Abstract*—Load imbalance among small and macro cells is a major challenge that undermines the gains of emerging ultradense heterogeneous networks (HetNets). Existing load balancing (LB) schemes have one common caveat which is operating in reactive mode i.e., cell parameters are tweaked reactively in accordance with the dynamics of cell loads. The inherent reactiveness of these LB schemes hinder in achieving promising quality of experience (QoE) gains from 5G and beyond. To cope with this issue, in this paper we propose a novel proactive load balancing framework "OPERA" empowered by mobility prediction paradigm for future ultra dense networks (UDNs). The pro-activeness of OPERA stems from its novel capability that instead of passively waiting for congestion indicators to be observed and then reacting to them, OPERA predicts future cell loads and then proactively optimizes key antenna parameters and cell individual offsets (CIOs) to preempt congestion before it happens. OPERA also incorporates capacity and coverage constraints and load aware association strategy for ensuring conflict free operation of LB and coverage and capacity optimization (CCO) self-organizing network (SON) functions. Simulation results show that compared to real network deployments settings and published state-of-the-art reactive schemes, OPERA can yield significant gain in terms of fairness in load distribution and percentage of satisfied users. Superior performance of OPERA on several fronts compared to current schemes stems from its following features: 1) It preempts congestion instead of reacting to it; 2) it actuates more parameters than any current LB schemes thereby increasing system level capacity instead of just shifting it among cells; 3) while performing LB OPERA simultaneously maximizes residual capacity while incorporating throughput and coverage constraints; 4) it incorporates a load aware association strategy for ensuring conflict free operation of LB and CCO SON functions; 5) the ahead of time estimation of cell loads allows ample time for heuristics search algorithms to find LB solutions with high gain.**

***Index Terms*—5G, load balancing, mobility prediction, proactive SON, small cells, CIOs.**

I. Introduction

THE race to 5G is on with massive impromptu densification by small cells, orchestrated by Self Organizing Networks (SON), being perceived as a cost-effective solution to the impending mobile capacity crunch. Although poor indoor coverage coupled with explosive cellular data growth—that were expected to generate the momentous demand—are still relevant, to date, hefty small cell deployments are not there as expected. One of the key challenge therein is the load imbalance issue that stems from low transmission power and height of small cells and the conventional max–received signal strength based user association [1]. Even with a targeted deployment where the small cells are placed in high-traffic zones, most users still end up receiving the strongest downlink signal from the tower-mounted macrocell. As a result, macrocells remain overloaded and small cells remain underloaded as they fail to achieve user association proportional to available bandwidth. This load imbalance also effects the user perceived rate which is the product of instantaneous rate and the radio resources assigned to users. In highly loaded macrocells, few resources are assigned to users and hence user perceived Quality of Experience (QoE) drastically degrades. Consequently, load imbalance has been a time persistent challenge that has thwarted the wide scale deployment and benefits of small cells.

A. Relevant Work

Load imbalance can be mitigated by shifting the traffic from high loaded cells to less loaded neighbors as far as interference and coverage situation allows. To exploit this approach, recently load balancing (LB) has gained attention as a prominent SON function by 3GPP [2] and has been focus of research for many works like in [3]–[12]. However, the existing LB approaches proposed in [3]–[12] have following four common limitations that hinders them achieving 5G ambitious QoE requirements:

1) Reactive Design: The state-of-the-art LB SON algorithms are designed to optimize the hard or soft network parameters such as tilts (hard parameter), transmission powers (hard parameter), cell individual offsets (example of soft parameter) based on current network conditions. Such solutions (e.g. [9]) offer improvement over fixed parameters settings in real networks that achieves LB at the cost of QoE. However, in the fast dynamical cellular environment, where the scheduling is done in order of milliseconds, by the time the realistic non-convex NP-hard LB algorithms come up with optimum network configuration, the scenario might have already changed, and optimized parameter values become outdated thus undermining gain achieved from

Manuscript received February 1, 2019; revised November 21, 2019; accepted January 8, 2020. Date of publication January 15, 2020; date of current version March 12, 2020. This work was supported in part by the National Science Foundation under Grants 1619346, 1718956, and 1730650 and in part by the Qatar National Research Fund under Grant NPRP12-S 0311-190302. The review of this article was coordinated by Dr. Y. Ji. *(Corresponding author: Hasan Farooq.)*

The authors are with the University of Oklahoma, Tulsa, OK 74135 USA (e-mail: hasan.farooq@ou.edu; ahmad.asghar@ou.edu; ali.imran@ou.edu).

Digital Object Identifier 10.1109/TVT.2020.2966725

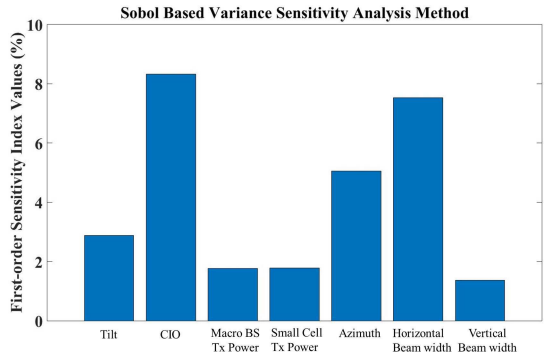


Fig. 1. Sobol method-based first-order sensitivity index values for tilts, CIOs, macro BS transmission power, small BS transmission power, azimuth, horizontal and vertical beam widths.

LB. This problem is bound to escalate further in 5G as delay intrinsic to a reactive LB solution means the stringent latency and QoE requirements cannot be met. Furthermore, in 5G and beyond the support for new mobility centric services such as intelligent transport systems and self-deriving cars, and smaller cell sizes mean even faster dynamics.

2) Limited Set of Optimization Parameters: Existing LB solutions use one or more of the only following three parameters as actuators for achieving LB: antenna tilts [9], [12], downlink transmission power [4], [12] and cell individual offsets (CIOs) [3], [4], [6]–[8], [12]. However with the evolution of smart antennas technology, new set of optimization parameters have emerged that are yet to be exploited. These includes beam widths (radiation pattern) that can be adapted on the fly by optimizing the phases of complex weight vectors—thanks to multi-array antennas technology. Similarly azimuth orientation of the antennas can be changed remotely and frequently to effectively change cell footprint, in addition to or in conjunction with the antenna tilts. In Fig. 1 we have quantified the ability of possible parameters to affect network performance (QoE) using Sobol based variance sensitivity analysis method [13]. It is observed that the CIOs, horizontal beam width and azimuth have the largest impact on the network performance. This observation calls for a shift from the legacy paradigm of mostly optimizing tilts and/or Tx power to maximize system performance and keeping other control knobs untouched.

3) SON Conflict Prone Design: Another issue with current LB SON solutions is the intrinsic conflicts or unexpected performance that results from concurrent operation of multiple SON use cases. Stand-alone LB solutions are bound to negatively conflict with Coverage and Capacity Optimization (CCO) SON function due to the overlap among their optimization parameters. For example, when CCO may try to improve coverage of cell by increasing its Tx power, this can force large number of users to associate to that cell thereby conflicting with LB SON objective. The interplay between CCO and LB becomes complicated considering that both CCO and LB resort to optimization of same parameters i.e., tilts, Tx power and CIO. For detailed analysis of this conflict, reader is referred to [14]. CIO, which unlike antenna parameters and Tx power is a soft parameter, has been recently introduced by 3GPP for LB and traffic steering in HetNets. However, adjustment of CIO by the LB algorithm may also cause conflict with CCO objectives as a user offloaded due to increased CIO may face higher interference (assuming intra-frequency offloading), and lower received power from the destination cell, compared to the origin cell. This may result into lower SINR and ultimately lower throughputs thereby conflicting CCO objective. Such conflict prone LB design can often end up increasing the complexity of network operation for RAN engineers and compromising the QoE instead of improving it [14].

4) Impractical Assumptions: There exist line of works such as [10], [11] that are more theoretical in nature aimed for LB or more precisely optimal cell association in HetNets while considering CCO in form of constraints and vice versa. While these works provide valuable theoretical insights often into the asymptotic behavior of the system, for tractability the analytical models used in these theoretical studies often build on overly-simplified and unrealistic assumptions such as uniformly distributed user equipments (UEs), spatially independent distribution of base stations, omnidirectional single-antenna transmission and reception, fixed transmit powers, same CIO for all cells in one tier, full load scenarios etc. These assumptions help to make the analysis tractable and make optimization problem convex, but render the end result less useful for practical implementation. Contrary to dense HetNet as the main motivation for LB SON function, some works on LB exist like [8], [9] wherein the solution is proposed and simulated mainly for macrocell scenarios, i.e., large cell individual offsets and Tx power disparities between small cells and macro cells are not considered. These approaches may work for current macro cell dominated network deployment but may not be applicable to dense HetNet envisioned for 5G.

In light of the aforementioned limitations, we propose OPERA framework (Fig. 2) that leverages a novel approach of transforming user mobility from being challenge to an advantage. OPERA exploits the knowledge gained from mobility/hand-off patterns to proactively and preemptively prevent load imbalance in emerging dense HetNets. It mines user mobility behavior from easily available logs such as hand over (HO) traces to anticipate future load conditions. This knowledge is then leveraged by a novel LB optimization problem to prevent load imbalance in a proactive way. The paper has following contributions:

1) In the proposed novel OPERA framework, spatio-temporal mobility prediction based on semi-Markov model complemented with vector theory based geomarker concept is leveraged to predict future loads of the cells. Transparency of the mobility model to cell types is an added advantage to make the model's accuracy robust and stable in presence of cell type diversity in HetNets.
2) Based on predicted utilization of cells, proactive optimization is performed to maximize the logarithmic sum of free resources in all the cells. The proposed proactive LB scheme leverages a judicious combination of hard parameters i.e., (tilts, azimuths, beam widths, Tx power) and soft parameters i.e., CIOs as optimization variables. Furthermore, a novel load aware association strategy for balancing load among cells is also proposed and used. This formulation is solved by the novel hybrid combination of genetic algorithms and patterns search and the proactivity of OPERA enables them to converge to high yielding LB

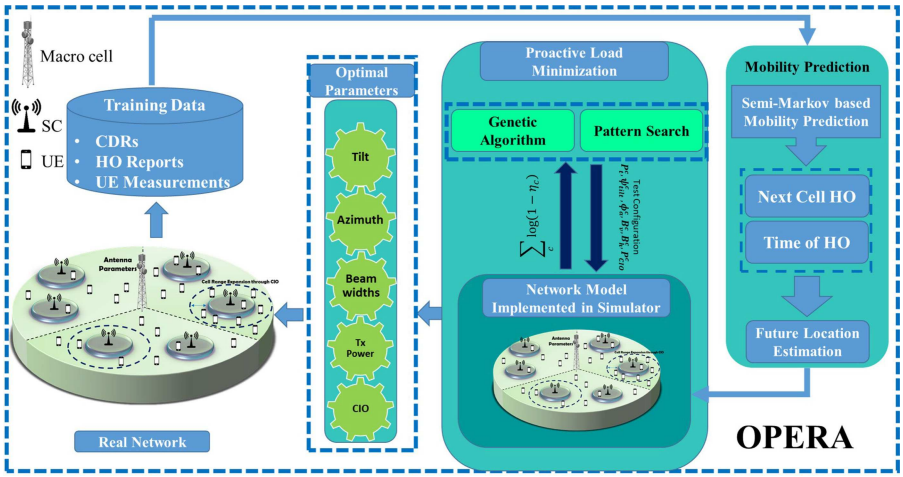


Fig. 2. OPERA framework.

solutions without affecting latency requirements in 5G and beyond.

3) Rigorous simulations are performed to benchmark performance of the proposed solution against several schemes including a real LTE network deployment and a reactive scheme from published study. OPERA significantly reduces number of un-satisfied users in the network and also achieves maximum residual capacity. Residual capacity i.e., resources available in cell to be allocated to a user, is metric that can be used to quantify the ability of emerging HetNets to cope with acute fluctuation in cell loads owing to mobility and decreased cell size. Load-aware association strategy provides robustness to OPERA against load estimation inaccuracies that is further verified by comparing it to near-optimal performance bound when future cell loads prediction accuracy is 100%.

II. OPERA Framework

This section describes the mobility prediction model considered and the load minimization optimization problem leveraged by the OPERA framework to minimize all network cells loads.

Network Assumptions: In this work, we only focus on the downlink of cellular systems. Macro cells are assumed to be equipped with smart directional antennas while UEs and small cells have omnidirectional antennas. Same spectrum is shared between the macro and small cells (co-channel interference). Each UE is assumed to be active all the time running a constant bit rate (CBR) service. OPERA builds on a centralized SON (C-SON) architecture to perform network wide optimization. The C-SON style implementation has access to all user reported measurements like time stamped HO reports, minimization of drive tests (MDT) measurments, call data records (CDRs) etc.

A. Cell-Level Prediction

Some phenomenal large scale studies like [15] have proven as high as 93% average predictability embedded in regular daily routines of humans. This in turn, provides a rational for predicting a person's movement given past trajectories. Backed by this fact, the basic building block of OPERA framework is a mobility prediction model that when given person's mobility history in terms of tuple of locations (cells) visited with corresponding pause times (cell sojourn times), it predict this person's next location, as well as his/her sojourn time. The mobility prediction model should meet two criteria: 1) It can be obtained with low complexity low latency online practically implementable algorithms; 2) It can predict future cell as well as HO time. In this paper, we leverage semi-Markov stochastic process for modeling and predicting human mobility because of 1) proven potential and suitability of Markov theory to model similar prediction problems [16]–[18], 2) their ability of modeling any arbitrary distributed sojourn time instead of being locked to impractical assumption of memroy-less exponentially distributed mobility that has been shown to be not true in general [19]. Some works do exist that have quantified the prediction accuracy of semi-Markov based predictor for mobility prediction [20]–[22]. However, like majority of recent studies on mobility prediction in cellular networks [23]–[27], the aim of studies in [20]–[22] is also limited to investigating the prediction accuracy of the leveraged mobility prediction scheme only. None of these studies further refine and exploit this information for optimization of the cellular network such as load balancing, as proposed in this paper.

We model user mobility as a semi-Markov renewal process $\{(C_n, J_n) : n \geq 0\}$ where C_n is the state (cell) at nth transition, J_n is the time of nth transition and a total of z cells with discrete state space $\mathcal{C} = \{\text{Cell}_1, \text{Cell}_2, \text{Cell}_3 \ldots, \text{Cell}_z\}$. Each state in the semi-Markov process represents a cell, wherein HO from a cell to another is modelled as a state transition. Random variable $J_n^{(u)}$ represent time instant of the transition $C_n^{(u)}$ to $C_{n+1}^{(u)}$ while random variable $J_{n+1}^{(u)} - J_n^{(u)}$ describes the cell sojourn time, or state holding time. The distribution of these random variables is not restricted to memoryless exponential distributions. It is assumed that the transition probabilities do not change when the model is being built. The associated time-homogeneous semi-Markov kernel for user u that is probability of u for transiting to

jth cell after staying in ith cell for no more than t time is defined as:

$$Q_{i,j}^{(u)}(t) = Pr\Big(C_{n+1}^{(u)} = j, J_{n+1}^{(u)} - J_n^{(u)} \leq t \,| \, C_0^{(u)}, \ldots, C_n^{(u)}; J_0^{(u)}, \ldots, J_n^{(u)}\Big) \quad (1)$$

$$= Pr\Big(C_{n+1}^{(u)} = j, J_{n+1}^{(u)} - J_n^{(u)} \leq t | C_n^{(u)} = i\Big) \quad (2)$$

Assuming that the cell sojourn time random variables are independent from the embedded state transition process $(C_{i,j})$, we get

$$Q_{i,j}^{(u)}(t) = Pr\Big(C_{n+1}^{(u)} = j | C_n^{(u)} = i\Big) \cdot Pr\Big(J_{n+1}^{(u)} - J_n^{(u)} \leq t | C_{n+1}^{(u)} = j, C_n^{(u)} = i\Big) \quad (3)$$

$$= h_{i,j}^{(u)} F_{i,j}^{(u)}(t) \quad (4)$$

where

$$h_{i,j}^{(u)} = \lim_{t\to\infty} Q_{i,j}^{(u)}(t) \quad (5)$$

$$= Pr\Big(C_{n+1}^{(u)} = j | C_n^{(u)} = i\Big) \quad (6)$$

and

$$F_{i,j}^{(u)}(t) = Pr\Big(J_{n+1}^{(u)} - J_n^{(u)} \leq t | C_{n+1}^{(u)} = j, C_n^{(u)} = i\Big) \quad (7)$$

Here $h_{i,j}^{(u)} \in \mathbf{H}^{(u)}$ is the probability of HO of user u from cell i to j while $\mathbf{H}^{(u)}$ is the probability transition matrix of the embedded discrete time Markov chain of user u. $F_{i,j}^{(u)}(t)$ is the sojourn time distribution of user u that is the probability that u will move from cell i to cell j at, or before time t. The probability of user u staying in cell i for no more than t time can be expressed as:

$$\Lambda_i^{(u)}(t) = Pr\Big(J_{n+1}^{(u)} - J_n^{(u)} \leq t | C_n^{(u)} = i\Big) \quad (8)$$

$$= \sum_{j=1}^{z} Q_{i,j}^{(u)}(t) \quad (9)$$

This also indicates the distribution of the sojourn time in cell i for user u, regardless of the next cell. Let $C^{(u)} = (C_t^{(u)}, t \in \mathbf{R_0^+})$ be another time-homogeneous semi-Markov process that describes the cell occupied by user u at time t. The transition probabilities for this process can be written as:

$$\chi_{i,j}^{(u)}(t) = Pr\Big(C_t^{(u)} = j | C_0^{(u)} = i\Big) \quad (10)$$

It gives the probability that a user u is in the cell j after the time instant t from the moment a transition to cell i has just been made. First for a special case that the user stays in cell i until the end of the period t is:

$$Pr(C_t^{(u)} = i | C_0^{(u)} = i, J_1 \geq t) \quad (11)$$

$$= Pr\Big(J_1 - J_0 \geq t | C_0^{(u)} = i\Big) = 1 - \Lambda_i^{(u)}(t) \quad (12)$$

For all other cases in which user u goes from from cell i to j through some intermediate cell $r \neq i$ is given as:

$$Pr\Big(C_t^{(u)} = j | C_0^u = i \text{ and at least one transition}\Big) \quad (13)$$

$$= \sum_{r=1}^{z} \int_0^t \frac{dQ_{i,r}^{(u)}(\tau)}{d\tau} \chi_{r,j}^{(u)}(t-\tau) d\tau \quad (14)$$

This is the Volterra equation of second kind and the integral is the convolution of $Q_{i,r}^{(u)}(.)$ and $\chi_{r,j}^{(u)}(.)$ i.e., $Q_{i,r}^{(u)} * \chi_{r,j}^{(u)}$. Here $Q_{i,r}^{(u)}(\tau)$ represents the probability of the user of staying in cell i for τ length of time and then transiting to cell r. Invoking the argument for the renewal of process here, expected behavior of user from here on is same irrespective of HO time to cell r. Therefore, $\chi_{r,j}^{(u)}(t-\tau)$ gives the probability of user being in cell j at time t given that user is in cell r at τ. Integration over τ takes care of all possible transition times [28]. Therefore,

$$\chi_{i,j}^{(u)}(t) = \Big(1 - \Lambda_i^{(u)}(t)\Big)\delta_{i,j} + \sum_{r=1}^{z} \int_0^t \frac{dQ_{i,r}^{(u)}(\tau)}{d\tau} \chi_{r,j}^{(u)}(t-\tau) d\tau \quad (15)$$

where $\delta_{i,j}$ is the Kronecker function that is only equal to 1 when $i = j$. We can solve equation (15) with approach given in [29]. To this end, the discrete-time version of evolution equation in (15) becomes:

$$\chi_{i,j}^{(u)}(s) = D_{i,j}^{(u)}(s) + \sum_{r=1}^{z} \sum_{\tau=1}^{s} \sigma_{i,r}^{(u)}(\tau) \chi_{r,j}^{(u)}(s-\tau) \quad (16)$$

where $D_{i,j}^{(u)}(s) = (1 - \Lambda_i^{(u)}(t))\delta_{i,j}$ and $\sigma_{i,r}^{(u)}(s) = \frac{dQ_{i,r}^{(u)}(\tau)}{d\tau}$ which is the probability to have a HO from cell i to r in the time s can be approximated as follows assuming unit time step:

$$\sigma_{i,r}^{(u)}(s) = \begin{cases} Q_{i,r}^{(u)}(1), & s = 1 \\ Q_{i,r}^{(u)}(s) - Q_{i,r}^{(u)}(s-1), & s > 1 \end{cases} \quad (17)$$

Due to $\mathbf{H}^{(u)}$ being a right stochastic matrix, $\mathbf{Q}^{(u)}(s)$ and $\boldsymbol{\chi}^{(u)}(s)$ will also be right stochastic matrices; i.e., $\sum_{j=1}^{z} Q_{i,j}^{(u)}(s) = \sum_{j=1}^{z} \chi_{i,j}^{(u)}(s) = 1, \forall i, j \in \mathcal{C}$. The $\chi_{i,j}^{(u)}(s)$ gives the probability that a user u is in the cell j in the time slot s counted from the moment a HO to cell i has just been made. In order to predict the location of a user in every s' time slots, we need to find the probability $\hat{\chi}_{i,j}^{(u)}(s,' o) = P(C_{o+s'}^{(u)} = j | C_0^{(u)} = i, t_{soj} = o)$ i.e., probability that a user is in cell j after s' time slot given that the current cell is i and user has stayed in cell i for sojourn time $t_{soj} = o$. $\hat{\chi}_{i,j}^{(u)}(s,' o)$ becomes [20]:

$$= \frac{P\Big(C_{o+s'}^{(u)} = j, t_{soj} = o, C_0^{(u)} = i\Big)}{P\Big(C_0^{(u)} = i, t_{soj} = o\Big)} \quad (18)$$

$$= \frac{P\Big(C_{o+s'}^{(u)} = j, t_{soj} = o | C_0^{(u)} = i\Big) P(C_0^{(u)} = i)}{P\Big(C_0^{(u)} = i, t_{soj} = o\Big)} \quad (19)$$

$$= \frac{P\left(C_{o+s'}^{(u)} = j, t_{soj} = o | C_0^{(u)} = i\right) P\left(C_0^{(u)} = i\right)}{P\left(t_{soj} = o | C_0^{(u)} = i\right) P\left(C_0^{(u)} = i\right)} \quad (20)$$

$$= \frac{P\left(C_{o+s'}^{(u)} = j, t_{soj} = o | C_0^{(u)} = i\right)}{P\left(t_{soj} = o | C_0^{(u)} = i\right)} \quad (21)$$

$$= \frac{D_{i,j}^{(u)}(o+s') + \sum_{r=1}^{z} \sum_{\tau=o+1}^{o+s'} \sigma_{i,r}^{(u)}(\tau) \chi_{r,j}^{(u)}(o+s'-\tau)}{1 - \Lambda_i^{(u)}(o)} \quad (22)$$

Note that just after HO i.e., $o = 0$, $\hat{\chi}_{i,j}^{(u)}(s,' o) = \chi_{i,j}^{(u)}(s)$. By mining the HO logs that contain information of the past handover information of user u, probability transition matrix $\mathbf{H}^{(u)}$ and sojourn time distribution matrix $\mathbf{F}^{(u)}$ are initialized as done in [22]. After each HO from cell i to j, $h_{i,j}^{(u)}$ and $F_{i,j}^{(u)}(s)$ are updated and $Q_{i,j}^{(u)}(s)$ is computed. Finally $\chi_{i,j}^{(u)}(s)$ and $\hat{\chi}_{i,j}^{(u)}(s,' o)$ are solved. In scope of this work, we choose future cell that has highest probability i.e.,$\max_{j \in \mathcal{N}_i} \hat{\chi}_{i,j}^{(u)}(s', o)$ where $\mathcal{N}_i$ is set of all cells whose coverage footprints overlap with cell i.

B. Coordinates-Level Location Estimation

Let $l_s^{(u)} = (x_s^{(u)}, y_s^{(u)})$ be the UE's current location coordinates in time slot s and $\{\mathcal{C}_N^{(u)}, \mathcal{T}_{HO}^{(u)}\}$ be the next cell HO tuple information for each UE wherein $\mathcal{C}_N^{(u)}$ is next probable cell of user u at time $\mathcal{T}_{HO}^{(u)}$. Leveraging future location estimation algorithm proposed by us in [30], future geographical coordinates at time step $s + s'$ are estimated as:

$$l_{s+s'}^{(u)} = l_s^{(u)} + \frac{\sqrt{\left(x_{\mathcal{C}_N^{(u)}}^g - x_s^{(u)}\right)^2 + \left(y_{\mathcal{C}_N^{(u)}}^g - y_s^{(u)}\right)^2}}{T_{HO}^{(u)}} * s' * \hat{u} \quad (23)$$

where $x_{\mathcal{C}_N^{(u)}}^g$ and $y_{\mathcal{C}_N^{(u)}}^g$ are the coordinates of most probable geomarker for UE u in next cell $\mathcal{C}_N^{(u)}$ (we utilize past mobility logs of UEs to estimate most probable geomarkers visited by each UE in each cell) and $\hat{u}$ is a unit vector pointing towards $(x_{\mathcal{C}_N^{(u)}}^g, y_{\mathcal{C}_N^{(u)}}^g)$.

C. Proactive Load Minimization Optimization

Leveraging predicted information $(\{\mathcal{C}_N^{(u)}, \mathcal{T}_{HO}^{(u)}\}, l_{s+s'}^{(u)})$ for all users, we formulate a load optimization problem for next time slot $s + s'$ in such a way that network load is minimized while meeting operator desired coverage ratio, QoE requirement of each UE and cell loads for next time window. The added advantage of targeting load minimization is that many QoS-related KPIs are monotonic functions of the average cell loads e.g., average throughput, latency and number of successful sessions etc. Due to monotonicity, minimizing cell loads improves network wide user throughputs and similar measures, and thus, LB minimization focused objective function can capture the goals of CCO objective too. Moreover, load minimization or load balancing increases the probability of the availability of free resources in all the cells that becomes advantageous for HetNets. To explain this point, consider a two cell scenario, for instance, wherein Cell X is bearing a load of 50% while cell Y is already at maximum load of 90%. If a mobile user enters Cell Y coverage area and requests service, the user will be denied and will have to be handed over to the cell X as the cell Y is already close to its maximum load utilization. This will result in lower QoE for the user compared to scenario where cell Y would have the free resources (residual capacity) to serve the oncoming user. A load minimization approach with minimum throughput guaranteed, solves this problem as it tries to minimize the load of the two cells in the first place without compromising QoE of existing users. Now as result of load balancing, if load utilization of both cells is at 70% and a new user enters any of the two cells, the cell will be able to accommodate this new user without additional delay.

The cell load η_c of a cell c can be defined based on the utilization of Physical Resource Blocks (PRBs) in the cell. The number of available PRBs at each base station (BS) is proportional to the available bandwidth and scheduling interval at that BS. The total load of cell c is the fraction of the total resources (PRBs) in the cell needed to provide required rate for all users of a cell and can be given as:

$$\eta_c = \frac{1}{N_b^c} \sum_{\mathcal{U}_c} \frac{\hat{\tau}_u}{\omega_B f(\gamma_u^c)} \quad (24)$$

where $\mathcal{U}_c$ is the number of users with active sessions connected to a cell c, $\hat{\tau}_u$ is the required/desired rate for user $u \in \mathcal{U}_c$, ω_B is bandwidth of a PRB, γ_u^c is the achievable SINR by the user u when conencted to cell c and N_b^c is the total number of PRBs in a cell. The function $f(\gamma_u^c)$ maps SINR to spectral efficiency of the user link and can be defined as $f(\gamma_u^c) = A \log_2(1 + B(\gamma_u^c))$. Here A and B constants can reflect post processing diversity gains through e.g., by MIMO and/or losses incurred in system. For sake of simplicity, without any loss in generality, we assume A and B as 1 in our simulations. The load in (24) by virtue of its definition is a virtual load since it can exceed one and thus can quantify how overloaded a cell is.

The SINR γ_u^c of user link to its cell c at its estimated location $l_{s+s'}^{(u)}$ in time slot $s + s'$ is defined as (25) shown at the bottom of the next page, where P_t^c is cell's transmission power; G_u is the gain of UE; λ_v is the weight assigned to the vertical beam pattern of the transmitter antenna; θ_u^c is the vertical angle of the user u in cell c with respect to horizon; θ_{tilt}^c is the tilt angle of the serving cell's antenna (at $\theta_{\text{tilt}}^c = 0^0$, BS antenna faces the horizon); φ_v is the vertical beam width of the transmitter antenna of cell c; λ_h is the weighting factor for the horizontal beam pattern; ϕ_u^c is the horizontal angle of user u in cell c with respect to absolute north; ϕ_a^c is the azimuth of the antenna of cell c ($\phi_a^c = 0^0$ corresponds to the absolute north); φ_h is the horizontal beam width of the transmitter antenna of cell c; δ_u^c denotes the shadowing observed at the location of user u from cell c; α is the path loss constant; d_u^c represents the distance of the estimated user location of u i.e., $l_{s+s'}^{(u)}$ from cell c; β is the path loss exponent; and κ is the

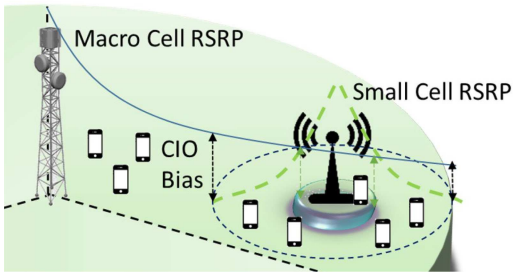


Fig. 3. CIO bias.

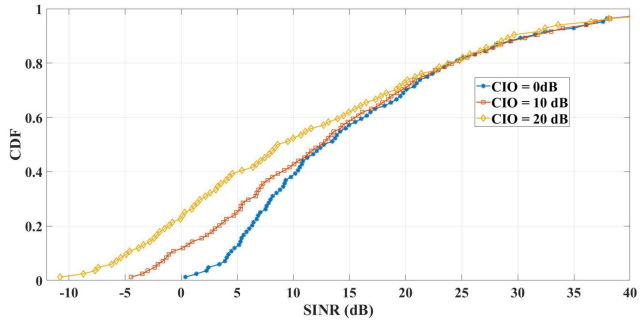


Fig. 4. Average UE SINR (dB) vs. CIOs.

noise variable. The time subscript on the right hand side of (25) denotes that all terms enclosed within square brackets $[.]_{s+s'}$ are considered for the next time slot $s+s'$. In current work, we assume that C-SON server running in the core network is able to estimate shadowing at all locations with normally distributed error by leveraging channel maps. These maps are built based on MDT reports, a 3GPP standardized feature, wherein all UEs report their geo-tagged time stamped channel measurements back to the network. In (25), the cell load utilization η_i in the denominator can be thought of as probability of transmission of BS i while the sum reflects the average interference power. In contrast to an exact time dependent SINR formulation that results into range of SINR values that vary depending upon the scheduling instants and load of other cells, with this approach of mean interference, we can easily evaluate SINR with low complexity and tractability. On average, more interference will come from cells that are more loaded. The UEs in idle or connected mode will be associated with the cell that ranks highest according to following user association criterion:

$$\mathcal{U}_j := \left\{ \forall u \in \mathcal{U} \mid j = \arg\max_{\forall c \in \mathcal{C}} (P^c_{r,u_{dBm}} + P^c_{\mathrm{CIO}_{dB}}) \right\} \quad (26)$$

where $P^c_{r,u_{dBm}}$ is the actual reference signal receivce power in dBm that user u is getting from cell c and $P^c_{\mathrm{CIO}_{dB}}$ is the small cell attraction bias parameter (CIO). The term CIO accounts for various biases used in idle and active mode procedures [9]. The CIO is attraction factor that is broadcasted by small cells to bias their ranking and attract users to camp on them. This way power disparity in macro and small cell transmissions powers is avoided and more load can be transferred to them (Fig. 3). CIO, as a stand-alone solution, addresses the selection between different network layers in HetNets; however, it has catastrophic affect on user SINR since through artificial biasing, UE is no longer connected the strongest cell. As a consequence, SINR deteriorates with higher values of CIO as illustrated in Fig. 4. Nevertheless, CIO is still relevant network parameter for load balancing albeit at cost of CCO if used in legacy way for LB [3], [4], [6]–[8]. The negative influence of degraded SINR on user throughput can be partially offset if small cell can allocate enough surplus PRBs compared to macro cell and thus satisfy required QoE. Hence CIO is a vital control parameter to balance the tradeoff between LB and CCO. Moreover, we also leverage user association criterion proposed by us in [12] that also takes the cell load into consideration defined as:

$$\mathcal{U}_j := \left\{ \forall u \in \mathcal{U} \mid j = \arg\max_{\forall c \in \mathcal{C}} \left(\left(\frac{1}{\eta_c} \right)^a * \left(P^c_{r,u_{dBm}} + P^c_{CIO_{dB}} \right)^{(1-a)} \right) \right\} \quad (27)$$

where η_c is the cell load and $a \in [0,1]$ is the weighting factor in order to associate a level of priority to load and RSRP metrics. Large value of a forces users to avoid highly loaded BSs even if they provide good RSRP. Note that setting $a = 0$ will make it equivalent to (26). With cell association method defined by (27), user is associated with such a cell with whom the product of the received power $(P^c_{r,u_{dBm}} + P^c_{CIO_{dB}})$ and reciprocal of cell load is maximum. Note for cell association criterion, η_c cannot be 0 therefore for unloaded cells, η_c can be set as a very small number $\epsilon \to 0$.

Note that in our case where all UEs are assumed to be active demanding constant bit rate service, user satisfaction ratio is more relevant performance metric then conventional throughput. The reason being that for load optimization with guaranteed QoS requirements, UEs either get exactly the desired constant bit rate or remain unsatisfied. The number of unsatisfied users (dropped/blocked) "N_{us}" is given as [31]:

$$N_{us}(s+s') = \left[\sum_c \max \left(0, \sum_{\mathcal{U}_c} \mathbb{1}. \left(1 - \frac{1}{\eta_c} \right) \right) \right]_{s+s'} \quad (28)$$

The η_c in (28) by definition from (24) has range $\eta_c \in [0, \infty)$ to quantify overloading in a cell. When cell is fully loaded i.e., $\eta_c = 1$, the inner sum in (28) will be zero which means all users in cell c are satisfied. If cell load exceeds 1 e.g., $\eta_c = 2$, inner

$$\gamma^c_u(s+s') = \left[\frac{P^c_t G_u 10^{-1.2\left(\lambda_v \left(\frac{\theta^c_u - \theta^c_{\mathrm{tilt}}}{\varphi_v}\right)^2 + \lambda_h \left(\frac{\phi^c_u - \phi^c_a}{\varphi_h}\right)^2\right)} \delta^c_u \alpha \left(d^c_u\right)^{-\beta}}{\kappa + \sum_{\forall i \in \mathcal{C}/c} \eta_i P^i_t G_u 10^{-1.2\left(\lambda_v \left(\frac{\theta^i_u - \theta^i_{\mathrm{tilt}}}{\varphi_v}\right)^2 + \lambda_h \left(\frac{\phi^i_u - \phi^i_a}{\varphi_h}\right)^2\right)} \delta^i_u \alpha \left(d^i_u\right)^{-\beta}} \right]_{s+s'} \quad (25)$$

sum will evaluate to half of the number of users of cell c. This means the cell in reality is fully loaded. Half of the users are satisfied while other half of oncoming users will be blocked.

Based on the works of [32], we use optimization objective function that is parameterized function of the BS loads. The objective function considered is:

$$\Phi(\eta) = \begin{cases} \sum_{i\in\mathcal{C}} \frac{(1-\eta_i)^{1-\xi}}{\xi-1}, & \text{for } \xi \neq 1 \\ \sum_{i\in\mathcal{C}} -\log(1-\eta_i), & \text{for } \xi = 1 \end{cases} \tag{29}$$

where $\xi \geq 0$ is a parameter that induces the desired degree of load balancing. For $\xi = 0$, (29) reduces to maximizing the arithmetic mean of the BS' free resources. When $\xi = 1$, (29) is equivalent to maximizing the geometric mean of the resources available in the network. When $\xi = 2$, the harmonic mean of the BSs' free resources is maximized. Increasing ξ further to ∞ minimizes the maximum utilization, i.e., min-max utilization which yields solutions with balanced loads. The value of ξ in general depends on network operators' preferences and policies. It should be noted that load balancing does not necessarily aim at equalizing the loads of all BSs since different values of ξ have different implications. In this work, we chose $\xi = 1$ since it prevents overload situation (logarithmic term tend to infinity for overloaded scenarios) and minimizes the total system load with notion of fairness rather than distributing load equally among cells.

The load minimization optimization problem formulated for next time slot $s + s'$ as (30)–(33) shown at the bottom of this page.

Since η_c denotes the resource utilization of cell c, term $(1 - \eta_c)$, hence forth noted as residual capacity, is fraction of resources in cell c ready to be allocated to users. The objective is to optimize the parameters P_t^c, θ_{tilt}^c, ϕ_a^c, φ_v^c, φ_h^c, P_{CIO}^c such that logarithmic sum of idle resources in all cells is maximized while ensuring coverage reliability and QoE requirements. The log utility function leads to a kind of proportional fair treatment of the individual cells while minimizing cell loads or maximizing residual capacity. The first six constraints (33a)–(33f) define

$$\min_{P_t^c,\theta_{\text{tilt}}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{\text{CIO}}^c} \sum_c [-\log(1-\eta_c(P_t^c,\theta_{\text{tilt}}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{\text{CIO}}^c))]_{s+s'} \tag{30}$$

$$\min_{P_t^c,\theta_{\text{tilt}}^c,\phi_a^c,\varphi_v^c,\varphi_h^c,P_{\text{CIO}}^c} \times \sum_c -\log\left[1 - \frac{1}{N_b^c}\sum_{u_c} \frac{\hat{\tau}_u}{\omega_B \log_2\left(1 + \frac{P_t^c G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^c-\theta_{\text{tilt}}^c}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^c-\phi_a^c}{\varphi_h}\right)^2\right)}\delta_u^c\alpha(d_u^c)^{-\beta}}{\kappa+\sum_{\forall i\in\mathcal{C}/c}\eta_i P_t^i G_u 10^{-1.2\left(\lambda_v\left(\frac{\theta_u^i-\theta_{\text{tilt}}^i}{\varphi_v}\right)^2+\lambda_h\left(\frac{\phi_u^i-\phi_a^i}{\varphi_h}\right)^2\right)}\delta_u^i\alpha(d_u^i)^{-\beta}}\right)}\right]_{s+s'} \tag{31}$$

where

$$\mathcal{U}_j := \left\{\forall u \in \mathcal{U} \mid j = \arg\max_{\forall c\in\mathcal{C}} \left(\left(\frac{1}{\eta_c}\right)^a * \left(P_{r,u_{dBm}}^c + P_{CIO_{dB}}^c\right)^{(1-a)}\right)\right\} \tag{32}$$

$$P_{t,\min} \leq P_t^c \leq P_{t,\max} \forall c \in \mathcal{C} \tag{33a}$$

$$\theta_{\min} \leq \theta_{\text{tilt}}^c \leq \theta_{\max} \forall c \in \mathcal{C} \tag{33b}$$

$$\phi_{\min} \leq \phi_t^a \leq \phi_{\max} \forall c \in \mathcal{C} \tag{33c}$$

$$\varphi_{v,\min} \leq \varphi_v^c \leq \varphi_{v,\max} \forall c \in \mathcal{C} \tag{33d}$$

$$\varphi_{h,\min} \leq \varphi_h^c \leq \varphi_{h,\max} \forall c \in \mathcal{C} \tag{33e}$$

$$P_{\text{CIO},\min} \leq P_{\text{CIO}}^c \leq P_{\text{CIO},\max} \forall c \in \mathcal{C} \tag{33f}$$

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

$$\tau_u \geq \hat{\tau}_u \forall u \in \mathcal{U} \tag{33h}$$

$$\eta_c < 1 \forall c \in \mathcal{C} \tag{33i}$$

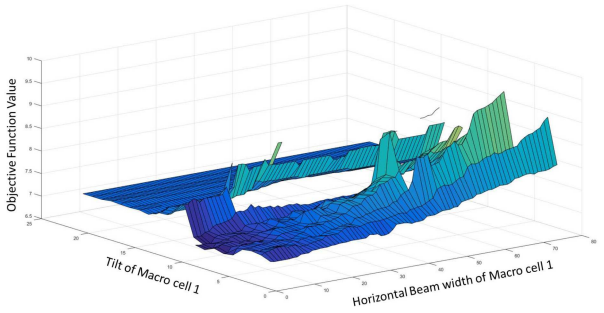


Fig. 5. Non-convexity behavior of the objective function.

the limits for the variation in the Tx power, tilts, azimuths, beam widths (vertical, horizontal) and CIOs respectively. These constraints determine the size of solution search space. The seventh constraint (33g) ensures that with new parameters settings, network meets at least minimum network coverage threshold $\bar{\omega}$, a QoS KPI set by the operator. P_{th}^{c} is the minimum acceptable threshold level for received power for user below which no session can successfully be established. The eighth constraint (33h) ensures each covered user is satisfied meaning it receives minimum guaranteed throughput that is required depending upon the subscription level or session types. This constraint is needed because for achieving LB objective, if CIO is leveraged to tune actual RSRP based cell association for the user, the received power $P_{r,u}^{c}$ for offloaded user may become worse, and consequently the SINR and throughput for that user will be impacted. The loss in SINR can be neutralized by allocating surplus resources given that the CIO biased user received power is above a certain threshold. Consequently, minimum throughput is assured for the users in network by this constraint (implicit CCO objective). This is possible only when cell has sufficient resources to meet total capacity requested, therefore, constraint in (33i) is needed to ensures that load for every cell has to be less then 1 $\eta_c < 1$.

The objective function, optimization variables and constraints indicate it is a large-scale non-convex NP-hard problem due to the inherent coupling of optimization parameters and the cell loads. Non convexity stems mainly from the fact that we are dealing with not one or two but five parameters per macrocell i.e., cell transmit power, antenna tilt, azimuth, horizontal beamwidth and vertical beamwidth and two independent paramters per small cell i.e., transmit power and CIO with inter-coupled effects on the objective function. In total, the solution space for the network system will have $147 = 21 \times 5 + 21 \times 2$ distinct and independent optimization parameters. This means that even if each parameter can take only 2 values, we will end up with 2^{147} distinct combinations in the solution space that becomes computationally prohibitive. The plot of the objective function for a sample topology of 42 cells is shown in Fig. 5 wherein tilt and horizontal beam width of a base station are varied while rest of all variables are kept constant. It can be observed that solution space is combination of multiple hills and valleys (non-convex). As the number of possible combinations for the optimization parameters considered increases exponentially with network density, a brute-force style strategy for search of the optimal parameters to achieve the load minimization may become impractical for large size network. In a practical network of 100 cells with only 10 tilt values per cell available as optimization variables, number of combinations 10^{100} become greater than total number of atoms in universe. Clearly this search space size is unfathomable, mostly filled with suboptimal points and is too large to be traversed by brute force algorithm in as short time as LTE's transmission time interval (TTI).

For solving the formulated proactive LB problem for next time slot $s + s'$ in real time, we experimented with several heuristics and found hybrid combination of Genetic Algorithm (GA) and Pattern Search (PS) to perform the best. Genetic Algorithms are class of artificial intelligence algorithms based on Darwin's "survival of the fittest, natural selection" theory of evolution. GA is population based search algorithm that uses randomized operators mimicking natural selection processes like crossover and mutation operating over a population of candidate solutions to generate new points in the search space. GA are theoretically and empirically proven to provide robust, efficient and effective search capabilities in complex multivariable combinatorial search spaces. The inherent randomness significantly increase the probability of jumping out local search space to achieve optimal solutions in global space. GA are known to find feasible regions relatively quickly but convergence time to find optimal point is usually very large. Therefore to overcome this issue, we used hybrid augmentation scheme wherein GA is first unleashed on unfathomable search space peculiar to cellular networks to find feasible region. Once there, the optimization search process is handed over to Pattern Search algorithm that are efficient for local search. Therefore based on estimated future network state (i.e. cell loads) in time slot $s + s'$, OPERA framework optimizes network parameters to their optimal values ahead of time such that load balancing is achieved. Note that for stability issues, optimization parameter values remain fixed from time slot s to s'. The optimization algorithms need some time to converge. However, thanks to proactiveness powered by load prediction instead of observation as is the case with most existing LB solutions [3]–[12], the proposed strategy gives considerable time s' to find feasible solution.

III. Performance Evaluation

In this section, we present the results for our proposed OPERA framework. We have gauged its performance against three benchmark schemes. (i) The first scheme comprises real mobile network deployment settings—RDS-A, RDS-B, and RDS-C that are the three most common configurations adapted from real network LTE deployment settings for one of USA's national mobile operator in city of Tulsa with RDS-A (Tilt: 3^0) and RDS-B (Tilt: 5^0) both using antenna [33] and RDS-C (Tilt: 4^0) using antenna [34]. (ii) The second scheme (a phenomenal work) is a Joint algorithm (referred to as Joint1 in [9]) that is quite relevant and has inspired the proposed work wherein LB is achieved via tilts with coverage constraints. It is used as a representative of state-of-art reactive schemes simulated by inducing artificial delay in getting user location information; i.e., the scheme is implemented for location information from the

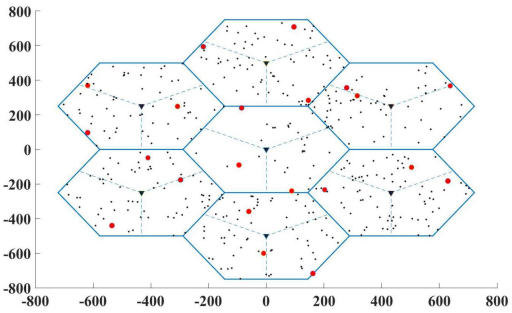


Fig. 6. Network topology with black dots indicating UEs and SCs are illustrated by red circles.

TABLE I
NETWORK SIMULATION PARAMETERS

System Parameters	Values
Topology	Hexagonal - 7 Macro Cells with 3 Sectors per Base Station
Number of Small Cells	1 per Sector
Number of UEs	84 Mobile, 336 Stationary
LTE System Parameters	Frequency = 2 GHz, Bandwidth = 10 MHz, ISD: 500m
Macro Cell Tx Parameters	Tx Power: 40 to 46 dBm, Tilt: 90^0 to 120^0 Azimuth: -45^0 to 45^0 Horizontal Beamwidth: 45^0 to 120^0 Vertical Beamwidth: 5^0 to 15^0
Small Cell Tx Parameters	Tx Power = 27 - 30 dBm, CIO = 0 - 10 dB
Node Heights	Macro Cell = 25 m, Small Cell = 10 m, UE = 1.5 m
Antenna Gains	Macro Cell = 17 dBi, Small Cell = 5 dBi
Macro Cell Antenna Side Lob level Suppression	Side lobe level suppression of combined antenna = 25 dB Side lobe level suppression of azimuth antenna = 25 dB Side lobe level suppression of elevation antenna = 20 dB
UE Noise figure	7 dB
Coverage Ratio	100%
Simulation Time	1 hour

previous 1 minute. The reactive style optimization is also done on a per minute basis just as for OPERA i.e., it is done every minute for the whole 7th day (a total of $60 \times 24 = 1440$ evaluation points). One thing to clarify here is that for fair comparison, we implemented the algorithm in [9] using load-aware user association (27). (iii) The third scheme is near-optimal performance bound (NARN) that is OPERA with 100% prediction accuracy. NARN (OPERA) leverage a conventional association strategy $(a = 0)$ in (27) while NARN*(OPERA*) uses a load-aware scheme with $a = 0.5$ in (27).

A. Simulation Settings

We generated typical macro and small cell based network topologies and UE distributions in matlab following 3GPP specifications that are widely used in industrial simulations found in [35] and [36]. The path loss and shadow fading vary with carrier frequency whether the UE link is LOS or NLOS. The detailed expressions for pathloss model used are given in table A1-2 of [36]. The typical flow of simulation is as follows: At each time slot, for a given network parameters configuration (that is set by the optimization algorithm) and UE position set by the mobility traces, (i) a large scale channel is generated between UEs and base stations (ii) the path loss, shadow fading, sectorized antenna gains and other miscellaneous losses are generated (iii) the combined gains of the horizontal (azimuth) sectorized and the vertical (elevation) antennas for a given UE to all base cells/sectors is generated (iv) Each UE is associated to one macro cell or small cell based on the association criterion used satisfying the handover margin and SINR is calculated (v) PRBs are assigned to UEs based on their required throughputs and achievable SINRs (vi) cell loads and KPIs of interest are calculated.

The multi-tier HetNet deployment simulated consists of a primary tier represented by macrocells, and secondary tier comprising of small cells that share the same spectrum with the primary tier. Snapshot of the network topology at of one of the instants is shown in Fig. 6, and the simulation parameter details are given in Table I. To eliminate any artifacts due to boundary effects limitations, a wrap around model is used to simulate an infinitely large network without requiring large number of cells. For realistic evaluations, clustered based UE depolyment is considered wherein some of the UEs are distributed non-uniformly by clustering them around a random hotspot in each sector. We capture the variation of the network conditions through Monte Carlo style simulations. The performance of OPERA highly depends on the movement patterns of simulated UEs. Majority of relevant works leverage random waypoint mobility model wherein trajectory is completely random and unrealistic. Naturally this kind of model is not suitable especially when objective is to assess performance of mobility prediction schemes. Therefore, for accurately gauging performance of the proposed work, selection of appropriate mobility model was key step since the performance analysis of OPERA done using realistic mobility traces is going to be plausible representative of its actual performance in the real environment. Recently some realistic mobility models have come to limelight such as SLAW, SMOOTH etc [37]. Among them, only SLAW-model-generated mobility traces [38] have been shown to capture all the statistical characteristics of mobility patterns in cellular networks like (i) truncated power-law distributed length of human flights, pause times and inter-contact times; (ii) each person having his/her own confined mobility region; (iii) attraction of people to famous landmarks. Therefore for realistic performance evaluation of our framework, we selected SLAW for our simulations. SLAW model based one week HO traces were generated for 84 mobile users. Six days data was used for building semi-Markov model. Since in real networks, 80% of traffic is generated indoor [39] therefore additional 336 stationary UEs are deployed to increase loading on the network. We consider uniformly distributed five different UE traffic requirement profiles corresponding to 24 kbps, 56 kbps, 128 kbps, 1024 kbps and 2048 kbps desired throughputs. Considering typical time period after which updating the parameter may be practical, we use 1 minute value for the prediction interval s' in our simulation study. Therefore, every minute, OPERA predicts future location of users for next time slot and perform optimization and this continues for whole day (a total of 1440 evaluation points).

B. Results and Discussion

We first evaluate prediction performance of the mobility predictor i.e., semi-Markov model trained on six days mobility patterns and tested on seventh day's dataset. The input training data for the semi-Markov predictor is time-stamped cell association record for all UEs containing two fields (Time and Serving

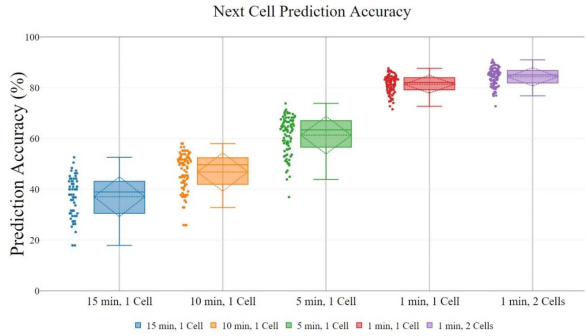


Fig. 7. Next Cell Mobility Prediction Accuracy for various {Prediction Interval, n-Cell Prediction} combinations.

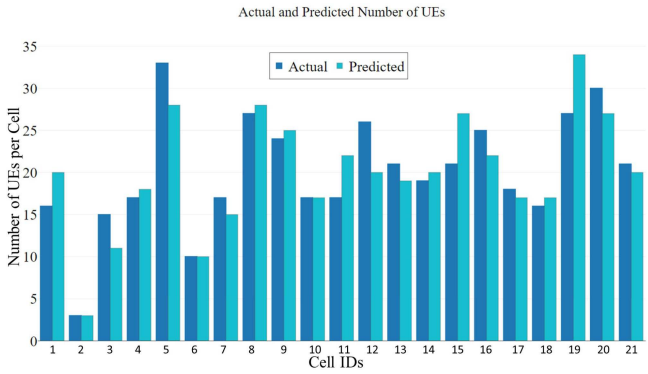


Fig. 8. Actual and Predicted Number of UEs per cell.

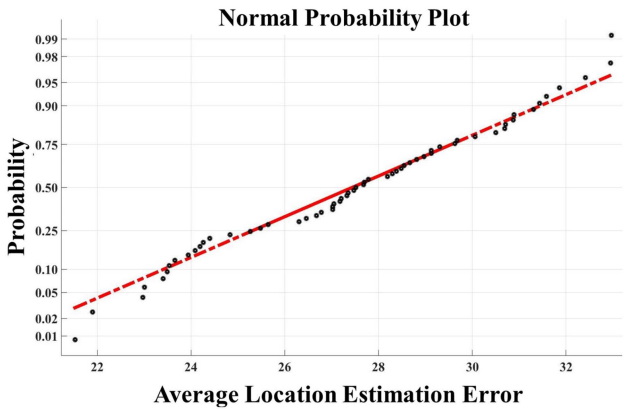


Fig. 9. Normal Probability Plot for Average Location Estimation Error.

Cell) i.e., at time t_1, UE_1 is associated with cell x. The time granularity chosen was 1 minute interval. In real networks, this record can be extracted from CDRs or handover reports. In each time slot s, next cell is predicted for next time slot $s + s'$ using (16) and (22) and prediction accuracy is computed which is measured as percentage of correct predictions of the next cell to visit in next time slot $s + s'$. Fig. 7 plots prediction accuracy for various combinations of prediction interval and number of most probable cells. Comparing 1-Cell prediction with 2-Cell prediction, we observe that prediction accuracy improves when for next time slot, we have more than one potential future location. The average value reach upto 84.39%. This is expected because spatial resolution has decreased (coarse prediction). On the other hand, given 1-cell prediction only, prediction accuracy improves (81.46% average prediction accuracy) with decrease in prediction interval length for $s' = 1$ min. With smaller prediction window size, UE is less probable to move to large distances and hence accuracy improves. These high accuracies observed with semi-Markov model trained/tested on SLAW generated traces are in line with other studies that are based on real HO traces collected from live cellular networks [22]. The prediction interval window size is constrained by the convergence time of Genetic Algorithm and Pattern Search heuristics algorithms. With the available resources for this study, minimum amount of 1 minute was required to find feasible solutions therefore we set $s' = 1$ minute in our simulations.

Next we compared the actual and predicted number of UEs per cell. Let $|\mathcal{U}_j(t+1)|$ be the number of users predicted to be in cell j at time $t + 1$. This consists of users who (i) just entered into cell i at time t and will be in cell j at time $t + 1$ given by the following equation:

$$\mathcal{U}_j(t+1) := \left\{ \forall u \in \mathcal{U} | j = \arg\max_{r \in \mathcal{C}} (\chi_{i,r}^{(u)}(s=1)) \right\} \quad (34)$$

and (ii) users who are in cell i and have stayed in cell i for sojourn time $t_{soj} = o$ and will be in cell j at time $t + 1$ given by the following equation:

$$\mathcal{U}'_j(t+1) := \left\{ \forall u \in \mathcal{U} | j = \arg\max_{r \in \mathcal{C}} (\hat{\chi}_{i,r}^{(u)}(s'=1, o)) \right\} \quad (35)$$

Therefore, the total number of UEs predicted to be in cell j at time $t + 1$ will be as follows:

$$|\mathcal{U}_j(t+1)| = |\mathcal{U}_j(t+1)| + |\mathcal{U}'_j(t+1)| \quad (36)$$

As evident in the Fig. 8, the mobility prediction model is able to predict the number of UEs in most of the cells at the next time interval with high accuracy. Algorithm 1 proposed by us in [30] was used to estimate the location of UEs for one hour simulation duration after every s' time slots. On average, location estimation algorithm exhibited distance error (distance between estimated and actual coordinates) of 27.5 meters with maximum value of around 33 meters. The normal probability plot for average location estimation error is shown in Fig. 9 that is a graphical technique to identify normality in observations. Samples from normal distribution follow straight line. As per the figure, error in location estimation can be approximated by normal distribution. Fig. 10 plots the histogram of difference (error) between predicted and actual load values with OPERA that leverage semi-Markov based future location algorithm [30]. It is observed that most of the error falls into 0.05 bin with root mean square error (RMSE) of 0.2711.

Next, the offered cell load CDFs for all the cells with Real Deployment Settings, Joint, and proposed schemes is shown in Fig. 11. It is evident from the plot that with Joint, majority of the cells remain overloaded. The reason can be attributed to (i) reactive approach and (ii) usage of only tilt as optimization parameter. This increases the overloading or the percentage of

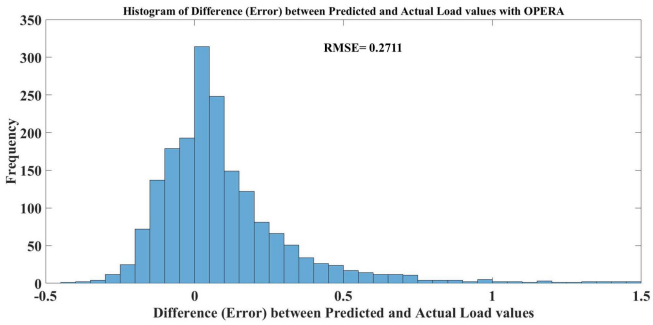


Fig. 10. Histogram of Error between Predicted and Actual Load values.

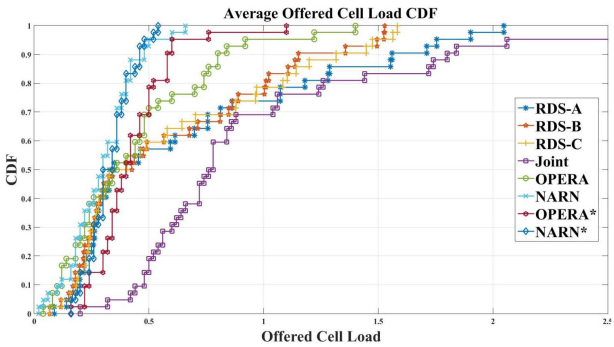


Fig. 11. Average Offered Cell Loads CDF of all cells.

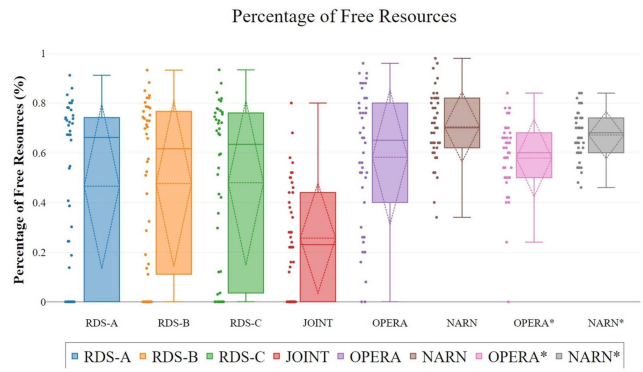


Fig. 12. Box plot of percentage of free resources in the cells.

unsatisfied users (as shown in Fig. 13). Same trend is observed for the Real Deployment Settings wherein cells remain overloaded with overloaded cells maximum in RDS-A (around 26%), followed by RDS-C (around 23%) and RDS-B (around 21%) respectively. Compared to these fixed configurations settings and Reactive schemes, the proposed solution OPERA and OPERA* achieve load reduction purely by increasing resource efficiency through dexterous optimization of antenna parameters (transmission power, tilts, azimuths, beam widths) and CIOs such that the cell loads are substantially reduced. Although slight overloading is observed with OPERA (OPERA*) of around 4% (2%) that is due to the prediction inaccuracies. This overloading is mitigated when prediction accuracy reaches 100% which is shown by NARN and NARN* wherein maximum cell loads are 66% and 54% respectively. It is observed that inclusion of load

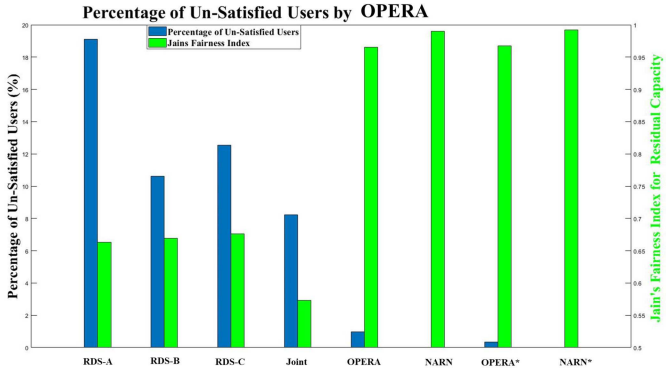


Fig. 13. Percentage of Un-Satisfied users.

metric in association criterion i.e., making the cell association scheme load aware as proposed in (27) [12] improves the residual capacity fairness in all cells. As a result of this even in presence of prediction inaccuracies, cells have more free capacity to accommodate actual extra load as compared to a less predicted load.

Fig. 12 shows the box plot of percentage of free resources among all the cells achievable with the RDS, Reactive and proposed schemes. The inclusion of load metric in user association criterion as defined by (27) in OPERA* and NARN* results in less variance in residual capacity as compared to rest of the schemes. Note that the OPERA and OPERA* result in some cells with no free resources. This is due to the prediction inaccuracies. This zero residual capacity scenario is avoided with NARN and NARN*. The variance in cell loads is further analyzed using Jains Fairness index calculated through (37) and plotted in Fig. 13 wherein the average percentage of un-satisfied users is visualized on left y-axis while Jains Fairness index for residual capacity is plotted on right y-axis achievable with the RDS, Reactive and proposed schemes.

$$\text{JFI}(1-\eta_c) = \frac{(\sum_c (1-\eta_c))^2}{(|C| \times \sum_c (1-\eta_c)^2)} \tag{37}$$

The result computed from (37) ranges from $(1/|\mathcal{C}|)$ (worst case) to 1 (best case), and it is maximum when all the cells have the same amount of free residual capacity. Due to maximum overloading experienced with conventional RDS and Reactive schemes, considerable number of users face blocking and become unsatisfied. Load aware association based proposed schemes OPERA* (NARN*) achieve maximum fairness of 0.967 (0.992) as compared to their contemporaries OPERA (NARN) with fairness of 0.965 (0.989). This fairness helps to reduce the percentage of unsatisfied users from 0.98% in OPERA to 0.35% in OPERA*. It is interesting to observe that even in presence of prediction inaccuracy, percentage of satisfied users is above 99% with OPERA. Fig. 14 plots the CDFs for the achievable UE SINRs with the RDS, reactive, and proposed schemes. For reactive and RDS schemes, SINR drops as compared to other schemes. The reason is that maximum loaded macro cells cause more network-wide interference,

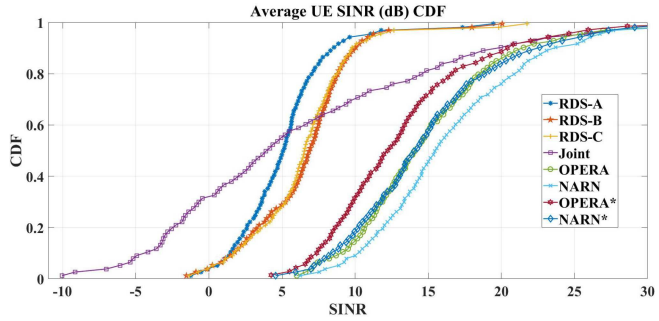


Fig. 14. Average UE SINR CDF.

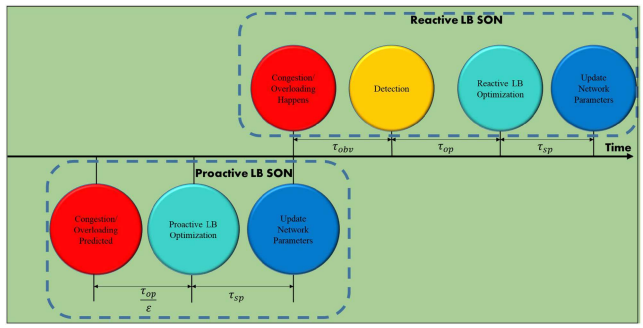


Fig. 15. Time line diagram for Proactive and Reactive LB SON functions.

which reduces the achievable SINR of the UEs. This interference footprint of macro cells becomes highly contained with the proposed schemes (OPERA and OPERA*) by optimizing the values of antenna parameters and CIOs such that SINR is enhanced and cell loads are minimized. Moreover, the inclusion of the load metric in the association scheme (OPERA* and NARN*) reduces the achievable SINR of the UEs, as the UEs are not connected to the strongest possible cell. Despite decreasing SINR for NARN*, as compared to NARN, the solution manages to deliver the gains observed in Fig. 13, mainly because of load fairness by optimizing the horizontal and vertical beam widths, tilts, azimuths, Tx power, and CIOs. Actually with CIO in use, SINR is bound to deteriorate; however, this can be taken care of if sufficient PRBs are available to offset the loss caused by the lower SINR. This compensating act is why OPERA* and NARN* outperform, hence the gain in resource utilization is observed.

C. Complexity Analysis

The complexity of OPERA framework depends upon time complexity of (i) semi-Markov based mobility prediction model, (ii) future location estimation algorithm, and (iii) the heuristic algorithm to solve optimization problem. As per [20], time complexity in computing (22) for all users in the network in each time slot of duration s' is $O(s'|\mathcal{U}||\mathcal{C}|^2)$ once all required parameters have been evaluated which is not a significant overhead. Time complexity of location estimation algorithm increases linearly with number of geo-markers $|L|$ in each cell. For the heuristic algorithm, considering GA alone with G as maximum number of iterations (generations) and P as the number of solution space samples per iteration, execution time complexity is $O(GP)$ [40]. Hence total runtime of OPERA framework can be generalized as $O(GP|L|s'|\mathcal{U}||\mathcal{C}|^2)$. The proactiveness of OPERA minimizes impact of this execution time on subscriber QoE. If τ_{obv} is the time needed to detect overloading in the network, τ_{op} the time needed to solve NP-hard non-convex load balancing problem and τ_{sp} as time needed to change network parameters to new settings then total degradation time in the network is sum of τ_{obv}, τ_{op} and τ_{sp}. With proactive optimization strategy of OPERA, this degradation time become zero if sum of τ_{op} and τ_{sp} is less than or equal to prediction window size s'. Moreover load-aware-user association and hybrid heuristic combination technique further reduces τ_{op} by some factor ϵ i.e., $\frac{\tau_{op}}{\epsilon}$ which lessens strain on selection of prediction window size (see Fig. 15).

IV. Conclusion & Future Works

In this paper we have proposed a novel spatiotemporal mobility prediction based proactive load balancing optimization framework for HetNets by jointly optimizing Tx power, tilts, azimuths, beam widths and CIOs. The proposed OPERA framework employs innovative concept of estimating future user locations and leverages that to estimate future cell loads. We then formulate a system level fairnesss aware load optimization problem for the estimated future cell specific loads. The majority of the current load balancing solutions are reactive and are designed to perform LB dynamically in real-time after observing the congestion. With reactive approach it is close to impossible to meet 5G ambitious QoS requirements even when substantial computing resources are available. Keeping this in view, the proposed approach makes it possible to solve LB optimization problem in real time without jeopardizing the QoE. Moreover, OPERA framework accounts for the interplay between two intertwined SON functions (LB and CCO) and thus ensures conflict free operation. A load aware association strategy that underpins OPERA further bolsters the framework against location estimation accuracies and maximizes system level capacity and QoE in addition to balancing load. Extensive simulations leveraging realistic mobility patterns indicate that, in best case, OPERA can reduce percentage of unsatisfied users to 0.35% despite of acute mobility and heterogeneity of cell sizes. The presented results highlight the value of prediction (AI) based proactive optimization.

For future work, vehicular mobility traces will be used since in case of vehicles, the trajectory direction of mobility traces will be more deterministic and regular as vehicles can only follow the road topology as compared to pedestrians who can go through any direction. On top of that, the knowledge of street/road layout and the navigation App data e.g., google maps navigation that determines the trajectory can be exploited to maintain accuracy. Thus intuitively, it is expected that by focusing on vehicular mobility the performance of proposed solution is likely to improve. However, superposing the road maps and speed data to achieve higher accuracy is a separate research study that is beyond scope of this paper and will be subject of future study. Moreover, machine learning predictors

like deep neural networks and gradient boosting trees will be employed in place of semi-Markov in OPERA framework that are recently being investigated heavily for cellular networks optimization like in [41], [42] and end-to-end gains will be evaluated.

REFERENCES

[1] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An overview of load balancing in hetnets: Old myths and open problems," *IEEE Wireless Commun.*, vol. 21, no. 2, pp. 18–25, Apr. 2014.

[2] 3GPP, "LTE; evolved universal terrestrial radio access network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (version 9.2.0 Release 9)," Sophia Antipolis, France, Tech. Rep. TR 36.902, 2010. [Online]. Available: https://www.etsi.org/deliver/etsi_tr/136900_136999/136902/09.02.00_60/tr_136902v090200p.pdf

[3] P. Kreuger, O. Gornerup, D. Gillblad, T. Lundborg, D. Corcoran, and A. Ermedahl, "Autonomous load balancing of heterogeneous networks," in *Proc. IEEE 81st Veh. Technol. Conf.*, May 2015, pp. 1–5.

[4] P. Muñoz, R. Barco, J. M. Ruiz-Avilés, I. de la Bandera, and A. Aguilar, "Fuzzy rule-based reinforcement learning for load balancing techniques in enterprise LTE femtocells," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1962–1973, Jun. 2013.

[5] Z. Li, H. Wang, Z. Pan, N. Liu, and X. You, "Heterogenous QoS-guaranteed load balancing in 3GPP LTE multicell fractional frequency reuse network," *Trans. Emerg. Telecommun. Technol.*, vol. 25, no. 12, pp. 1169–1183, Dec. 2014. [Online]. Available: http://dx.doi.org/10.1002/ett.2676

[6] A. Lobinger, S. Stefanski, T. Jansen, and I. Balan, "Load balancing in downlink LTE self-optimizing networks," in *Proc. IEEE 71st Veh. Technol. Conf.*, May 2010, pp. 1–5.

[7] M. Sheng, C. Yang, Y. Zhang, and J. Li, "Zone-based load balancing in LTE self-optimizing networks: A game-theoretic approach," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2916–2925, Jul. 2014.

[8] I. Viering, M. Dottling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2009, pp. 1–6.

[9] A. J. Fehske, H. Klessig, J. Voigt, and G. P. Fettweis, "Concurrent load-aware adjustment of user association and antenna tilts in self-organizing radio networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1974–1988, Jun. 2013.

[10] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[11] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, May 2013.

[12] A. Asghar, H. Farooq, and A. Imran, "Concurrent optimization of coverage, capacity and load balance in hetnets through soft and hard cell association parameters," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8781–8795, Sep. 2018.

[13] I. Sobol, "Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates," *Math. Comput. Simul.*, vol. 55, no. 1, pp. 271–280, 2001, the second IMACS seminar on Monte Carlo methods. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378475400002706

[14] H. Y. Lateef, A. Imran, M. A. Imran, L. Giupponi, and M. Dohler, "LTE-advanced self-organizing network conflicts and coordination algorithms," *IEEE Wireless Commun.*, vol. 22, no. 3, pp. 108–117, Jun. 2015.

[15] 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.

[16] S. Gambs, M.-O. Killijian, and M. N. del Prado Cortez, "Next place prediction using mobility markov chains," in *Proc. First Workshop Meas., Privacy, Mobility*, New York, NY, USA, 2012, pp. 3:1–3:6.

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

[18] 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 *Proc. IEEE Int. RF Microw. Conf.*, Dec. 2013, pp. 105–110.

[19] X. Zhou, Z. Zhao, R. Li, Y. Zhou, J. Palicot, and H. Zhang, "Human mobility patterns in cellular networks," *IEEE Commun. Lett.*, vol. 17, no. 10, pp. 1877–1880, Oct. 2013.

[20] J.-K. Lee and J. C. Hou, "Modeling steady-state and transient behaviors of user mobility: Formulation, analysis, and application," in *Proc. 7th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, New York, NY, USA, 2006, pp. 85–96.

[21] H. Abu-Ghazaleh and A. S. Alfa, "Application of mobility prediction in wireless networks using markov renewal theory," *IEEE Trans. Veh. Technol.*, vol. 59, no. 2, pp. 788–802, Feb. 2010.

[22] H. Farooq and A. Imran, "Spatiotemporal mobility prediction in proactive self-organizing cellular networks," *IEEE Commun. Lett.*, vol. 21, no. 2, pp. 370–373, Feb. 2017.

[23] Q. Lv, Y. Qiao, N. Ansari, J. Liu, and J. Yang, "Big data driven hidden markov model based individual mobility prediction at points of interest," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5204–5216, Jun. 2017.

[24] D. Zhang, D. Zhang, H. Xiong, L. T. Yang, and V. Gauthier, "Nextcell: Predicting location using social interplay from cell phone traces," *IEEE Trans. Comput.*, vol. 64, no. 2, pp. 452–463, Feb. 2015.

[25] A. Nadembega, A. Hafid, and T. Taleb, "A destination and mobility path prediction scheme for mobile networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 6, pp. 2577–2590, Jun. 2015.

[26] D. Zhang, M. Chen, M. Guizani, H. Xiong, and D. Zhang, "Mobility prediction in telecom cloud using mobile calls," *IEEE Wireless Commun.*, vol. 21, no. 1, pp. 26–32, Feb. 2014.

[27] Y. Lin, C. Huang-Fu, and N. Alrajeh, "Predicting human movement based on telecom's handoff in mobile networks," *IEEE Trans. Mobile Comput.*, vol. 12, no. 6, pp. 1236–1241, Jun. 2013.

[28] I. Schumm, "Lessons learned from Germany's 2001–2006 labor market reforms," Ph.D. dissertation, Dept. Econ., Univ. Wurzburg, Würzburg, Germany, 2009.

[29] G. Corradi, J. Janssen, and R. Manca, "Numerical treatment of homogeneous semi-markov processes in transient cases-a straightforward approach," *Methodol. Comput. Appl. Probability*, vol. 6, no. 2, pp. 233–246, 2004.

[30] H. Farooq, A. Asghar, and A. Imran, "Mobility prediction-based autonomous proactive energy saving (AURORA) framework for emerging ultra-dense networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 958–971, Dec. 2018.

[31] I. Viering, M. Dottling, and A. Lobinger, "A mathematical perspective of self-optimizing wireless networks," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2009, pp. 1–6.

[32] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Distributed α-optimal user association and cell load balancing in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 1, pp. 177–190, Feb. 2012.

[33] A. A. Solutions, HTXCW631819x000, 2019. [Online]. Available: http://66.201.95.79/amphenolantennas/product/htxcw631819x000/

[34] Powerwave, P45-17-XLH-RR, 2012. [Online]. Available: http://raycom-w.ru/files/import_pdf/P45-17-XLH-RR.ru.pdf

[35] 3GPP, "3rd Generation Partnership Project; Physical layer aspects for evolved universal terrestrial radio access (E-UTRA)," Sophia Antipolis, France, TR 25.814 V7.1.0 Release 7, 2006.

[36] "M.2135 Guidelines for evaluation of radio interface technologies for IMT-advanced," ITU Radiocommunications Sector, Geneva, Switzerland, Tech. Rep. ITU-R M.2135-1, Dec. 2009. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2135-1-2009-PDF-E.pdf

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

[38] K. Lee, S. Hong, S. J. Kim, I. Rhee, and S. Chong, "SLAW: A new mobility model for human walks," in *Proc. IEEE INFOCOM 28th Conf. Comput. Commun.*, Apr. 2009, pp. 855–863.

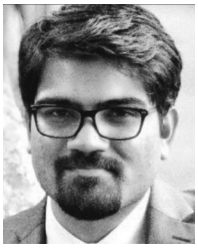
[39] "Five Trends to Small Cell 2020," Shenzhen, China, Huawei, and Barcelona, Spain, MWC, White Paper, 2016. [Online]. Available: http://www-file.huawei.com/-/media/CORPORATE/PDF/News/Five-Trends-To-Small-Cell-2020-en.pdf?la=en

[40] R. L. Haupt and S. E. Haupt, *Practical Genetic Algorithms*, vol. 2. New York, NY, USA: Wiley 1998.

[41] Y. Tang, N. Cheng, W. Wu, M. Wang, Y. Dai, and X. Shen, "Delay-minimization routing for heterogeneous vanets with machine learning based mobility prediction," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3967–3979, Apr. 2019.

[42] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A deep-learning-based radio resource assignment technique for 5G ultra dense networks," *IEEE Netw.*, vol. 32, no. 6, pp. 28–34, Nov. 2018.

Hasan Farooq (Member, IEEE) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2009, the M.Sc. (by Research) degree in information technology from Universiti Teknologi PETRONAS, Seri Iskandar, Malaysia, in 2014, and the Ph.D. degree in electrical and computer engineering from The University of Oklahoma, Norman, OK, USA, in 2018, while working in AI4Networks Research Center as a Graduate Research Assistant. His dissertation was awarded Gallogly College of Engineering Dissertation Excellence Award by The University of Oklahoma in 2018 and his dissertation's core idea won IEEE GREEN ICT Young Professional Award in 2017. He has authored or coauthored more than 30 publications in high impact journals, book chapters and proceedings of IEEE flagship conferences on communications. His research interests include 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 tools, and data analytics. He has been involved in multinational QSON Project on Self Organizing Cellular Networks (SON) as well as four NSF funded projects on 5G SON. He was the recipient of Internet Society (ISOC) First Time Fellowship Award towards Internet Engineering Task Force (IETF) 86th Meeting held in USA, in 2013.



Ahmad Asghar (Student Member, IEEE) received the B.Sc. degree in electronics engineering from the Ghulam Ishaq Khan Institute of Science and Technology, Khyber Pakhtunkhwa, Pakistan, in 2010, the M.Sc. degree in electrical engineering from the Lahore University of Management and Technology, Lahore, Pakistan, in 2014, and the Ph.D. degree in electrical and computer engineering from The University of Oklahoma, Norman, OK, USA, in 2018. His research interests include studies on self-healing and self-coordination of self-organizing functions in future big-data empowered cellular networks using analytical and machine learning tools.

Ali Imran (Senior Member, IEEE) is the Founding Director of AI4Networks Research Center, The University of Oklahoma, Norman, OK, USA, where he is leading several multinational and industry lead projects on AI for wireless networks. His research on AI enabled wireless networks has played seminal role in this emerging area and has been supported by $4 M in nationally and internationally competitive research funding, and recognized by several prestigious awards such as IEEE Green ICT Young Professional Award in 2017, VPR Outstanding International Impact Award at The University of Oklahoma, in 2017 and Best Paper Award IEEE CAMAD in 2013. He has authored or coauthored more than 100 refereed journal and conference papers and has several patents granted and pending on this topic. Since 2018, he has been named William H. Barkow Presidential Professor with The University of Oklahoma. He is routinely invited to serve as an Advisor to key stakeholder in cellular network eco-system and as a Speaker and a Panelist on international industrial fora and academic conferences on this topic. Before joining OU in January 2014, for three years he has worked as a Research Scientist with QMIC, Qatar. Between October 2007 and 2011, he worked with the Institute of Communications Systems (5GIC) University of Surrey, England, U.K. In that position, he has contributed to a number of pan-European and international research projects while working in close collaboration with key industrial players. He is an Associate Fellow of Higher Education Academy (AFHEA), U.K., the President of ComSoc Tulsa Chapter, a member of Advisory Board for Special Technical Community on Big Data at IEEE Computer Society, and Board Member of ITERA.