Concurrent Optimization of Coverage, Capacity, and Load Balance in HetNets Through Soft and Hard Cell Association Parameters

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

***Abstract*—Ultradense heterogeneous networks (HetNets) are emerging as an inevitable approach to tackle the capacity crunch in cellular networks. However, imbalanced load among small and macrocells and poor resource utilization as a consequence in HetNets remains a long-standing problem. This paper addresses this problem by presenting a solution for maximization of coverage and capacity while minimizing load imbalance among macro and small cells. Most recent studies on the topic focus on either optimization of coverage, capacity or load, or a combination of two of these three intertwined objectives. We formulate the optimization problem as a function of two hard parameters namely antenna tilt and transmit power, and a soft parameter, cell individual offset, that affect the coverage, capacity, and load directly. The resulting solution is a combination of the otherwise conflicting coverage and capacity optimization (CCO) and load balancing (LB) self-organizing network (SON) functions. In the presented joint CCO-LB solution, a conflict free operation of CCO and LB is ensured by designing a novel load aware user association methodology and resolving the effects of shadowing on coverage probability using stochastic approximation. The problem is proven to be nonconvex and is solved using genetic algorithm, sequential quadratic programming, and pattern search algorithms. The proposed CCO-LB solution is compared against two recently proposed CCO and CCO-LB solutions in the literature. Results show that the proposed solution can yield significant gain in terms of throughput, spectral efficiency, and load distribution.**

***Index Terms*—Heterogeneous networks, self-organizing networks, coverage, capacity, load balancing, joint optimization, 5G mobile cellular networks.**

I. Introduction

DESPITE recent advancements in many physical layer techniques and possible exploitation of new spectrum at higher frequencies, network densification remains the most yielding means to meet capacity demands of future 5G cellular networks. Densification, in one form or another, has also emerged as the most prolific defense against the energy and spectral efficiency challenges that plague modern cellular networks [1]–[3]. However, network densification is not without limitations itself. One of the biggest challenges facing dense heterogeneous networks (HetNets) is the imbalance of load between macro cells and small cells [4]–[10]. This load imbalance mainly stems from received power disparity between macro cell and small cells and causes poor utilization of system capacity.

A. Background and Motivation

State-of-the-art cellular systems use Reference Signal Received Power (RSRP) based user association mechanism where the cell with highest RSRP is the serving cell. Most academic studies on HetNets also build on the same user association method. The problem with this method is that it does not consider several factors that determine the overall performance of the network. These factors include load in the candidate cell, Signal to Interference and Noise Ratio (SINR) from the candidate cell, the effective load generated by the user to be associated, available free resources in the candidate cell, as well as the impact of new user association on interference and hence overall system capacity.

The problems caused by RSRP based user association become more pronounced in HetNets because compared to macro cells, small cells have much shorter range due to their low transmission power and shorter antenna heights. Thus, given a uniform user distribution, a small cell in dense HetNets is likely to attract much smaller number of users compared to macro cells. Full spectrum reuse between small cells and macro cells can lead to serious load imbalance, resource inefficiency and degradation in Quality of Experience (QoE). Cell individual offset (CIO) has been proposed and standardized by the 3GPP [11] to address this problem. A positive value of CIO artificially extends the range of a cell, thereby allowing additional users to be associated with a cell as long as the RSRP from that cell is smaller than the RSRP of the strongest neighbor only by CIO value or less.

However, several recent studies [12]–[15] show that CIO is not the panacea for the load imbalance and resultant resource inefficiency problem in HetNets either. The aftermath of CIO enabled association is illustrated in Figs. 1(a) and 1(b). In Fig. 1(a), small cells do not use any CIO, and thus have a marginal share of associated users compared to macro cells. In Fig. 1(b), small cells are given CIO of 10 dB each which gives them a range boost proportional to CIO, thus increasing their associated user share. However, note that users who have been shifted from macro

Manuscript received August 18, 2017; revised April 5, 2018; accepted May 13, 2018. Date of publication June 12, 2018; date of current version September 17, 2018. This work was supported by the National Science Foundation under Grants 1559483, 1619346, and 1730650. For more details about these projects, visit www.ai4networks.com. A portion of this work has been accepted for publication in IEEE PIMRC '17 [45]. The review of this paper was coordinated by Prof. C. Zhang. *(Corresponding author: Ahmad Asghar.)*

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

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVT.2018.2846655

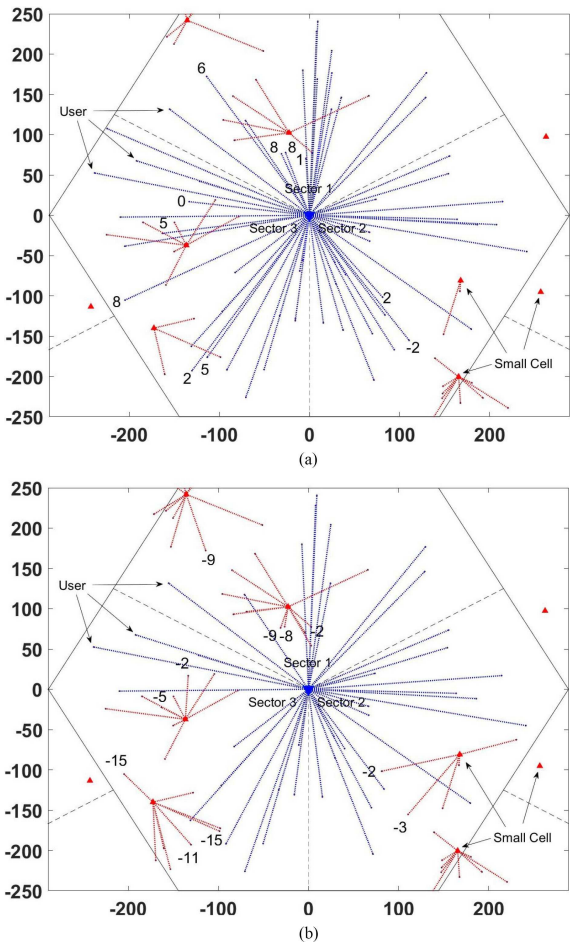


Fig. 1. RSRP-based user association.

TABLE I
SINR VALUES OF RE-ASSOCIATED USERS BEFORE AND AFTER CIO CHANGE

User ID	Pre-CIO SINR (dB)	Post-CIO SINR (dB)
1	1.68	-2.48
7	0.61	-2.39
20	5.63	-9.36
25	1.15	-1.73
56	1.86	-10.89
106	4.33	-4.55
110	7.62	-14.82
127	-2.48	-2.92
129	-0.2	-2.03
132	8.01	-8.21
136	8.3	-8.57
148	4.82	-15.35

cells to small cells due to CIO suffer a significant drop in SINR. The pre and post CIO change SINRs of users whose associations changed are compared in Table I. This example demonstrates that blanket use of empirically determined CIO values can affect overall resource efficiency in the system negatively, thereby causing the same problem that CIOs were introduced to solve in the first place. Instead, CIOs need to be determined through a method that considers user traffic demands and current cell loads. Most importantly CIO values should be determined in conjunction with two other key hard parameters that affect SINR as well as cell association i.e., transmit powers and antenna tilts.

B. Relevant Work

In commercial networks, to-date CIO values are set using ad-hoc methods. Some recent academic studies on LB that do consider CIO as a parameter of interest [12]–[15], provide heuristic solutions for finding CIO values to balance cell loads but fall short of assessing the impact of CIO on user QoE. These studies effectively verify the conclusion drawn from the above example that heuristic or ad-hoc settings of CIO can create more problems than they solve.

Additionally, an effective solution towards LB or CCO in emerging HetNets cannot simply focus on one or two parameters out of CIO, antenna tilt, and transmit powers, but must take into account the interplay among all three parameters. There exist some recent studies on CCO [16]–[18] that focus purely on SINR optimization with one or two of the aforementioned parameters. However, as demonstrated previously in Figs. 1(a) and 1(b) and further explained in Section III, in HetNets SINR alone is not always the suitable optimization criteria as it can contribute to overall performance degradation.

Compared to CCO, LB has been studied more extensively, e.g., see the survey on LB and the references therein in [19]. However, most of these studies on LB consider hard parameters only i.e., either tilt, and/or transmit power for optimization and do not include CIO (see Fig. 5, in [19]). Tilt and transmit power based LB may work for macro cell only networks, but these two parameters alone cannot offset the imbalance of cell load between macro and small cells in HetNets. Only a handful of studies consider CIO as an optimization parameter for LB [7], [8], [12], [13], [20] alone or with at most one other parameter.

While CCO [6], [16], [17], [21] and LB [7], [8], [12]–[15], [22]–[26] problems have been addressed individually and extensively in literature, the co-design of CCO and LB, the only practically viable way to implement both CCO and LB in a HetNet concurrently, has received limited attention. Co-design of other SON functions such as Mobility Robustness Optimization and Mobility Load Balancing [27], and LB and Handover Optimization [28], [29] has been studied from various perspectives. However, co-design of CCO and LB is particularly challenging because of their parametric overlap as explained in [30] as well as objective conflict between the two SON functions as expatiated in [31]. As a result, most LB solutions presented in literature often balance load at the cost of CCO, and vice versa. An effective solution has to be a judicious combination of both CCO and LB.

The most relevant study to our work is presented in [32]. This study proposes joint optimization of CCO and LB by minimizing the log sum of cell loads in the network while maintaining minimum coverage threshold via a constraint. The solution makes use of only two of the three parameters of interest i.e., antenna tilts and CIOs. A heuristic algorithm is used to first optimize antenna tilts for load balancing. The cell partitions thus obtained are further refined using CIOs since the authors argue that accurate cell coverage mapping with antenna tilts is not possible in real world scenarios. Contrary to the approach presented in [32], our proposed CCO-LB solution leverages transmit powers, antenna tilts and CIOs within one formulation. Our approach is fundamentally different than that in [32] in the sense that instead of focusing on load minimization, our objective function

is focused on throughput maximization, but embeds LB into the optimization problem through a built-in load fairness measure among cells as well as through introduction of a novel load aware cell association mechanism. A comparative analysis of our proposed solution with the solution in [32] is presented in Section V.

C. Proposed Approach and Contributions

In 3GPP Release 9 of Radio Access Network specifications [33], CCO SON function is designed to ensure automatic service and coverage reliability. to network subscribers whereas LB SON function is aimed at minimizing cell congestion in the network. The ideal goals of CCO SON function can be given as:

- Minimum received downlink power $P^c_{r,u}$ for a predefined percentage ϖ of users should meet or exceed the minimum coverage threshold i.e, RSRP value P^c_{th};
- Each user *u* is served with a data rate that should meet or exceed a predefined data rate $\hat{\tau}_u$ for that user or class of users.

On the other hand, the primary goal of LB SON function can be given as:

- Load distribution among cells should remain such that no cell becomes congested as long as there are cells with free radio resources in the neighborhood of that cell.

CCO and LB SON functions, if designed and deployed correctly, have the capability to enhance the QoE and resource efficiency in HetNets tremendously. However, designing a CCO solution that can work in-tandem with LB and vice versa, remains an open problem so far. The difficulty in designing a combined CCO-LB solution stems from the following facts:

1) CCO and LB SON functions have different objectives but leverage the same optimization parameters including transmit powers, CIOs and antenna tilts;
2) CCO and LB impact user QoE in two distinct and conflicting ways [30], [31], [34];
3) In HetNets relationships between load, capacity and SINR become intertwined, as highlighted in several recent studies [4], [8], [32], [35]. The load in a cell for given traffic demand depends on SINR perceived by the users associated with that cell. On the other hand, SINR of a user is affected by the load of neighboring cells;
4) While both CCO and LB can leverage CIO, CIO boost set to balance load can result in poorer SINR as shown in Figs. 1(a) and (b) leading to the previously mentioned problem.

Our proposed approach tackles these challenges by embedding the goals of CCO and LB into a single objective function by introducing a load aware user association method and by jointly optimizing soft parameter CIO and hard parameters tilt and transmit power. The contributions of this paper can be summarized as follows:

1) Modeling and Analysis: We formulate two versions of the optimization problem both of which capture the goals of both CCO and LB SON functions in terms of antenna tilt, transmit power and CIO to reflect the cases of known and unknown user traffic demand. We resolve the uncertainty in user coverage, and consequently the coverage constraint of CCO, due to shadowing by employing stochastic approximation to transform the coverage probability constraint into a deterministic coverage constraint. We analyze the convexity of our objective function and show that the problem is a non-convex large scale NP-hard problem. However, since the objective function in our formulation provides a quickly evaluable quantitative measure of the impact of optimization parameters on network performance, we demonstrate that techniques to solve large scale problems such as genetic algorithm, sequential quadratic programming and pattern search can be employed to effectively solve the problem.

2) A New Cell Association Methodology: We propose and evaluate a novel user association technique that incorporates cell load into the user association decision in addition to RSRP. While the proposed user association scheme is mainly intended for emerging 5G HetNet deployments, we also present a methodology to implement this scheme in legacy cellular networks such as LTE without requiring any change to the standard. The proposed load-aware user association scheme also offers a mechanism to set the priority level between CCO and LB at cell level or in a centralized fashion as per operator's policy. We also compare our proposed load-aware user association scheme against state of the art Max RSRP and Max SINR user association schemes.

3) System Level Performance Analysis and Benchmarking: We use multi-tier system level simulations to conduct a comprehensive performance analysis of proposed joint CCO-LB solution in realistic HetNet settings using 3GPP compliant simulation parameters. We compare the results of our solution with the current industrial practice of using fixed parameter settings, and with the two most relevant studies in [6] and [32] respectively that present solutions for CCO and CCO-LB respectively. Our comparative analysis investigates performance in terms of a range of key performance indicators (KPIs) that includes network loading, user throughput, SINR and spectral efficiency.

4) New Insights for HetNet Design and Standardization in 5G: The analysis and results presented in this paper also provide the following design insights for radio efficiency improvement in legacy networks and standardization in emerging 5G based HetNets:

1) Joint optimization of antenna tilts, transmit powers and CIOs yields better performance than optimization of individual parameter;
2) State-of-the-art user association methodology needs an evolution beyond RSRP(+CIO) based user association to include new factors such as cell loads, amplifier operating point (for energy efficiency considerations), expected traffic of incumbent user, mobility pattern estimations etc.;
3) There is a need for paradigm shift from SINR focused network parameter optimization since SINR optimization in HetNets becomes almost meaningless in the face of imbalanced cell loads;
4) CIO can be used for more than just biasing RSRP. Our results suggest that CIO can be modulated with information about the residual capacity in the cell in dynamic fashion to implement the proposed new load aware user association methodology. This would allow the proposed load-aware user association to be implemented without requiring any change in the current standard.

The paper is organized as follows: Section II presents the system model used for the joint formulation of CCO-LB SON function problem, Section III provides the problem formulation, Section IV presents the solution methodologies used to solve the joint CCO-LB SON function problem, and Section V presents the results of proposed solution as well as comparison with the solutions in [6] and [32].

TABLE II
KEY SYMBOL DEFINITIONS

Sym.	Definition	Sym.	Definition
η_u^c	PRB allocated to user u at cell c	$\mathbb{C}$	Set of all cells
ω_B	Bandwidth per PRB	$\mathbb{U}$	Set of all active and idle users
$\hat{\tau}_u$	Desired user throughput	$\mathbb{U}$	Set of all active users
γ_u^c	SINR of user u at cell c	$\hat{\mathbb{U}}$	Set of all active satisfied users
N_b^c	Total PRBs at cell c	κ	Thermal noise
δ_u^c	Shadowing of user u from cell c	$P_{r,u}^c$	Downlink Rx power to user u from cell c
a	Pathloss constant	ϖ	Network coverage threshold
β	Pathloss exponent	P_{th}^c	Downlink Rx power threshold
d_u^c	Distance of user u from cell c	η_{lth}^c	Cell load threshold
P_c^t	Tx power of cell c	α	User association exponent
μ	Antenna gain constant	Ω	Objective value for CLASS solutions
ψ_u^c	Vertical angle between user u and cell c	ψ_{tilt}^c	Antenna tilt of cell c

II. SYSTEM MODEL

In this Section, we describe the system model employed in the formulation of the joint CCO-LB SON function and the underlying assumptions. Furthermore, Table II provides the list of key symbols used in the problem formulation.

A. Network and User Specifications

For formulating the joint CCO-LB problem, we consider a network of hexagonal macro base stations with at least one randomly deployed small cell in the coverage area of each macro cell. 100% frequency reuse is considered between macro and small cells. Macro cells use directional antennas while small cells employ omni-directional antennas. An Orthogonal Frequency Division Multiple Access (OFDMA) based system with resources divided into physical resource blocks (PRBs) of fixed bandwidth, is assumed. For conciseness, the downlink direction is chosen for the analysis as this is where most imbalance in coverage of macro and small cells occurs. It is assumed that users in the network are stationary. It is further assumed that requested user data rate is known which gives a lower-bound on the desired instantaneous user throughput. Desired user throughput can be modeled as a spatio-temporal function of subscriber behavior, subscription level, service request patterns, as well as the applications being used with the help of big data analytics as recently proposed in [36]. Our formulation is not dependent on particular scheduling techniques.

B. Parameters and Measurements

1) Cell Loads: We can define instantaneous cell load as the ratio of PRBs occupied in a cell during a Transmission Time Interval and total PRBs available in the cell. This information is available as a standard measurement from 3GPP as "UL/DL total PRB usage" [33] and can be broadcast to the users. To define cell load η_c for our system model, we first calculate minimum number of PRBs η_u^c to be allocated to a user:

$$\eta_u^c = \frac{1}{\omega_B}\left(\frac{\hat{\tau}_u}{f(\gamma_u^c)}\right) \tag{1}$$

where $\hat{\tau}_u$ represents the (desired) throughput of user $u \in \mathbb{U}_c$, where $\mathbb{U}_c$ is the set of all active users associated with cell c. γ_u^c represents the SINR of user *u* when associated with cell *c* and ω_B is the bandwidth per PRB. $f(\gamma_u^c)$ denotes the spectral efficiency of the user link for given SINR. If we consider features such as MIMO or coding scheme gains and scheduling gains, $f(\gamma_u^c)$ can be defined as $f(\gamma_u^c) := A\log_2(1 + B\gamma_u^c)$, where *A* and *B* are constants that can capture throughput gains (per PRB) achievable from various types of diversity schemes, or losses incurred by signaling overheads and hardware inefficiencies. For the sake of simplicity and without loss of generality, we assume $A = B = 1$. Thus, we can define residual cell capacity and cell load as:

$$\text{Residual Capacity} = \Lambda_c = N_b^c - \frac{1}{\omega_B}\left(\sum_{\mathbb{U}_c}\frac{\hat{\tau}_u}{\log_2(1+\gamma_u^c)}\right) \tag{2}$$

$$\text{Cell Load} = \eta_c = \frac{1}{N_b^c}\left(\frac{1}{\omega_B}\left(\sum_{\mathbb{U}_c}\frac{\hat{\tau}_u}{\log_2(1+\gamma_u^c)}\right)\right) \tag{3}$$

where N_b^c is the total PRBs at cell c. Consequently, the range of cell load is $\eta_c \in [0, \infty)$. If the cell load exceeds 1, the cell in reality will be fully loaded and incoming users will be blocked. The value of load η_c is therefore referred to as virtual load and $\eta_c > 1$ reflects congestion in cell *c*.

2) Received Power: In LTE networks, downlink RSRP from nearby base stations is continuously monitored by the users and reported to the serving cell for a number of purposes. In our proposed CCO-LB approach we use the RSRP to calculate coverage probability in the network.

3) Cell Individual Offset: CIO can be defined as a combination of multiple cell association parameters introduced by the 3GPP [11] including cell hysteresis, cell offsets and event related offsets which are used to decide user association. CIO information is by each cell and decoded by the users as part of standard operation. For the purpose of this paper we treat CIO as a simple virtual boost in RSRP.

III. PROBLEM FORMULATION

To incorporate QoE into the joint CCO-LB optimization, we choose to formulate our problem as a per cell per user throughput optimization problem. The first step towards this goal is to build a SINR model as function of all three optimization parameters under consideration.

A. User SINR as Function of Tilt, Transmit Power and CIO

Downlink SINR $\hat{\gamma_u^c}$ of a reference signal at user location *u* when associated with cell *c* can be expressed as the ratio of RSRP $P_{r,u}^c$ measured by user *u* from cell *c* to the sum of RSRP measured by user *u* from all interfering cells *i* such that $\forall i \in \mathbb{C}/c$, and the noise power κ:

$$\hat{\gamma_u^c} = \frac{P_t^c G_u G_u^c \delta_u^c a\,(d_u^c)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} P_t^i G_u G_u^i \delta_u^i a\,(d_u^i)^{-\beta}} \tag{4}$$

where P_t^c and P_t^i are the transmit powers of serving cell *c* and interfering cell *i*, G_u is the gain of user equipment, G_u^c and G_u^i are the gains of transmitter antenna of the cells *c* and *i* towards user *u*, δ_u^c and δ_u^i is the shadowing observed at the location of user u from serving cell c and interfering cell i, a is the pathloss constant, d_u^c and d_u^i represent distance of user *u* from cell *c* and *i*, and β is the pathloss exponent. The numerator in (4) is obtained from the standard exponential pathloss model while δ_u^c and δ_u^i in equation (4) can be modeled as random variables with either Gaussian or log-normal distribution varying over both space and time.

The expression in (4) is only useful when estimating the quality of reference signals which are always being transmitted by all the cells. Thus, $\hat{\gamma_u^c}$ is not a true measure of SINR on the PRBs where interference generated is dependent on utilization of that same PRB in other cells at the same time. We assume user arrival in the system follows a general distribution, thus the exact interference becomes a function of time. Therefore, to obtain an SINR estimate independent of time, a reasonable low complexity substitute for average downlink interference from a cell *i* is to use the ratio of occupied PRBs in the cell. The expression for SINR estimate for user *u* in cell *c* can then be given as:

$$\gamma_u^c = \frac{P_t^c G_u G_u^c \delta_u^c a \left(d_u^c\right)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \hat{\eta}_i P_t^i G_u G_u^i \delta_u^i a \left(d_u^i\right)^{-\beta}} \tag{5}$$

where $\hat{\eta}_i$ denotes actual cell load in a cell, that for a cell i can be obtained by modifying (3) as:

$$\hat{\eta}_i = \frac{1}{N_b^i}\left(\frac{1}{\omega_B}\left(\sum_{\hat{\mathbb{U}}_i} \frac{\hat{\tau}_u}{\log_2\left(1+\gamma_u^i\right)}\right)\right) \tag{6}$$

where $\hat{\mathbb{U}}_c \subseteq \mathbb{U}_c \subseteq \mathbb{U}$ is the set of all active user associated with cell *c*. Here $\mathbb{U}$ represents the complete set of users in the network and the difference set $\mathbb{U}_c - \hat{\mathbb{U}}_c$ represents users who requested but were denied resources by the cell *c* due to congestion which implies $\hat{\eta}_c \in [0,1]$. Note that in SINR expression (5) above, we do not use the virtual cell load from (3), but the actual cell load which can never exceed 1.

As macro cells in the system under consideration use directional antennas, using the expression for 3D antenna gain from [37], the gain from base station to user G_u^c can be given as:

$$G_u^c = 10^{-1.2\left(\lambda_v\left(\frac{\psi_u^c - \psi_{tilt}^c}{B_v}\right)^2 + \lambda_h\left(\frac{\phi_u^c - \phi_{azi}^c}{B_h}\right)^2\right)} \tag{7}$$

where λ_h and λ_v are the weights of horizontal and vertical beam patterns of the antenna, ψ_u^c is the vertical angle between user *c* and the antenna of cell *c*, ψ_{tilt}^c is the tilt angle of serving cell antenna, ϕ_u^c is the horizontal angle of user *u* from cell *c*, ϕ_{azi}^c is the azimuth of antenna of cell *c*, and B_h and B_v are horizontal and vertical beam widths of the transmitter antenna of cell *c*. As our variable of interest in (7) is tilt angle and the rest of the antenna parameters can be treated as constants, for the sake of conciseness we can simplify (7) using the following substitution:

$$x_u^c = \frac{(B_v)^2 \lambda_h}{\lambda_v}\left(\frac{\phi_u^c - \phi_{azi}^c}{B_h}\right)^2 \tag{8}$$

and re-write the SINR expression from (5) as:

$$\gamma_u^c = \frac{P_t^c G_u 10^{\mu\left(\left(\psi_u^c - \psi_{tilt}^c\right)^2 + x_u^c\right)} \delta_u^c a \left(d_u^c\right)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \hat{\eta}_i P_t^i G_u 10^{\mu\left(\left(\psi_u^i - \psi_{tilt}^i\right)^2 + x_u^i\right)} \delta_u^i a \left(d_u^i\right)^{-\beta}} \tag{9}$$

where μ is consolidated constant based on fixed antenna characteristics.

Finally, we address CIO in the SINR expression. This offset parameter is used for cell association as:

$$P_{r,u_{\mathrm{dBm}}}^c = \acute{P}_{r,u_{\mathrm{dBm}}}^c - P_{CIO_{\mathrm{dB}}}^c \tag{10}$$

where $P_{r,u_{\mathrm{dBm}}}^c$ is the true signal power in dBm received by user *u* from cell *c* and $\acute{P}_{r,u_{\mathrm{dBm}}}^c$ is the received power reported back by user *u* to cell *c* in dBm. This value includes $P_{CIO_{\mathrm{dB}}}^c$ (the CIO value of cell *c* in dB) which is then subtracted by the cell to retrieve $P_{r,u_{\mathrm{dBm}}}^c$.

The motivation behind introduction of CIO was to allow load balancing among cells. However, as described previously in Section I, if CIO has to be invoked to alter natural RSRP based cell association for the user, the SINR for that user is bound to be lower (see Figs. 1(a) and 1(b)). Nevertheless, CIO is a necessary means to balance cell loads while capacity loss due to drop in SINR can partially be offset if the cell association takes into account cell load in addition to RSRP.

B. An Improved Load-Aware User Association Mechanism

The state-of-the-art method of determining user associations $\mathbb{U}_c$ is to use the RSRP measurements along with CIO values as given in (10). However, this method overlooks the key role of user association in overall capacity and QoS through cell load and SINR distributions. To overcome this challenge, we propose to establish user association with cell *j* not only based on received power but also load in that cell. More specifically, this load-aware user association with cell *j* can be determined as:

$$\mathbb{U}_j := \left\{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}}\left(\left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}_{r,u_{\mathrm{dBm}}}^c\right)^{(1-\alpha)}\right)\right\} \tag{11}$$

where $\mathbb{U}_j$ is a set of all users for whom (a scaled version of) the product of the RSRP(+CIO) in Watts $\acute{P}_{r,u}^c$ and the normalized residual cell capacity is maximized for cell *j*. $\alpha \in [0,1]$ is a weighting factor introduced to allow trading between the impact of RSRP and cell load measurements in the user association. As established in (3), cell load is dependent on the SINR of users in the cell i.e., better the SINR of users in candidate cell, lesser the load in the cell for given traffic demand. Note that in (11), to make new user association decision with a cell we use the virtual load and not the actual load. While, using actual cell load that has range $\hat{\eta}_c \in [0,1]$ can indicate the current load in a cell, it cannot help take into account the users that are already associated with that cell but were not served. On the other hand, virtual cell load as defined in (3) with range $\eta_c \in [0,\infty)$, provides a truer picture of effective potential load in the candidate cell.

The expression in (11) gives the set $\mathbb{U}_j$ of users to be associated with the cell *j* and thus represents both active and idle users. On the other hand, the set $\mathbb{U}_c$ used in the expression for SINR in

(9) represents the set of only active users associated with the cell c. With $\alpha = 1$, the user association simply becomes a function of cell load and SINR at the time of association. Consequently this cell association espouses the LB SON function only. On the other hand, if $\alpha = 0$, the proposed user association method simply represents state-of-the-art RSRP based cell association method which helps achieve coverage optimization aspect in the CCO SON function. Determining the optimal value of weighting factor α is an optimization problem worth investigating in itself. In Section V, we evaluate KPIs with a range of α and discover interesting trends that can be used to develop some practical design guidelines.

C. Problem Statement

A common approach towards throughput maximization in LB or CCO is to use a problem formulation that maximizes the mean throughput per user per cell. However, if we try to maximize the arithmetic mean of user throughput determined by SINR expression derived above, users with no throughput and cells with no load will be equally acceptable as users with very high throughputs and cells with full loads. While such formulation will achieve the objectives of CCO, it will not perform load balancing, and hence cannot be suitable approach for joint CCO-LB. To simultaneously reflect the goals of both CCO and LB in a single objective function, we propose the objective function to be modeled as:

$$\max_{\boldsymbol{P_t^c},\boldsymbol{\psi_{tilt}^c},\boldsymbol{P_{CIO}^c}} \left(\prod_{\mathbb{C}} \left(\prod_{\mathbb{U}_c} \omega_u^c \log_2\left(1+\gamma_c^u\right) \right)^{\frac{1}{|\mathbb{U}_c|}} \right)^{\frac{1}{|\mathbb{C}|}} \tag{12}$$

The outer geometric mean in this formulation dampens the load disparity among cells, and thus integrates LB goal into the optimization objective. This formulation is intended for scenarios where user required rates are not known or predicted. Thus, use of inner geometric mean instead of arithmetic mean for user throughput protects users with lower SINR from being unfairly treated, while maximizing the overall throughput.

If, however, the desired user throughput is already known or can be predicted, for example using the framework presented in [36], we can adopt a more greedy approach by replacing the inner geometric mean with arithmetic mean as it is bound to provide an improved or equivalent result [38]. The new objective function with this assumption is given as:

$$\max_{\boldsymbol{P_t^c},\boldsymbol{\psi_{tilt}^c},\boldsymbol{P_{CIO}^c}} \left(\prod_{\mathbb{C}} \left(\frac{\sum_{\mathbb{U}_c} \omega_u^c \log_2\left(1+\gamma_c^u\right)}{|\mathbb{U}_c|} \right) \right)^{\frac{1}{|\mathbb{C}|}} \tag{13}$$

A comparison between performance of both formulations is presented in Section V. The formulations in (12) and (13) inherit two basic constraints to achieve full objectives of CCO and LB SON function i.e.:

i) The ratio of covered users C must meet or exceed the minimum network coverage threshold ϖ i.e. $C \geqslant \varpi$ where C is dependent on the number of users satisfying the equation $P_{r,u}^c \geqslant P_{th}^c$;
ii) Cell load, as defined in (3), for every cell has to be less than or equal to the cell load thresholds set by operator policies: $\eta_c \leqslant \eta_{th}^c \forall c \in \mathbb{C}$
We introduce an additional constraint in the formulation to avoid blocking any users i.e.:
iii) The set of served active users $\hat{\mathbb{U}}_c$ by cell c must be equal to the total set of active users $\mathbb{U}_u$ associated with the cell c: $\hat{\mathbb{U}}_c = \mathbb{U}_c$.

The satisfaction of constraint (i) depends heavily on the pathloss model employed in (4). Despite the assumption that user location remains the same over time, random variations in shadowing δ_u^c over space introduce uncertainty into the determination of $P_{r,u}^c$. Consequently C becomes a function of the distribution of δ_u^c such that constraint (i) becomes $Pr(C(\delta_u^c)) \geqslant \varpi$. This also implies that the evaluation of $P_{r,u}^c \geqslant P_{th}^c$ is a probabilistic problem rather than a deterministic one which can make

$$\max_{\boldsymbol{P_t^c},\boldsymbol{\psi_{tilt}^c},\boldsymbol{P_{CIO}^c}} \Omega = \tag{14}$$

$$\max_{\boldsymbol{P_t^c},\boldsymbol{\psi_{tilt}^c},\boldsymbol{P_{CIO}^c}} \left(\prod_{\mathbb{C}} \left(\prod_{\mathbb{U}_c} \omega_u^c \log_2 \left(1 + \frac{P_t^c G_u 10^{\mu\left(\left(\psi_u^c - \psi_{tilt}^c\right)^2 + x_u^c\right)} \delta_u^c a\left(d_u^c\right)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \hat{\eta}_i P_t^i G_u 10^{\mu\left(\left(\psi_u^i - \psi_{tilt}^i\right)^2 + x_u^i\right)} \delta_u^i a\left(d_u^i\right)^{-\beta}} \right) \right)^{\frac{1}{|\mathbb{U}_c|}} \right)^{\frac{1}{|\mathbb{C}|}} \tag{14a}$$

OR

$$\max_{\boldsymbol{P_t^c},\boldsymbol{\psi_{tilt}^c},\boldsymbol{P_{CIO}^c}} \left(\prod_{\mathbb{C}} \left(\frac{\sum_{\mathbb{U}_c} \omega_u^c \log_2 \left(1 + \frac{P_t^c G_u 10^{\mu\left(\left(\psi_u^c - \psi_{tilt}^c\right)^2 + x_u^c\right)} \delta_u^c a\left(d_u^c\right)^{-\beta}}{\kappa + \sum_{\forall i \in \mathbb{C}/c} \hat{\eta}_i P_t^i G_u 10^{\mu\left(\left(\psi_u^i - \psi_{tilt}^i\right)^2 + x_u^i\right)} \delta_u^i a\left(d_u^i\right)^{-\beta}} \right)}{|\mathbb{U}_c|} \right) \right)^{\frac{1}{|\mathbb{C}|}} \tag{14b}$$

$$\textbf{subject to} = \begin{cases} \frac{1}{|\mathbb{C}|} \sum_{\mathbb{C}} \frac{1}{|\mathbb{U}_c|} \sum_{\mathbb{U}_c} 1\left(P_{r,u}^c \geqslant P_{th}^c\right) \geqslant \varpi, \\ \eta_c \leqslant \eta_{th}^c \forall c \in \mathbb{C} \\ \hat{\mathbb{U}}_c = \mathbb{U}_c \end{cases} \tag{14c}$$

$$\mathbb{U}_j := \left\{ \forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}} \left(\left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}_{r,u_{\mathrm{dBm}}}^c \right)^{(1-\alpha)} \right) \right\} \tag{14d}$$

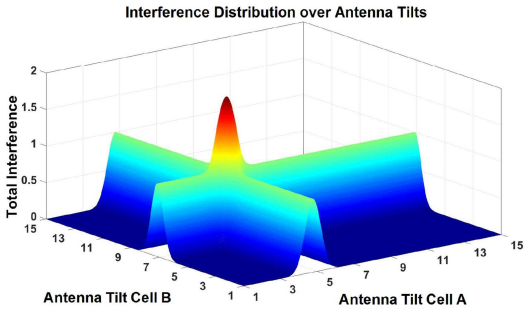


Fig. 2. Interference distribution over antenna tilts of interferers.

the overall problem intractable. In order to overcome this issue, we propose to reformulate constraint (i) such that it becomes deterministic.

Proposition 1: For Gaussian distributed shadowing δ_u^c, the probable coverage ratio $Pr(C(\delta_u^c))$ can be estimated using the transformation $\frac{1}{|\mathbb{C}|}\sum_{\mathbb{C}}\frac{1}{|\mathbb{U}_c|}\sum_{\mathbb{U}_c} 1\left(P_{r,u}^c \geqslant P_{th}^c\right)$.

Proof: The complete proof of proposition 1 is provided in Appendix A. ■

Substituting the expression for SINR from (9) in (12) gives the fair joint CCO-LB formulation given in (14a), while substituting SINR from (9) in (13) gives the greedy joint CCO-LB formulation given in (14b). Combining the two formulations with the above problem constraint and user association expression in (11) gives the final formulation in (14), (14a)–(14d), shown at the bottom of previous page.

Algorithm 1: Objective Function (14) Implementation Routine.

Input: $\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}$
Output: $\Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right)$ (14a) or (14b)
for $u \in \mathbb{U}$ **do**
  Find serving cell $j = \arg\max_{\forall c \in \mathbb{C}}\left(\acute{P}_{r,u_{\mathrm{dBm}}}^c\right)$
  Calculate SINR $\hat{\gamma}_u^j$ and η_u^j
end for
for $c \in \mathbb{C}$ **do** Calculate cell load η_c
end for
for $u \in \mathbb{U}$ **do**
  Find new serving cell $j =$ $\arg\max_{\forall c \in \mathbb{C}}\left(\left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}_{r,u_{\mathrm{dBm}}}^c\right)^{(1-\alpha)}\right)$
  Find updated SINR γ_u^j and η_u^j
end for
for $c \in \mathbb{C}$ **do** Calculate cell load η_c
end for
if $\frac{1}{|\mathbb{C}|}\sum_{\mathbb{C}}\frac{1}{|\mathbb{U}_c|}\sum_{\mathbb{U}_c} 1\left(P_{r,u}^c \geqslant P_{th}^c\right) \geqslant \varpi$ **then**
  if $\eta_c \leqslant \eta_{th}^c \forall c \in \mathbb{C}$ **then**
    if $\hat{\mathbb{U}}_c = \mathbb{U}_c$ **then**
      Calculate $\Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right)$
    end if
  end if
else
  $\Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) = -\infty$
end if

IV. Solution Methodology

In this Section, we first analyze the convexity of the joint **C**CO **L**B user **As**sociation aware **S**ON function (**CLASS**) presented in (14) and then present methodologies to implement it.

A. Convexity Analysis

Assuming we have a network of macro cells only, we can define the range of transmission powers $P_t^c \in [20\text{ W}, 40\text{ W}]$, antenna tilts as $\psi_{tilt}^c \in 90° + [0°, 15°]$ and CIOs as $P_{CIO}^c \in [0\text{ dB}, 10\text{ dB}]$. Affine sets are convex sets, therefore, the first requirement for convexity for problem (14) i.e., the constraints should be convex, is fulfilled. We know that geometric and arithmetic means preserve convexity of a function. We also know that the logarithmic function is also a convex function over the interval $(0, \infty)$. This leaves, the SINR expression in (9) to be examined to see if the formulation in (14) is convex or not.

Proposition 2: SINR as a function of antenna tilts as given in (9), is a non-convex function.

Proof: Fig. 2 plots the interference (denominator of (9)) as function of antenna tilts of two neighboring cells. Clearly it is not a convex function implying proposition 2. A more formal proof of proposition 2 is provided in Appendix B ■.

B. Alternate Solution Methodologies

Given the non-convexity and large scale of the problem, we must resort to heuristic approaches that can find optimal or near optimal solution of the formulation in (14).

Algorithm to Implement Proposed Cell Association: Before delving into possible non-convex optimization techniques to solve (14), an algorithm to practically implement the proposed user associations for given values of the three optimization parameters and obtain the updated value of objective function with new user associations is presented in Algorithm 1. This routine has to be called at each iteration of the heuristic optimization techniques to be discussed in the following.

1) Sequential Quadratic Programming (SQP): One way to solve non-convex problems of the type (14) that have linear constraints is to approximate it piece-wise with a convex quadratic function and then use convex optimization to solve it, a method also known as sequential quadratic programming. To leverage SQP we can re-write the problem in (14) as:

$$\max_{\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}} -\Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right)$$

$$\textbf{subject to}$$

$$= \begin{cases} W\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) \\ \quad := \varpi - \frac{1}{|\mathbb{C}|}\sum_{\mathbb{C}}\frac{1}{|\mathbb{U}_c|}\sum_{\mathbb{U}_c} 1\left(P_{r,u}^c \geqslant P_{th}^c\right) \geqslant 0, \\ X\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) := \eta_c - \eta_{th}^c \leqslant 0 \forall c \in \mathbb{C} \\ Y\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) := \hat{\mathbb{U}}_c - \mathbb{U}_c = 0 \end{cases}$$

$$\mathbb{U}_j := \left\{\forall u \in \mathbb{U} \mid j = \arg\max_{\forall c \in \mathbb{C}}\left(\left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}_{r,u_{\mathrm{dBm}}}^c\right)^{(1-\alpha)}\right)\right\} \quad (15)$$

As compared to unconstrained problem or problem with inequality constraint, equality constraints can reduce the search space of optimization problem significantly. We express user

association as an equality constraint such that for $u \in \mathbb{U}_c$

$$Z\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) := \sum_{i \in \mathbb{C}/c} 1\left(\left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}^c_{r,u_{\mathrm{dBm}}}\right)^{(1-\alpha)} \geqslant \left(\frac{1}{\eta_c}\right)^{\alpha} * \left(\acute{P}^c_{r,u_{\mathrm{dBm}}}\right)^{(1-\alpha)}\right) - |\mathbb{C}| + 1 = 0. \tag{16}$$

The expression in (16), where $1(.)$ is the indicator function, means that for a user u to be associated with cell c, the association function of the user with that cell must be greater than all the other cells. Lagrangian of (15) can be given as:

$$\begin{aligned}
&\mathcal{L}\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}, \lambda^1, \boldsymbol{\lambda^2}, \boldsymbol{\lambda^3}, \boldsymbol{\lambda^4}, \boldsymbol{\lambda^5}, \boldsymbol{\lambda^6}, \boldsymbol{\lambda^7}\right) \\
&= \Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) - \lambda^1 W\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) \\
&- \sum_{c \in \mathbb{C}} \lambda_c^2 X\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) - \sum_{c \in \mathbb{C}} \lambda_c^3 Y\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) \\
&- \sum_{u \in \mathbb{U}} \lambda_u^4 Z\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) - \sum_{c \in \mathbb{C}} \lambda_c^5 \left(P_t^c - P_{t,\min}^c\right) \\
&- \sum_{c \in \mathbb{C}} \lambda_c^6 \left(\psi_{tilt}^c - 90\right) - \sum_{c \in \mathbb{C}} \lambda_c^7 \left(P_{CIO}^c\right)
\end{aligned} \tag{17}$$

where $\boldsymbol{\lambda^x}$ represents the x-th vector of Lagrangian multipliers for the constraints in (15) and (16). Thus, the quadratic sub-problem to be solved at each iteration of SQP is given by (18) shown at the bottom of this page, where $\hat{\boldsymbol{H}}$ represents the approximate Hermitian matrix, which is updated at each iteration using the Broyden-Fletcher-Goldfarb-Shanno approximation method [39].

2) Other Heuristic Techniques: Through results presented in Section V, we found that SQP returns an acceptable solution with low number of iterations in most instances at the cost of a lack of guarantee that the solution is optimal due to the large dimensions of the problem in (14). Furthermore, the enormous search space size of (14) makes validation of the results produced through brute force almost impossible. Therefore, we tried a number of heuristic techniques that are known to converge to optimal solutions given enough iterations. In the following, we discuss two heuristics which yielded most promising results for this problem.

a) Genetic Algorithms: Genetic algorithms are known to be one of the most suitable heuristic algorithms available for solving complex combinatorial problems of kind of (14). It is important to note that the genetic algorithm starts from a random parameter set in the solution space, therefore, for each run, the time to find the feasible space is different. However, once found, the algorithm can quickly move towards the optimal solution in the feasible space. Algorithm 2 represents the pseudo code for the genetic algorithm used to solve (14).

Algorithm 2: Genetic Algorithm for CLASS Implementation.

Input:
Algorithm 1 to solve (14)
Parameter set space $\mathbf{S}(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c})$,
Maximum iterations **G**,
Solution space samples per iteration **P**,
Key samples per iteration **E**,
Mutation ratio **M**.
Output:
Solution $\mathbf{X} = [\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}]$

```
Generate |P| parameter sets from S randomly;
Generate values of Ω for each set in P
Create an empty set Pop and save the sets from P in it;
for i = 1 to G do
    Number of elite members in Pop num_elite = E;
    Select the best num_elite sets in Pop in terms of the
    value of Ω and save them in Pop_1;
    Number of crossover solutions num_crossover =
    (|P| * num_elite)/2;
    for j = 1 to num_crossover do
        Randomly select 2 parameter sets X_1 and X_2
        from Pop;
        Generate X_3 and X_4 by one-point crossover to
        X_1 and X_2;
        Save X_3 and X_4 to Pop_2;
    end for
    for j = 1 to num_crossover do
        Select a parameter set X_j from Pop_2;
        Mutate each element of X_j at a rate M and
        generate new solution X́_j;
        if X́_j is non-feasible then Update X́_j with a
        feasible solution by repairing X́_j;
        end if
        Update X_j with X́_j in Pop_2;
    end for
    Update Pop = Pop_1 + Pop_2;
end for
Return the set X which has the best value of Ω in Pop;
```

b) Pattern Search: Another effective solution methodology to solve (14) is Pattern Search Method, a simpler ver-

$$\min_{\boldsymbol{y}} \left(\frac{1}{2}\right) \boldsymbol{y^T} \hat{\boldsymbol{H}} \left(\mathcal{L}\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}, \lambda^1, \boldsymbol{\lambda^2}, \boldsymbol{\lambda^3}, \boldsymbol{\lambda^4}, \boldsymbol{\lambda^5}, \boldsymbol{\lambda^6}, \boldsymbol{\lambda^7}\right)\right) \boldsymbol{y} + \nabla\Omega\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right)$$

$$\mathbf{subject\ to} = \begin{cases}
y_i + W\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) \leqslant 0, \text{for } i = 1 \\
y_i + X\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) \leqslant 0, \text{for } i = 2, \ldots, |\mathbb{C}| + 1 \\
y_i + Y\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) = 0, \text{for } i = |\mathbb{C}| + 2, \ldots, 2|\mathbb{C}| + 1 \\
y_i + Z\left(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}\right) = 0, \text{for } i = 2|\mathbb{C}| + 2, \ldots, 2|\mathbb{C}| + |\mathbb{U}| + 1 \\
y_i + P_t^c - P_{t,\min}^c \leqslant 0, \text{for } i = 2|\mathbb{C}| + |\mathbb{U}| + 2, \ldots, 3|\mathbb{C}| + |\mathbb{U}| + 1 \\
y_i + \psi_{tilt}^c - 90^\circ \leqslant 0, \text{for } i = 3|\mathbb{C}| + |\mathbb{U}| + 2, \ldots, 4|\mathbb{C}| + |\mathbb{U}| + 1 \\
y_i + \psi_{tilt}^c \leqslant 0, \text{for } i = 4|\mathbb{C}| + |\mathbb{U}| + 2, \ldots, 5|\mathbb{C}| + |\mathbb{U}| + 1
\end{cases} \tag{18}$$

Algorithm 3: Pattern Search Algorithm for CLASS Implementation.

Input:
Algorithm 1 to solve (14)
Parameter space $\mathbf{S}(\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c})$
Output: Solution $\mathbf{X} = [\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}]$
$k = 0$;
while $k < iteration_{max}$ **do**
  Determine a step size s_k using exploratory search algorithm;
  Test Ω at parameter set $\boldsymbol{x_0}$ and two more points $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$ in a triangle;
  Label best, good and worst points as $\boldsymbol{x_B}$, $\boldsymbol{x_G}$ and $\boldsymbol{x_W}$;
  Reflect $\boldsymbol{x_W}$ on the plane as $\boldsymbol{x_R}$;
  if $\Omega(\boldsymbol{x_R}) > \Omega(\boldsymbol{x_G})$ **then**
    if $\Omega(\boldsymbol{x_R}) > \Omega(\boldsymbol{x_B})$ **then** replace $\boldsymbol{x_W}$ with $\boldsymbol{x_R}$;
    else Find $\boldsymbol{x_E}|2\boldsymbol{x_R} - (\boldsymbol{x_B} + \boldsymbol{x_G})/2$, find $\Omega(\boldsymbol{x_E})$
      if $\Omega(\boldsymbol{x_E}) > \Omega(\boldsymbol{x_B})$ **then** replace $\boldsymbol{x_W}$ with $\boldsymbol{x_E}$;
      end if
    end if
  else
    if $\Omega(\boldsymbol{x_R}) < \Omega(\boldsymbol{x_W})$ **then** replace $\boldsymbol{x_W}$ with $\boldsymbol{x_R}$;
      Compute $\boldsymbol{x_C} = ((\boldsymbol{x_B} + \boldsymbol{x_G})/2) + \boldsymbol{x_R})/2$, find $\Omega(\boldsymbol{x_C})$
    else Compute $\boldsymbol{x_C} = ((\boldsymbol{x_B} + \boldsymbol{x_G})/2) + \boldsymbol{x_W})/2$, find $\Omega(\boldsymbol{x_C})$
    end if
    if $\Omega(\boldsymbol{x_C}) < \Omega(\boldsymbol{x_W})$ **then** replace $\boldsymbol{x_W}$ with $\boldsymbol{x_C}$;
    else Compute $\boldsymbol{x_S} = (\boldsymbol{x_B} + \boldsymbol{x_W})/2$ and replace $\boldsymbol{x_W}$ with $\boldsymbol{x_S}$ and $\boldsymbol{x_G} = (\boldsymbol{x_B} + \boldsymbol{x_G})/2$
    end if
  end if
  Compute $p_k = \Omega(\boldsymbol{x_k}) - \Omega(\boldsymbol{x_k} + s_k)$
  if $p_k > 0$ **then** $\boldsymbol{x_{k+1}} = \boldsymbol{x_k} + s_k$
  else $\boldsymbol{x_{k+1}} = \boldsymbol{x_k}$
  end if
  Update pattern vectors and step size $k = k + 1$
end while
Return $\mathbf{X} = [\boldsymbol{P_t^c}, \boldsymbol{\psi_{tilt}^c}, \boldsymbol{P_{CIO}^c}]$

sion of Powell's method [40]. Algorithm 3 presents a generic pseudo-code which describes the main elements of a pattern search method [41] where we use Nelder-Mead algorithm as the exploratory search algorithm within each iteration of pattern search [42].

V. System Level Performance Analysis

A. Simulation Setup

We employ a LTE 3GPP standard compliant network topology simulator [37] to generate typical macro and small cell based network and user distributions. The simulation parameters details are given in Table III.

We use wrap around model to simulate interference in an infinitely large network thus avoiding boundary effect. To model realistic networks, users are distributed non-uniformly in all the sectors such that a fraction of users are clustered around randomly located hotspots in each sector. Monte Carlo simulations are used to estimate average performance of the algorithms. We consider five different user traffic requirement profiles corresponding to 24 kbps, 56 kbps, 128 kbps, 512 kbps and 1024 kbps desired throughput.

TABLE III
Parameter Settings for Simulation

System Parameters	Value
No. of Macro Base Stations	7
Sectors per Base Station	3
Small Cells per Sector	1
Number of Users per Sector	25
Transmission Frequency	2 GHz
Transmission Bandwidth	10 MHz
Network Topology	Hexagonal
Macro Cell Transmit Power	Max: 46 dBm, Min: 40 dBm
Macro Cell Antenna Tilt	Max: 15°, Min: 0°
Small Cell Transmit Power	Max: 30 dBm, Min: 27 dBm
Small Cell CIO	Max: 10 dB, Min: 0 dB
Fixed Parameter Settings (FPSs)	Macro Transmit Power: 43 dBm; Small Transmit Power: 27 dBm; Tilt: 0° (FPS - 0), 10° (FPS - 10), 15° (FPS - 15), 20° (FPS - 20); CIO: 0 dB
Cellular System Standard	LTE
Macro Cell Height	25 m
Small Cell Height	10 m
Inter-site Distance (Macro)	500 m
Macro Cell Antenna Gain	17 dBi
Small Cell Antenna Gain	5 dBi
Coverage Threshold P_{th}^c	95%
Load Threshold η_{th}^c	100%

B. Results

In this Section, we evaluate the impact of different α values used in load-aware user association on CLASS along with a comparison of load-aware user association with state-of-the-art-maximum RSRP and maximum SINR user association methods. Using the proposed load-aware user association with best performing α value, we then compare results from 4 Fixed Parameter Settings (FPSs) against the optimal parameter values returned by both CLASS equations using SQP, genetic algorithm and pattern search to demonstrate their gain. For simplicity, the CLASS solution in equation (14 a) is henceforth referred to as CLASS1 and solution in equation (14 b) as CLASS2. The results of proposed solutions are further compared with the two algorithms that are most relevant to this work i.e., the distributed tilt-based CCO solution presented in [6] and the tilt-based CCO-LB function given in [32]. It is important to note here explicitly that due to the use of virtual loads in our system, the user association from [32] returns undefined results. Therefore, the algorithm in [32] is implemented using load-aware user association.

1) Impact of Load-aware User Association: The proposed load-aware user association (11) is dependent on 3 features: cell loads at the time of association, downlink received power with CIO and the association exponent α. The impact of cell loads and received powers on user association are obvious from (11); however, the impact of exponent value on user association requires quantitative evaluations of system KPIs for different values α. A very relevant KPI in this case is the cell load and its

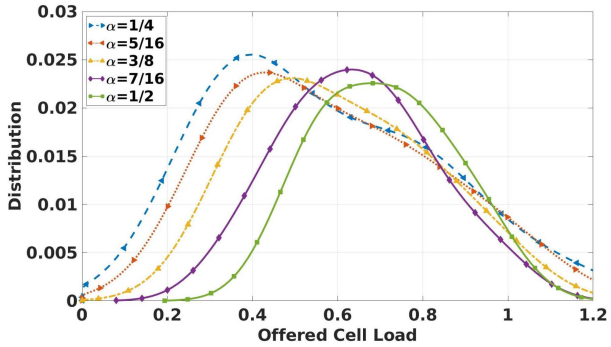


Fig. 3. Comparison of offered cell load distribution for α values in load-aware user association.

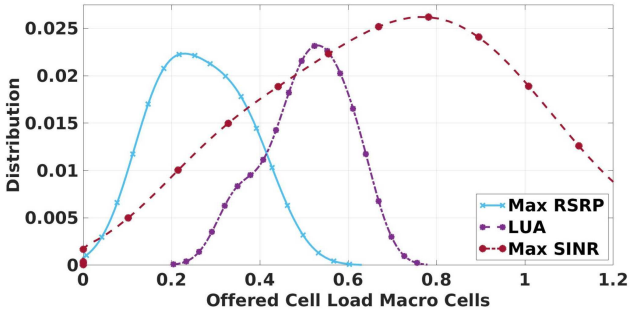


Fig. 4. Comparison of offered macro cell load distribution for load-aware (LUA) vs. Max RSRP and Max SINR user association.

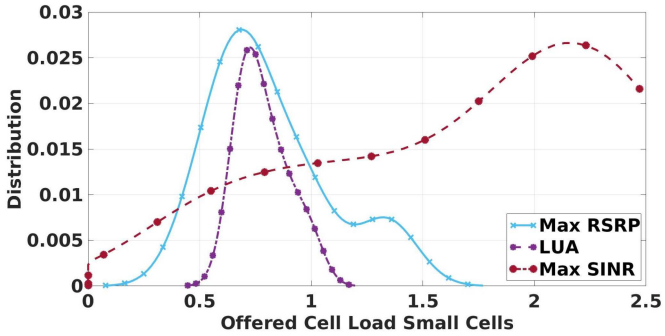


Fig. 5. Comparison of offered small cell load distribution for load-aware (LUA) vs. Max RSRP and Max SINR user association.

distribution among cells for given total traffic in the network. A lower average cell load and smaller load variance among cells for given traffic reflects a better performing user association scheme and vice versa. Though we have performed a comparison of $\alpha \in [0, 1]$ for both CLASS formulations, for brevity Fig. 3 only presents cell load distribution for $\alpha \in [\frac{1}{4}, \frac{1}{2}]$.

From the results in Fig. 3 it can be seen that the load distribution improves and becomes the most compact at $\alpha = \frac{7}{16}$ and starts to spread beyond it. Using $\alpha = \frac{7}{16}$ Figs. 4 and 5 present a comparison of the proposed load-aware user association with coverage based Max RSRP user association and quality based Max SINR user association techniques for macro and small cells.

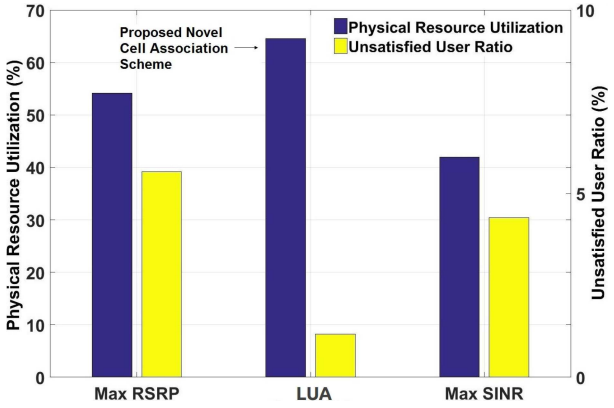


Fig. 6. Comparison of network utilization and unsatisfied user ratio for load-aware (LUA) vs. Max RSRP and Max SINR user association.

The results in Fig. 4 show that the proposed load-aware user association manages to keep macro cell loads within 80%, Max RSRP keeps macro cell loads to within 60%, while Max SINR association overloads a number of macro cells due to their stronger signals. In comparison, Fig. 5 shows that the proposed load-aware user association technique attempts to distribute load evenly between macro and small cells, with only a few small cells marginally overloaded. On the other hand, due to a lack of load awareness, both Max RSRP and Max SINR association overload the small cells with more than half the small cells overloaded. The even load distribution offered by the load-aware user association methodology also results in fewer unsatisfied users i.e., users who are unable to achieve their desired throughput due to a lack of physical resources at the serving cell.

This is evidenced by the ratio of unsatisfied users in the network and the utilization of physical resources in the network given in Fig. 6. We can see that while the load-aware user association occupies more resources, it is able to minimize the ratio of unsatisfied users by evenly distributing the load between cells. On the other hand, the Max RSRP and Max SINR user association schemes are oblivious to the needs of the users and blindly associate them with cells offering best coverage and quality. This leads to cells becoming overloaded and higher ratio of unsatisfied users. The results in Figs. 4, 5, and 6, also demonstrate that the flexibility in the design of the proposed load-aware user association scheme allows it to be an effective coverage, capacity and load optimization solution, even when deployed independently in a cellular network.

2) Comparative Analysis of Proposed Solutions:

a) Downlink SINR: To compare the performance of the two CLASS formulations, we use downlink SINR as the benchmark performance indicator. In Fig. 7 we compare the results for CLASS1 obtained using SQP, genetic algorithm and pattern search against different fixed parameter settings defined in Table III. The results show that 50th percentile users achieve 14 dB SINR with CLASS1-PS compared to 10 dB for top performing FPS-20. In Fig. 8, the same comparison is presented for CLASS2 which shows that 50th percentile users achieve 4.5 dB higher SINR with CLASS2 compared to FPS-20. Recall that using CIOs alone for LB has negative impact on SINR as demonstrated in Figs. 1(a) and (b). But when CIOs are adapted through the proposed load-aware user association in conjunction with transmit power and antenna tilts, we still achieve a gain in

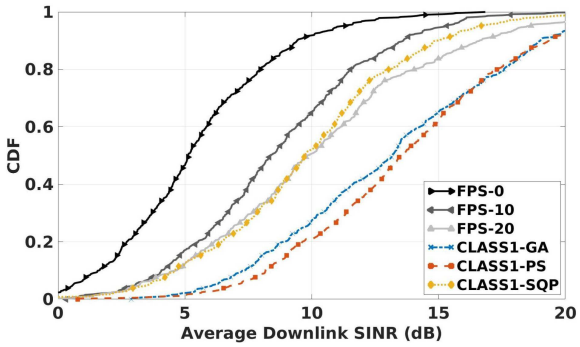


Fig. 7. Downlink SINR CDF - FPSs vs. CLASS1-genetic algorithm (GA), pattern search (PS) and SQP.

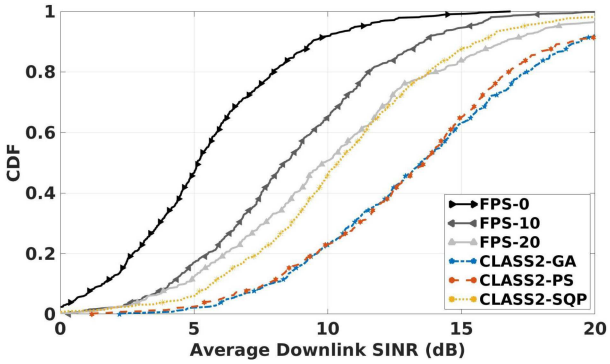


Fig. 8. Downlink SINR CDF - FPSs vs. CLASS2-genetic algorithm (GA), pattern search (PS) and SQP.

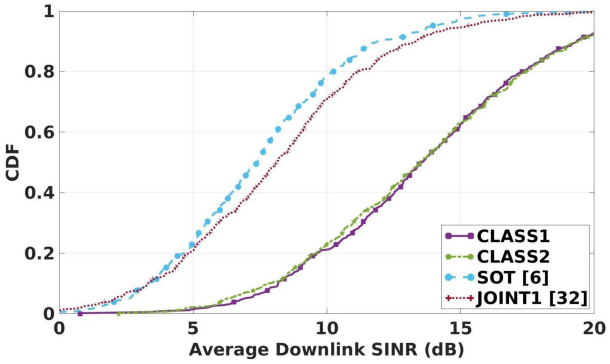


Fig. 9. Downlink SINR CDF - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

SINR. This rationalizes the need to include all three optimization parameters in the proposed CCO-LB solution, compared to existing studies which use one or two parameters at a time. Another key results to point out here is that the solutions obtained using genetic algorithm and pattern search perform better for both CLASS1 and CLASS2 compared to SQP. This is due to the fact that the genetic algorithm and pattern search attempt to find the global optimum whereas SQP is a gradient driven process that is vulnerable to convergence to local extrema.

Fig. 9 compares the best solution obtained for CLASS1 (pattern search) and CLASS2 (genetic algorithm) against the CCO algorithm proposed in [6] referred to by the authors as SOT, and

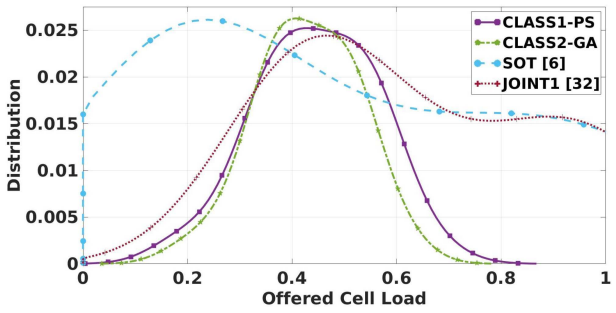


Fig. 10. Offered cell load distribution - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

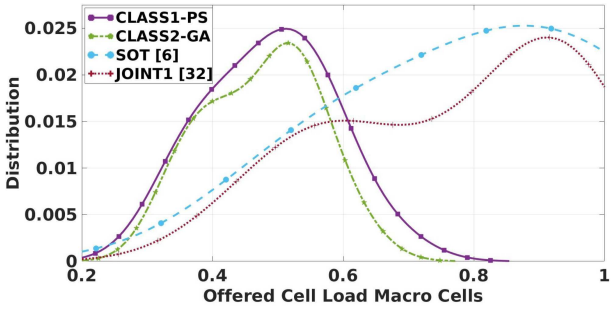


Fig. 11. Offered macro cell load distribution - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

the CCO-LB algorithm JOINT1 presented in [32]. Results show that CLASS1 and CLASS2 offer SINR $>$ 10 dB for almost 80% of users. In comparison, with SOT and JOINT1 only 20% and 30% of users have SINR above 10 dB respectively. We also see that CLASS1 performs slightly better compared to CLASS2 for cell edge users i.e., the lower half of users with CLASS2 giving slightly better performance for the top half. This is because of the use of geometric mean in CLASS1 which forces fairness in all user throughputs, whereas the use of arithmetic mean attempts to maximize the extreme throughput values.

b) Offered Cell Load: Fig. 10 compares offered cell loads for CLASS1, CLASS2, SOT and JOINT1. The results show that for CLASS1, the cell loads range from 10% to 80%, and from 10% to 70% for CLASS2. This difference is due to the higher focus of CLASS1 on fairness which means it attempts to increase throughput of low SINR users by allocating them more resources compared to CLASS2 which only focuses on maximizing total throughput. By comparison, SOT shows the widest disparity among cell loads. This is primarily due to the fact that SOT is a CCO-only algorithm that only optimizes antenna tilts, thus highlighting the importance of formulating LB and CCO jointly with all three parameters. JOINT1 being a CCO-LB solution that incorporates two parameters i.e., antenna tilts and CIOs, offers better load balancing compared to SOT, but is still significantly outperformed by both CLASS1 and CLASS2.

Figs. 11 and 12 show the performance of the proposed CCO-LB solution in terms of LB and QoS by showing load distributions for macro and small cells separately. While none of the macro or small cells are overloaded by the CLASS solutions, SOT heavily favors macro cells over small cells for loading

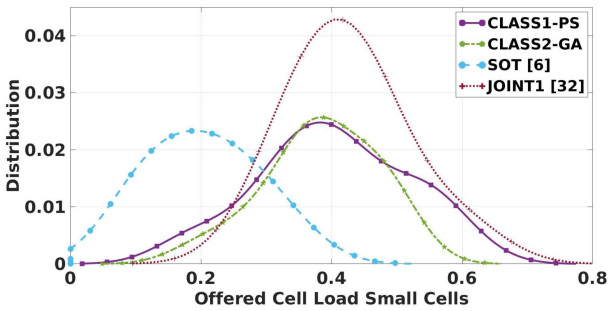


Fig. 12. Offered small cell load distribution - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

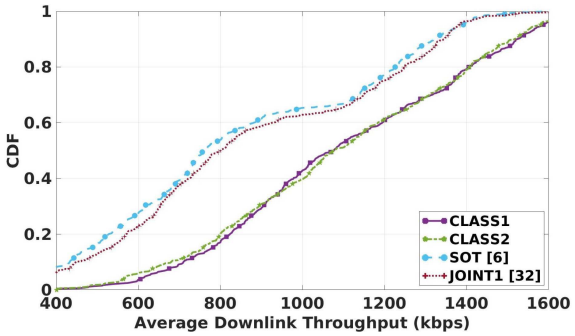


Fig. 14. Downlink throughput CDF - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

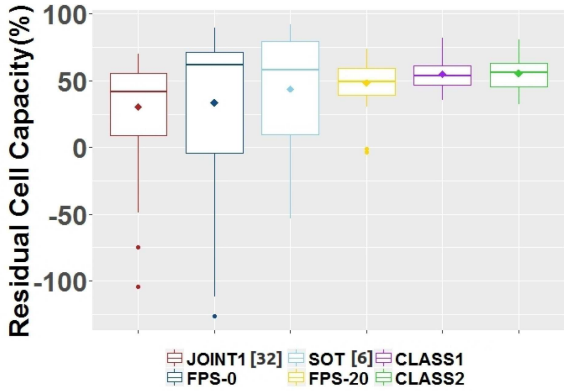


Fig. 13. Residual cell capacity - FPS-0, FPS-20, SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

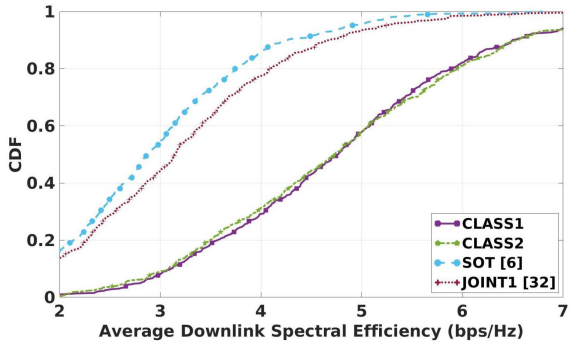


Fig. 15. Downlink spectral efficiency CDF - SOT [6], JOINT1 [32] vs. CLASS1 and CLASS2.

causing almost 50% of the macro cells to become overloaded. Similarly, since JOINT1 only optimizes CIOs and antenna tilts, it also favors macro cells for load bearing over small cells. Another key insight here is that contrary to existing load balancing schemes [7], [8], [12]–[15], [22]–[26], the proposed solution not only balances loads between macro and small cells but actually increases capacity in the system by jointly optimizing soft and hard parameters, thereby satisfying CCO objective at the same time.

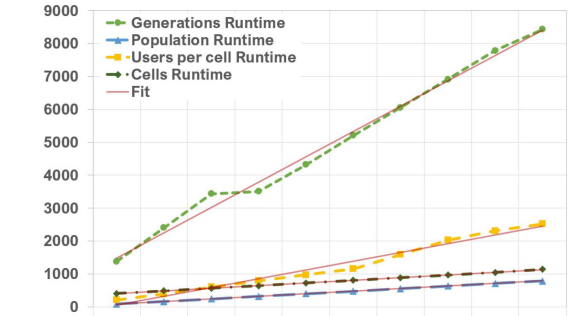
This is further put into perspective when we observe the residual cell capacity across the network, as shown in Fig. 14. The box plots show the median residual capacity value along with the distance between 1st and 3rd quartiles, whereas the points inside the box plots signify the mean residual capacity. The average residual cell capacity of the proposed CLASS1 and CLASS2 solutions are 54.8% and 55.5% respectively, which is 20% more than the average residual capacity of the algorithm in [6], and over 45% more than the residual capacity of the algorithm in [32]. However, the key observation in Fig. 13 is compactness of the 1st and 3rd quartile, and the outer fences for CLASS solutions compared to the residual capacities of other solutions. The increased residual capacity creates additional space for transit users within each cell, a feature that is highly desirable in ultra-dense HetNets due to the expected high user mobility.

c) Downlink User Throughout: Fig. 14 plots the average downlink user throughput CDF for all the users in the network with CLASS1, CLASS2, SOT and JOINT1. We observe a significant gain in user throughput for CLASS1 and CLASS2 compared to both SOT and JOINT1. Note that the observed gain in throughput offered by CLASS solutions is despite the fact desired user throughputs are pre-set and that PRBs are allocated to each user based on that requirement. The observed gain in throughput occurs due to the user SINR at the time of cell association in calculation of PRBs required to serve a user. The same PRBs later result in better throughput for the user when the user SINR improves as a result of the parameter optimization by the proposed solution. In real system, this throughput increase beyond desired user throughput can be controlled by doing SINR calculations more frequently e.g., using CQI reports.

d) Downlink Spectral Efficiency: Fig. 15 shows the CDF for downlink spectral efficiency in the network. CLASS solutions provide the highest spectral efficiency. As spectral efficiency is a function of throughput, the same logic as for user throughput applies here too. However, the impact of SINR on spectral efficiency is also visible with the plot for spectral efficiency following similar trend as SINR.

C. Performance Analysis of the Proposed CLASS Solutions

The complexity of the proposed CLASS solution depends on two factors: 1) the execution time of algorithm 1, and 2) the execution time of the optimization algorithm. The execution time of algorithm 1 comes out to be $O(|\mathbb{U}| + 2|\mathbb{C}| + 2|\mathbb{U}||\mathbb{C}|)$ which can be generalized as $O(|\mathbb{U}||\mathbb{C}|)$. This means that the runtime of algorithm 1 increases linearly with increase in the number of users $|\mathbb{U}|$ and cells $|\mathbb{C}|$. Any additional execution time depends on the optimization algorithm being used. Assuming genetic algorithm is used to optimize the cell parameters, its



Generations	50	100	150	200	250	300	350	400	450	500
Samples per Population	50	100	150	200	250	300	350	400	450	500
Users per Cell	5	10	15	20	25	30	35	40	45	50
Cells	42	63	84	105	126	147	168	189	210	231

Fig. 16. Actual vs. fitted runtimes for CLASS algorithms for different values of $\boldsymbol{G}$, $\boldsymbol{P}$, $|\mathbb{U}|$ and $|\mathbb{C}|$.

execution time can be obtained from [43], which comes out to be $O(\boldsymbol{GP})$.

Thus, the total runtime of the proposed solution is $O(\boldsymbol{GP}|\mathbb{U}||\mathbb{C}|)$ which is linear in all four variables. This is also demonstrated in Fig. 16 which shows the experimental algorithm runtimes for varying values of $\boldsymbol{G}, \boldsymbol{P}, |\mathbb{U}|$ and $|\mathbb{C}|$. Given the computational powers of state-of-the-art network controllers, this execution time is easily manageable. Furthermore, network operators can use big data analytics, as proposed in [36] to predict cell loads and obtain optimal parameters proactively to minimize the impact of computation delay on subscriber QoE. Apart from this, the implementation of the proposed load-aware user association requires only one additional multiplication step on top of calculating RSRP(+CIO) for each UE. This, given the capabilities of today's smartphones, is not a significant computational burden.

D. Practical Implementation of Proposed CLASS Solutions in Current and Future Mobile Cellular Networks

To implement CLASS solutions in a real network, idle (disconnected) users must be informed about cell loads at the time of association whereas association decision for active (connected) users will be made by the network based on user measurement reports and cell load data. State-of-the-art networks have this information in the form of Total PRB Usage [44], that can act as a proxy for cell load until a tailor-made measurement is made available in future standards to implement CCO and LB.

Also, to successfully balance cell loads across the network, it helps to generate and leverage user traffic prediction model. Most existing operators already construct some form of this model on their own. Current standardization includes a traffic classification parameter called Number of Active users in the DL per QCI [44]. This measure can act as proxy for expected data rate and or QoS requirements until a custom measurement to facilitate CCO-LB and other SON function that can benefit from intelligence of QoS expectations, is standardized for future networks.

Moreover, in this paper, we considered same maximum load threshold $\eta^c_{th} = 100\%$ for all cells. However, in real networks and in advanced implementation of proposed CCO-LB, setting individual cell load thresholds can be useful in scenarios, where different cells are known to have different user arrival rate, sojourn times and traffic statistics. This can also be useful, where the power consumption model of the BS in different cells are different and loading points that return optimal energy efficiency in individual cells are different.

VI. Conclusion

In this paper, we presented a framework for joint CCO and LB SON functions with transmit powers, antenna tilts and CIOs as the optimization parameters. The proposed CCO-LB solution (CLASS) not only provides significant gains in terms of downlink SINR and throughput, it also provides balanced distribution of cell loads in a heterogeneous network which is key to meeting overall resource efficiency demands. We also show that the key metrics for quantifying gains for the joint CCO-LB function are not merely user SINR, or throughput or spectral efficiency, but also, and most importantly, the amount of free resources in the network after all users are satisfied, what we call residual capacity. Maximization of residual capacity is the key to achieving temporal stability in the network optimization process due to the acute mobility dynamics of HetNets. Further gain in throughput and spectral efficiency may become possible by softening the constraint of desired user throughput and by incorporating scheduling level decisions in the future. Nevertheless, the proposed CCO-LB solution substantially outperforms the comparable algorithms proposed in literature for all KPIs without exception because unlike prior works: 1) it exploits joint optimization of all three parameters that influence coverage and cell association; thus in addition to just shifting load, it shifts load in a way that increases overall system capacity; 2) it leverages a smarter load aware cell association mechanism, and 3) though the objective function targets throughput maximization and thus aims for CCO, the formulation is designed to incorporate LB in the objective function itself through use of geometric mean. This yields better results compared to solutions that target CCO and take LB as a constraint and vice versa, because a goal included as constraint is likely to yield acceptable but not optimal results.

The joint CLASS formulation presented in this work can pave the way for several future studies and SON function developments. For example, it is possible to incorporate energy efficiency (EE) and mobility robustness optimization (MRO) SON functions into the formulation by setting the load thresholds for intelligently selected cells to zero based on user mobility and activity profiles. Incorporation of big data aided knowledge like optimal cell load thresholds for each cell by considering spatio temporal prediction of oncoming traffic is another promising research direction.

Appendix A

The dowlink received power based on standard exponential pathloss model with Gaussian distributed shadowing for user *u* associated with cell *c* is:

$$P^c_{r,u} = P^c_t G_u G^c_u \delta^c_u a \left(d^c_u\right)^{-\beta}$$

Due to the randomness of δ^c_u, the coverage constraint of user *u* i.e. $C_u := P^c_{r,u} \geqslant P^c_{th}$ becomes a function of δ^c_u such that for user *u*, the coverage constraint will be satisfied with some probability i.e. $Pr(P^c_{r,u}(\delta^c_u) \geqslant P^c_{th})$. We can then calculate the minimum value of δ^c_u above which the coverage constraint for

an individual user will be satisfied.

$$Pr\left(P_t^c G_u G_u^c \delta_u^c a\left(d_u^c\right)^{-\beta} \geqslant P_{th}^c\right)$$

$$Pr\left(\delta_u^c \geqslant \frac{P_{th}^c}{P_t^c G_u G_u^c a\left(d_u^c\right)^{-\beta}}\right) \tag{19}$$

Based on the assumption that δ_u^c has a Gaussian distribution, the value of δ_u^c inside the parentheses in (19) gives the Z-score below which the coverage constraint of an individual user will be violated. If p gives the probability $Pr(\delta_u^c \geqslant \frac{P_{th}^c}{P_t^c G_u G_u^c a(d_u^c)^{-\beta}})$, we can remodel the event that a user is inside the coverage of its serving cell as a Bernoulli variable with probability p.

The consequence of modeling $Pr(P_{r,u}^c \geqslant P_{th}^c)$ as a Bernoulli random variable is that the network coverage C can be modeled as a Binomial random variable with chance of success p per user. Thus the probability of having $\varpi * |\mathbb{U}|$ users or more in coverage can be given as:

$$Pr(k \geqslant \varpi * |\mathbb{U}|) \geqslant \sum_{i=\varpi * |\mathbb{U}|}^{|\mathbb{U}|} \binom{|\mathbb{U}|}{i} p^i (1-p)^{|\mathbb{U}|-i} \tag{20}$$

If the desired value for $Pr(k \geqslant \varpi * |\mathbb{U}|) \rightarrow 1$, and $\varpi \rightarrow 1$, then

$$\lim_{\substack{\varpi \longrightarrow 1 \\ Pr(k \geqslant \varpi * |\mathbb{U}|) \longrightarrow 1}} p = 1$$

In such a scenario, we can substitute the probability p with the indicator function $1(.)$, thus giving us the following formulation for constraint (i):

$$\frac{1}{|\mathbb{C}|} \sum_{\mathbb{C}} \frac{1}{|\mathbb{U}_c|} \sum_{\mathbb{U}_c} 1\left(P_{r,u}^c \geqslant P_{th}^c\right)$$

Appendix B

The simplified form of the SINR function in antenna tilts is given as:

$$\gamma_u^c = \frac{10^{\mu\left(\left(\psi_u^c - \psi_{tilt}^c\right)^2 + x_u^c\right)}}{\kappa + 10^{\mu\left(\left(\psi_u^i - \psi_{tilt}^i\right)^2 + x_u^i\right)}}$$

The expression for antenna gains expressed as a function of tilts is given as:

$$G_u^c = 10^{\mu\left(\left(\psi_u^c - \psi_{tilt}^c\right)^2 + x_u^c\right)}$$

We treat x_u^c and μ as constants and assign then unit value and -1.2 respectively which gives us the resulting function of ψ_{tilt}^c:

$$f(\psi_{tilt}^c) = 0.0631 * 10^{-1.2\left(\psi_u^c - \psi_{tilt}^c\right)^2}$$

Taking derivative of $f(\psi_{tilt}^c)$ gives us:

$$f'(\psi_{tilt}^c) = 0.0631 l_n(10) * 10^{-1.2(\psi_u^c - \psi_{tilt}^c)^2} * (2.4(\psi_u^c - \psi_{tilt}^c))$$

Taking the second derivative:

$$f''(\psi_{tilt}^c) = 0.3634 l_n(10) * 10^{-1.2\left(\psi_u^c - \psi_{tilt}^c\right)^2} * [(\psi_u^c - \psi_{tilt}^c) l_n(10) - 0.417]$$

For a function to be convex, the second derivative has to be non-negative which is only possible in the range $[(\psi_u^c - \psi_{tilt}^c) \leqslant -0.4254, (\psi_u^c - \psi_{tilt}^c) \geqslant 0.4254]$. Hence, the antenna gain function is a non-convex function and by extension, the SINR expression with antenna gain is non-convex.

References

[1] V. Chandrasekhar, J. G. Andrews, and A. Gatherer, "Femtocell networks: A survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[2] X. Ge, S. Tu, G. Mao, C.-X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[3] R. Hu, Y. Qian, S. Kota, and G. Giambene, "HetNets—A new paradigm for increasing cellular capacity and coverage [Guest Editorial]," *IEEE Wireless Commun.*, vol. 18, no. 3, pp. 8–9, Jun. 2011.

[4] 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.

[5] A. J. Fehske, I. Viering, J. Voigt, C. Sartori, S. Redana, and G. P. Fettweis, "Small-cell self-organizing wireless networks," *Proc. IEEE*, vol. 102, no. 3, pp. 334–350, Mar. 2014.

[6] A. Imran, M. A. Imran, A. Abu-Dayya, and R. Tafazolli, "Self organization of tilts in relay enhanced networks: A distributed solution," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 764–779, Feb. 2014.

[7] 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.

[8] 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.

[9] R. Balakrishnan and B. Canberk, "Traffic-aware QoS provisioning and admission control in OFDMA hybrid small cells," *IEEE Trans. Vehicular Technol.*, vol. 63, no. 2, pp. 802–810, Feb. 2014.

[10] R. Hernandez-Aquino, S. A. R. Zaidi, D. McLernon, M. Ghogho, and A. Imran, "Tilt angle optimization in two-tier cellular networks—A stochastic geometry approach," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 5162–5177, Dec. 2015.

[11] 3rd Generation Partnership Project, "Radio Resource Control (RRC); Protocol Specification," 3GPP TS 36.331, Rev. V8.6.0 Release 8, 2009.

[12] X. Zhao, W. Zhang, and C. Wang, "A load prediction based virtual cell breathing scheme for LTE—A system," in *Proc. IEEE Military Commun. Conf.*, 2013, pp. 1296–1301.

[13] I. Siomina and D. Yuan, "Load balancing in heterogeneous LTE: Range optimization via cell offset and load-coupling characterization," in *Proc. IEEE Int. Conf. Commun.*, 2012, pp. 1357–1361.

[14] H. Du *et al.*, "A load fairness aware cell association for centralized heterogeneous networks," in *Proc. IEEE Int. Conf. Commun.*, 2015, pp. 2178–2183.

[15] C. A. S. Franco and J. R. B. de Marca, "Load balancing in self-organized heterogeneous LTE networks: A statistical learning approach," in *Proc. IEEE 7th Latin-Amer. Conf. Commun.*, 2015, pp. 1–5.

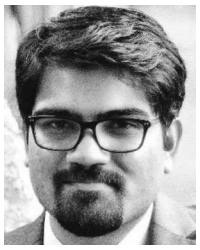
[16] A. Engels, M. Reyer, X. Xu, R. Mathar, J. Zhang, and H. Zhuang, "Autonomous self-optimization of coverage and capacity in LTE cellular networks," *IEEE Trans. Veh. Technol.*, vol. 62, no. 5, pp. 1989–2004, Jun. 2013.

[17] A. Awada, B. Wegmann, I. Viering, and A. Klein, "A mathematical model for user traffic in coverage and capacity optimization of a cellular network," in *Proc. IEEE 73rd Veh. Technol. Conf.*, 2011, pp. 1–5.

[18] S. Berger, M. Simsek, A. Fehske, P. Zanier, I. Viering, and G. Fettweis, "Joint downlink and uplink tilt-based self-organization of coverage and capacity under sparse system knowledge," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2259–2273, Apr. 2016.

[19] O. G. Aliu, A. Imran, M. A. Imran, and B. Evans, "A survey of self organisation in future cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 336–361, Jan./Mar. 2013.

[20] S. Singh and J. G. Andrews, "Joint resource partitioning and offloading in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 888–901, Feb. 2014.
[21] V. Buenestado, M. Toril, S. Luna-Ramírez, J. M. Ruiz-Avilés, and A. Mendo, "Self-tuning of remote electrical tilts based on call traces for coverage and capacity optimization in LTE," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4315–4326, May 2017.
[22] Z. Arslan, M. Erel, Y. Özcevik, and B. Canberk, "SDoff: A software-defined offloading controller for heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2014, pp. 2827–2832.
[23] I. Siomina and D. Yuan, "Analysis of cell load coupling for LTE network planning and optimization," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2287–2297, Jun. 2012.
[24] W. Song, W. Zhuang, and Y. Cheng, "Load balancing for cellular/WLAN integrated networks," *IEEE Netw.*, vol. 21, no. 1, pp. 27–33, Jan./Feb. 2007.
[25] Q. Song and A. Jamalipour, "A network selection mechanism for next generation networks," in *Proc. IEEE Int. Conf. Commun.*, vol. 2, 2005, pp. 1418–1422.
[26] A. Imran, E. Yaacoub, M. A. Imran, and R. Tafazolli, "Distributed load balancing through self organisation of cell size in cellular systems," in *Proc. IEEE 23rd Int. Symp. Person. Indoor Mobile Radio Commun.*, 2012, pp. 1114–1119.
[27] Z. Liu, P. Hong, K. Xue, and M. Peng, "Conflict avoidance between mobility robustness optimization and mobility load balancing," in *Proc. IEEE Global Telecommun. Conf.*, 2010, pp. 1–5.
[28] P. Mu, R. Barco, and S. Fortes, "Conflict resolution between load balancing and handover optimization in LTE networks," *IEEE Commun. Lett.*, vol. 18, no. 10, pp. 1795–1798, Oct. 2014.
[29] K. T. Dinh and S. Kuklinski, "Joint implementation of several LTE-SON functions," in *Proc. IEEE Globecom Workshops*, 2013, pp. 953–957.
[30] 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.
[31] H. Y. Lateef, A. Imran, and A. Abu-Dayya, "A framework for classification of self-organising network conflicts and coordination algorithms," in *Proc. IEEE 24th Int. Symp. Person. Indoor Mobile Radio Commun.*, 2013, pp. 2898–2903.
[32] 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.
[33] 3rd Generation Partnership Project, *Self-Configuring and Self-Optimizing Network (SON) Use Cases and Solutions*, 3GPP TR 36.902, Rev. V9.2.0 Release 9, 2010.
[34] T. Jansen *et al.*, "Embedding multiple self-organisation functionalities in future radio access networks," in *Proc. IEEE 69th Veh. Technol. Conf.*, 2009, pp. 1–5.
[35] H. Kim, G. de Veciana, X. Yang, and M. Venkatachalam, "Alpha-optimal user association and cell load balancing in wireless networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, 2010, pp. 1–5.
[36] A. Imran, A. Zoha, and A. Abu-Dayya, "Challenges in 5G: How to empower SON with big data for enabling 5G," *IEEE Netw.*, vol. 28, no. 6, pp. 27–33, Nov./Dec. 2014.
[37] 3rd Generation Partnership Project, "Further advancements for E-UTRA physical layer aspects," 3rd Generation Partnership Project TR 36.814, Rev. V9.0.0 Release 9, 2010.
[38] R. Muirhead, "Proofs that the arithmetic mean is greater than the geometric mean," *Math. Gazette*, vol. 2, no. 39, pp. 283–287, 1903.
[39] D. F. Shanno, "On Broyden-fletcher-Goldfarb–Shanno method," *J. Opt. Theory Appl.*, vol. 46, no. 1, pp. 87–94, 1985.
[40] M. J. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, 1964.
[41] W. E. Hart, "Evolutionary pattern search algorithms," Sandia Nat. Laboratories Rep. SAND95-2293, Albuquerque, New Mexico, Sep. 1995.
[42] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, no. 4, pp. 308–313, 1965.
[43] R. L. Haupt and S. E. Haupt, *Practical genetic algorithms*. New York: Wiley, vol. 2, 1998.
[44] 3rd Generation Partnership Project, "Evolved Universal Terrestrial Radio Access (E-UTRA); Layer 2 - Measurements," 3GPP TS 36.314, Rev. V8.3.0 Release 8, 2010.
[45] A. Asghar, H. Farooq, and A. Imran, "A novel load-aware cell association for simultaneous network capacity and user QoS optimization in emerging HetNets," in *Proc. IEEE 28th Annu. Int. Symp. Person., Indoor, Mobile Radio Commun.* (PIMRC), Montreal, QC, 2017, pp. 1–7.



Ahmad Asghar (S'17) received the B.Sc. degree in electronics engineering from Ghulam Ishaq Khan Institute of Science and Technology, Khyber Pakhtunkhwa, Pakistan, in 2010 and the M.Sc. degree in electrical engineering from Lahore University of Management and Technology, Punjab, Pakistan, in 2014. He is currently working toward the Ph.D. degree in electrical and computer engineering with the University of Oklahoma, OK, USA, as well as contributing to multiple NSF funded studies on 5th Generation Cellular Networks. His research work includes studies on self-healing and self-coordination of self-organizing functions in future big-data empowered cellular networks using analytical and machine learning tools.

Hasan Farooq (S'14) received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 2009 and the M.Sc. degree by Research degree in information technology from Universiti Teknologi PETRONAS, Malaysia, in 2014 wherein his research focused on developing adhoc routing protocols for smart grids. He is currently working toward 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 a multinational QSON project on Self Organizing Cellular Networks and is currently contributing to two NSF funded projects on 5G SON. He is the recipient of Internet Society First Time Fellowship Award towards Internet Engineering Task Force 86th Meeting held in USA, 2013.

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 currently an Assistant Professor in telecommunications with the University of Oklahoma, Tulsa, OK, USA where he is the Founding Director of Big Data and Artifcial Intelligence (AI) Enabled Self Organizing (BSON) 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. He is leading four NSF funded Projects on 5G amounting to over $2.2 million. He has authored more than 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. He is an Associate Fellow of the Higher Education Academy, U.K., and a member of the Advisory Board to the Special Technical Community on Big Data of the IEEE Computer Society.