Standards Compliant CoMP Scheduling over LTE Heterogeneous Wireless Networks

R. Arnott*, A. Gurung†, C. Liang *, Y. Maruta‡, D. Pham†, N. Prasad§, S. Rangarajan§, P. Wells*
*NEC Telecom Modus, UK, e-mail: {Rob.Arnott, Caroline.Liang, Patricia.Wells}@emea.nec.com
†NEC Australia, e-mail: {Arun.Gurung, Duong.Pham}@nec.com.au
‡NEC Corporation, e-mail: {y-maruta@cq.jp.nec.com}
§NEC Labs America, e-mail: {prasad,sampath}@nec-labs.com

***Abstract*—We analyze Coordinated Multi-Point (CoMP) transmission and reception over heterogeneous wireless networks (HetNets). Our focus is on the design of a practical algorithm for joint resource allocation (joint scheduling) in each HetNet cluster. We present one such algorithm together with its detailed analysis and evaluation under two different approaches for obtaining and utilizing channel feedback from the users. The considered approaches as well as the evaluation methodology are fully compliant with the 3GPP LTE standard. Our evaluations reveal several important insights that are essential to achieve significant CoMP gains over HetNets.**

I. Introduction

Cell splitting is the strategy that all network operators have chosen to meet the anticipated explosive growth in data traffic. Under this strategy, in each cell as opposed to the conventional deployment of a single macro base station, multiple transmission points (TPs) are deployed [1]. While each such TP can be a high power macro enhanced base-station (a.k.a. eNB), it is much more likely to be a low-power remote radio head (RRH) or a pico node of limited capabilities. The resulting networks formed by such disparate TPs are referred to as heterogeneous networks (a.k.a. HetNets) and indeed, a vast majority of next generation wireless networks will be HetNets that conform to the 3GPP LTE standard [2]. The "baseline" scheme for resource allocation in such HetNets is one where each user independently associates to the TP from which it receives the strongest average signal power. Each TP then performs dynamic scheduling at a fine subframe level time scale (every millisecond in LTE networks) over the pool of users that have associated to it, without coordinating with any other TP. Over many realistic network topologies this simple baseline scheme itself captures most of the average spectral efficiency (SE) gains promised by cell splitting [1]. Indeed, simulation results obtained as part of the LTE Release 11 standardization suggest that the potential benefit of more sophisticated coordinated resource allocation schemes is mainly to achieve significant gains in the $5-$percentile SE, while retaining the average SE gains of the baseline. To facilitate implementation of such coordinated resource allocation schemes, the favored architecture is one where the set of TPs is partitioned into multiple clusters. Each cluster represents a basic coordination unit comprising of multiple TPs that is tasked with serving a set of users. Furthermore, each user can be associated with only one cluster and the association of users to clusters needs to be determined at a much coarser time-scale (once every hundred or more milliseconds). In the scenarios that were considered during downlink CoMP standardization in LTE Release 11 [3] the TPs within each cluster were assumed to be connected via a reliable ultra-low latency backhaul such as optical fiber. Moreover, the TPs in each cluster were assumed to rely on directions from a central scheduler. These scenarios consequently are ideal for using centralized algorithms for coordinated resource allocation.

Our interest in this paper is in the design of algorithms for subframe level coordination within each cluster. Since clustering and user association happen on time scales which are orders of magnitude coarser, we assume them to be given and fixed. The design of joint resource allocation within a cluster of multiple TPs has received considerable attention in recent years [4]–[12]. These techniques range from those assuming perfect global knowledge of user channels states and their respective data at a central processer [4], to ones where only user channel states are shared among TPs in a cluster to jointly optimize downlink transmission parameters (such as beam-vectors and precoders) under several practical limitations [12]. The wisdom accrued from all these works is that substantial performance gains are possible via coordinated resource allocation provided that the underlying system constraints are properly addressed.

Our goal in this work is to assess the gains that can be achieved over real HetNets via resource allocation schemes that fully adhere to the LTE standard. A significant fraction of these HetNets will use the FDD mode. Indeed, according to the latest 3GPP statistics, 312 out of 360 operators only deploy the FDD mode. Further, most deployments of pico nodes (or small cells) in the foreseeable future will be ones where those nodes are equipped with 2 transmit antennas. Consequently, the major challenges that need to be addressed in order to achieve the possible CoMP gains, are those posed by quantized and delayed channel state information (CSI) feedback from the users (which must also be obtained in a standards compliant manner) and limited scope for precoding or beamforming. In this work we emulate such deployments following the detailed methodology of the 3GPP LTE standard. We formulate a resource allocation problem that is tailored towards CoMP transmission schemes that provide robustness against CSI errors, and which accommodates important mandatory constraints pertaining to the choice of transmission parameters. We propose two approaches for obtaining and

978-3-9018-8273-9/15/ ©2015 IFIP

utilizing CSI feedback from the users, that are both fully standards compliant. The other main challenge that needs to be addressed is the need for very low complexity resource allocation algorithms that can be implemented in the fine time-scale of interest. To meet this last challenge, we systematically develop a low-complexity resource allocation algorithm by first ascertaining the hardness of the problem at hand and then adopting an iterative approach, wherein each iteration involves solving a sub-problem on submodular set-function maximization. System evaluations of our algorithm and proposed approaches to obtain and exploit feedback, over a HetNet comprising of several hundred TPs serving a few thousand users reveal several important insights, such as:

- *Utilizing ACK/NACK feedback from the users is essential*
- *Controlling the interference estimates employed by the users to determine their CSI reports is essential*
- *The feedback approach that is geared towards coordinated muting yields the best results*
- *Exploiting a rank restriction feature is beneficial.*

With these insights in hand, we demonstrate that significant CoMP gains (around 34% $5-$percentile SE gain along with about 2% gain in the average SE) over the baseline scheme, can indeed be realized in realistic HetNets.

II. System Model

Consider the downlink in a HetNet with universal frequency reuse and focus on a cluster of M TPs that are synchronized and connected via a fiber backhaul. These TPs can simultaneously transmit on N orthogonal resource block groups (RBGs) during each scheduling interval (each subframe) in order to serve a pool of K active users. Each RBG comprises of a fixed number of resource elements (REs). It spans the duration of a subframe and is a slice of the available bandwidth. For our purposes an RBG represents the minimum unit that can be assigned to any user in each subframe. Each TP as well as each user can be equipped with multiple antennas. We let $\mathcal{U}$ denote the set of K users, $\mathcal{M}$ denote the set of M TPs and $\mathcal{N}$ denote the set of N RBGs. Then, the signal received by user k on the ℓ^{th} RE of RBG j in a subframe can be modelled as

$$\mathbf{y}_{j,k,\ell} \quad = \quad \sum_{m=1}^{M} \mathbf{H}_{j,k,m}\mathbf{t}_{j,m,\ell} + \mathbf{v}_{j,k,\ell}, \tag{1}$$

where, for notational convenience we have dropped the sub-frame index. The matrix $\mathbf{H}_{j,k,m}$ models the MIMO channel between TP m and user k on RBG j and is constant over all REs within that RBG. This channel matrix includes small-scale fading, large-scale fading and path attenuation. The vector $\mathbf{v}_{j,k,\ell}$ captures the additive noise as well as the out-of-cluster interference, while $\mathbf{t}_{j,m,\ell}$ denotes the signal vector transmitted by TP m on the ℓ^{th} RE in RBG j. [1] Considering the latter signal vector, we impose a restriction that each TP is allowed to serve at-most one user on each RBG. This restriction is imposed here to maintain readability but it is commonly used as a safeguard against inaccurate (quantized) feedback from the users. Then, we can express $\mathbf{t}_{j,m,\ell}$ as

$$\mathbf{t}_{j,m,\ell} = \mathbf{W}_{j,m}\mathbf{b}_{j,m,\ell}, \tag{2}$$

where $\mathbf{b}_{j,m,\ell}$ is the complex-valued symbol vector transmitted by TP m that is intended for some user and satisfies $E[\mathbf{b}_{j,m,\ell}\mathbf{b}_{j,m,\ell}^{\dagger}] = \mathbf{I}$. The transmit precoding matrix $\mathbf{W}_{j,m}$ is constrained (as per LTE) to remain fixed across all REs in RBG j and also satisfy a power constraint ($\mathrm{tr}(\mathbf{W}_{j,m}^{\dagger}\mathbf{W}_{j,m}) \leq P_m$ for some given $P_m > 0$). The row (column) dimension of $\mathbf{b}_{j,m,\ell}$ ($\mathbf{W}_{j,m}$) is the number of data symbol streams transmitted on each RE in RBG j and is referred to as the *rank* assigned to the intended user. As per another important LTE constraint, this rank must be invariant across all RBGs assigned to that user. In this work, we allow only one TP to serve a scheduled user in any sub-frame, which precludes joint transmission (JT) wherein a user can receive data from more than one TP on any RBG. In addition, we also preclude frequency selective dynamic point selection (DPS) (unlike [11]), wherein a user can receive data from different TPs on different RBGs. Our contention is that this is indeed a wise choice because for both these schemes ensuring a reliable channel estimation at the user end prior to demodulation is very challenging. This assertion holds since for both these schemes a quasi co-location indication that would allow a user to obtain reliable channel estimates by reusing some parameters that are estimated using dense cell-specific reference signals, cannot be correctly sent [2]. Moreover, in the case of JT the available feedback is insufficient to exploit coherent combining gains. Note then that due to the broadcast nature of the wireless channel, the signal intended for some user transmitted by some TP on an RBG, is also received as interference by all other users co-scheduled on that RBG. Thus, in order to coordinate signal transmissions the cluster scheduler requires CSI about all relevant channel and interference covariance matrices. In an FDD system the CSI must be obtained via over-the-air feedback signalling sent from the users in the cluster. This necessitates an estimation followed by quantization and feedback procedure at each user that is described next.

III. CSI Feedback Procedure

In networks conforming to the LTE standard, the CSI feedback procedure must be accomplished using a two step method [3] described below:

In the first step a *CoMP set* is configured for each user in the cluster. In particular, for each user $k \in \mathcal{U}$ a set $\mathcal{C}_k \subseteq \mathcal{M} : \ 1 \leq |\mathcal{C}_k| \leq 3$ is configured and referred to as the *CoMP set* for user k. This set is held fixed for several frames (several tens of milliseconds). It includes the TP in the cluster from which user k receives the highest average signal power, denoted here by a_k and referred to as the *anchor TP* of user k. In addition, the CoMP set can include up-to two other TPs in the cluster from whom that user receives the second and third highest average signal powers, respectively, provided those powers are greater than a (pre-set) fraction of the highest average received signal power. The main motivation of configuring such a CoMP set is that the

[1] Notice that the model in (1) holds for the orthogonal frequency-division-multiplexing-based downlink if the maximum signal propagation delay is within the cyclic prefix.

TPs in that set represent those in the cluster from which the user can receive a signal with a relatively significant average power. Thus, on any RBG on which user k is scheduled, the scheduler must coordinate the transmissions by the TPs in $\mathcal{C}_k$ to ensure a received signal quality close to the intended one at user k. The transmissions from the remaining TPs in the cluster can be regarded to have much lesser impact. As a result, we express the model in (1) as

$$\mathbf{y}_{j,k,\ell} \quad = \quad \sum_{m \in \mathcal{C}_k} \mathbf{H}_{j,k,m} \mathbf{t}_{j,m,\ell} + \boldsymbol{\eta}_{j,k,\ell}, \tag{3}$$

where the vector $\boldsymbol{\eta}_{j,k,\ell}$ captures the additive noise as well as the out of CoMP set interference. Accordingly, each user $k \in \mathcal{U}$ sends CSI feedback only for TPs in its CoMP set $\mathcal{C}_k$. Note here that the set of all users $\mathcal{U}$ can be partitioned into two parts $\mathcal{U} = \mathcal{U}^{\text{non.comp}} \cup \mathcal{U}^{\text{comp}}$, such that a user k belongs to $\mathcal{U}^{\text{non.comp}}$ ($\mathcal{U}^{\text{comp}}$) if and only if its CoMP set size is equal to (greater than) one.

In the second step, one or more *CSI-processes* are configured for each user k (again once in several frames). Each CSI-process indicates a set of REs (that is periodic in both time and frequency) on which pilots from a TP in $\mathcal{C}_k$ are sent. The periodicity in time (frequency) is decided a-priori based on the coherence time (bandwidth). In addition, that CSI-process also includes another set of REs (having the same respective time and frequency periodicities) for interference measurement. The latter REs are referred to as interference measurement resources (IMRs). We note that since these sets of indicated REs are periodic, they can each be viewed as being formed by repeating a base pattern of REs over time-frequency windows. Over each such window, the user estimates a channel matrix (using pilots) and and an interference covariance matrix (using IMRs) in order to determine its CSI report (corresponding to that window of that CSI-process), which it then feeds back via over-the-air signalling.

A. Configuring and Utilizing CSI processes

We now describe two ways of configuring and utilizing these CSI-processes that are evaluated in this paper.

Class A: For each user k, one CSI-process is configured for each TP $m \in \mathcal{C}_k$ in its CoMP set in the following manner. Let us focus on the CSI feedback that is determined under that process over a time-frequency window. Suppose that pilots from TP m are sent on an RE ℓ' of RBG j' within that window. Then, the received observation at user k is given by

$$\mathbf{y}_{j',k,\ell'} \quad = \quad \mathbf{H}_{j',k,m} \mathbf{s}_{j',m,\ell'} + \boldsymbol{\eta}_{j',k,\ell'}, \tag{4}$$

where $\mathbf{s}_{j',m,\ell'}$ denotes the pilot symbol sent from TP m. Note that all other TPs in the CoMP set $\mathcal{C}_k$ are muted on that RE in order to ensure a reliable channel estimation by the user. Next, the received observation at another RE ℓ'' in that RBG that is configured for interference measurement, is given by

$$\mathbf{y}_{j',k,\ell''} \quad = \quad \boldsymbol{\eta}_{j',k,\ell''}, \tag{5}$$

where we note that all TPs in the CoMP set $\mathcal{C}_k$ are muted on that RE in order to let the user measure out of CoMP set interference. From the observations in (4) and (5) the user can obtain the channel and interference covariance estimates $\mathbf{H}^{\text{est.}}_{j',k,m}$ and $\Sigma^{\text{est.out.}}_{j',k,\ell''}$, respectively. The user then obtains a whitened channel estimate $(\Sigma^{\text{est.out.}}_{j',k,\ell''})^{-1/2}\mathbf{H}^{\text{est.}}_{j',k,m}$, quantizes it and sends the quantized estimate $\boldsymbol{\mathcal{Q}}((\Sigma^{\text{est.out.}}_{j',k,\ell''})^{-1/2}\mathbf{H}^{\text{est.}}_{j',k,m})$ as its feedback. The quantized estimate is represented by a rank value which indicates the number of preferred channel directions (or modes) and identifies the quantization codebook used, as well as a precoder from that codebook whose columns indicate preferred directions, and a corresponding gain value for each such direction. This process is repeated for all time-frequency windows and for each CSI-process.

Let us now consider how all the received feedback is employed by the cluster scheduler. At the scheduler, the feedback received from all users is used to approximate the received observations in (3). In particular, for each RBG j the received observations are approximated as

$$\mathbf{y}_{j,k,\ell} \quad \approx \quad \sum_{m \in \mathcal{C}_k} \hat{\mathbf{H}}_{j,k,m} \mathbf{t}_{j,m,\ell} + \hat{\boldsymbol{\eta}}_{j,k,\ell}, \tag{6}$$

where $\hat{\mathbf{H}}_{j,k,m}, \ \forall \ k \in \mathcal{U}, m \in \mathcal{C}_k$ are the re-constructed whitened channel estimates and $E[\hat{\boldsymbol{\eta}}_{j,k,\ell}\hat{\boldsymbol{\eta}}^{\dagger}_{j,k,\ell}] = \mathbf{I}$. [2] This model is then used to an obtain an estimate of the weighted rate of each user scheduled on RBG j under any tentative scheduling decision. In order to detail the estimation of these weighted rates, we let $x_{j,k,m,r} \in \{0,1\}$ denote an indicator variable that is one if user $k \in \mathcal{U}$ is scheduled to receive data from a TP $m \in \mathcal{C}_k$ on an RBG $j \in \mathcal{N}$ and assigned rank $r \in \mathcal{R}_{k,m}$, and is zero otherwise. $\mathcal{R}_{k,m}$ denotes the set of ranks that can be assigned to user k when it is served by TP m. We collect all such variables corresponding to an RBG j into a vector $\mathbf{X}_{(j)} = [x_{j,k,m,r}]_{k \in \mathcal{U}, m \in \mathcal{C}_k, r \in \mathcal{R}_{k,m}}$. Note that for each RBG j and any given $\mathbf{X}_{(j)}$, we can see from the model in (6) that we have a Gaussian interference channel (GIFC) where the ranks assigned to all scheduled users is given. Then, an estimate of the weighted rate that can be obtained on RBG j for each user k under any tentative scheduling decision $\mathbf{X}_{(j)}$ (denoted here by $R(\mathbf{X}_{(j)}, j, k, m, r)$), can for example be determined using the relation

$$R(\mathbf{X}_{(j)}, j, k, m, r) = \\ \log \left| \mathbf{I} + o_k x_{j,k,m,r} \mathbf{M}_{j,k,m,r} (\mathbf{M}_{j,k,m,r})^{\dagger} \mathbf{C}^{-1}_{j,k,m} \right|,$$

$\mathbf{C}_{j,k,m} = \mathbf{I} + \sum_{k',m',r':k' \neq k} x_{j,k',m',r'} \mathbf{M}_{j,k',m',r'} \mathbf{M}^{\dagger}_{j,k',m',r'}$ and $\mathbf{M}_{j,k',m',r'} = \hat{\mathbf{H}}_{j,k',m'} \mathbf{W}_{j,k',m',r'}$. $\mathbf{W}_{j,k',m',r'}$ denotes a transmit precoder that has r' columns and satisfies the norm constraint $\text{tr}(\mathbf{W}^{\dagger}_{j,k',m',r'} \mathbf{W}_{j,k',m',r'}) \leq P_{m'}$. Note that any algorithm (specified a-priori) to compute these transmit precoding matrices (given the assigned ranks and the power constraints) can be employed. For instance, such an algorithm can be based on SLNR criterion [13], interference alignment [14], or can be an iterative one to optimize the weighted sum rate [9]. o_k denotes the outer-loop-link-adaptation (OLLA) factor associated with user k. These factors

[2] These reconstructed estimates are based on most recently received CSI reports corresponding to CSI processes configured for user k and TPs $m \in \mathcal{C}_k$.

are essential to compensate for feedback inaccuracy and are updated based on the ACK/NACK received from that user. Thus, $\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} R(\mathbf{X}_{(j)},j,k,m,r)$ denotes the estimated weighted sum rate for RBG j under $\mathbf{X}_{(j)}$.

Class B: This approach is geared towards exploiting coordinated muting, thereby simplifying the rate computation. In particular, for each user $k \in \mathcal{U}$ one CSI-process for each TP $m \in \mathcal{C}_k$ in its CoMP set is configured exactly as in the previous approach. In addition, for each user $k \in \mathcal{U}^{\text{comp}}$, and each TP $m \in \mathcal{C}_k$, a second CSI-process is configured, under which the feedback is determined over a time-frequency window that includes some RBG j' as follows. The received observations at user k on any RE ℓ' within that RBG, on which pilots from TP m are sent are given as before by (4). However, the received observation at another RE ℓ''' in that RBG that is configured for interference measurement is now given by

$$\mathbf{y}_{j',k,\ell'''} = \sum_{m'\in\mathcal{C}_k:m'\neq m} \mathbf{H}_{j',k,m'}\mathbf{t}_{j',m',\ell'''} + \boldsymbol{\eta}_{j',k,\ell'''}, \quad (7)$$

where we note that all TPs in the CoMP set $\mathcal{C}_k$ (other than m) transmit on that RE in order to let the user measure total (intra CoMP set and out of CoMP set) interference. From the observations in (4) and (7) the user can obtain the channel and interference covariance estimates, respectively. The user then quantizes and feeds back the whitened channel estimate and the process is repeated for all time-frequency windows.

Let us now consider how all the received feedback is employed by the cluster scheduler in this case. Here, the scheduler directly relies on the user feedback in order to obtain weighted rate estimates instead of constructing an approximate GIFC model. In particular, for each CoMP user $k \in \mathcal{U}^{\text{comp}}$ the weighted rate estimate is obtained as $R(\mathbf{X}_{(j)},j,k,m,r) = x_{j,k,m,r}R^{No.Int.}(j,k,m,r)$, when no user is scheduled (as per $\mathbf{X}_{(j)}$) on RBG j by any TP $m' \neq m$ that also lies in the CoMP set $\mathcal{C}_k$. $R^{No.Int.}(j,k,m,r)$ is an estimate of the weighted rate that can be achieved by user k if it receives data from TP m (with rank r) on RBG j in the absence of intra-cluster interference. Notice that $R^{No.Int.}(j,k,m,r)$ is directly computed using only the feedback from the first CSI process configured for user k and TP m. On the other hand, when at-least one other TP in its CoMP set is active (i.e., is scheduled as per $\mathbf{X}_{(j)}$ to serve some other user on RBG j) then the weighted rate estimate is obtained as $R(\mathbf{X}_{(j)},j,k,m,r) = x_{j,k,m,r}R^{Int.}(j,k,m,r)$, where $R^{Int.}(j,k,m,r)$ is directly computed using only the feedback from the second CSI process configured for user k and TP m. Finally, notice that only one CSI process is configured for each non-CoMP user $k \in \mathcal{U}^{\text{non.comp}}$. Then, the weighted rate estimate for such a user is directly computed using only the feedback from its configured CSI process, which we express as $R(\mathbf{X}_{(j)},j,k,m,r) = x_{j,k,m,r}R^{No.Int.}(j,k,m,r)$. Further, since $|\mathcal{C}_k| = 1$ for such a user, for convenience in exposition and with some abuse of our adopted convention, we also set $R^{Int.}(j,k,m,r) = R^{No.Int.}(j,k,m,r)$ so that $R(\mathbf{X}_{(j)},j,k,m,r) = x_{j,k,m,r}R^{Int.}(j,k,m,r)$ for all $k \in \mathcal{U}^{\text{non.comp}}$.

Comparing the two approaches, we see that the class B one configures more CSI processes to obtain more feedback. It uses these CSI reports to determine weighted rate estimates for each user (under any scheduling decision) based solely on that user's feedback. This aspect avoids the estimation of such rates based on determining transmit precoding matrices using the approximate models as in the class A approach, thereby reducing implementation complexity. Indeed, in the class-B approach the transmit precoder selected by a TP on any RBG, simply follows the recommendation made by the user it seeks to serve. In particular, the transmit precoder employed is either the zero matrix if no user is served, or it is (up-to scaling and column subset selection) one of the at-most two matrices recommended by the user depending on whether or not any of the other TPs in that user's CoMP set are active on that RBG. Hence, we can view the transmit precoding in the class-B approach as more akin to coordinated muting. In Table I we summarize the notation that is employed in this paper.

B. Influencing CSI estimation

Injecting Isotropic Interference An important aspect of influencing the CSI that is estimated by each user (according to its configured CSI processes), that applies to both the aforementioned approaches, is that of injecting interference on the IMRs. In particular, the out of CoMP set interference in (5) and the total interference in (7) can be controlled by the network. We consider two ways of control in this work. The first one is where each interfering TP (from the measuring user's perspective) simply transmits data on the IMRs, to the user that it is serving on that RBG. We note here that the signal transmitted by each such TP on an IMR is generated using the same precoder that is used to generate signals on all other REs in the RBG that includes that IMR. The second one is where each interfering TP transmits an *isotropically distributed signal on each IMR*, so that the covariance matrix of the signal transmitted on any IMR is a scaled identity matrix. The key advantage in the latter case is that the feedback that is sent is no longer dependent on the precoders that were being employed for data transmission by the interferers during the measurement instance. *Our results reveal that the latter method of controlling interference is essential.*

Rank Restriction As mentioned before, the quantized feedback reported by each user under each of its configured CSI processes, includes a rank value which indicates the number of preferred channel directions (or modes) being quantized. Since the feedback load remains fixed across all ranks, the quantization error is smaller for lower ranks since a fewer number of modes need to be quantized there. A useful feature provided in the LTE standard is one where a rank restriction can be independently imposed on each CSI process. For instance, by configuring a user to use only lower rank values, the scheduler can obtain the first few dominant modes of the channel with higher accuracy albeit at the expense of not obtaining the remaining ones. Furthermore, in LTE a CSI process configured for a user can be defined as a reference rank process for one or more of the other CSI processes configured for that user. In that case the user will first determine a rank for obtaining the quantized feedback corresponding to the

reference process, and then use the same rank for those other processes. Finally, recall that $\mathcal{R}_{k,m}$ denotes the set of ranks that can be assigned to user user k when it is served by TP m. Our formulation allows full freedom in choosing this set.

IV. Problem Formulation

We introduce our optimization problem in (8), which seeks to maximize the weighted sum-rate system utility. Note that the first constraint in (8) enforces that on each RBG, each TP can serve at-most one user. The second and the third constraints together enforce that each user can be assigned only one rank and can receive data from only one TP across all the RBGs on which it is scheduled. We remind here that the constraint of allowing only one rank per scheduled user is required in LTE. In summary, the formulation in (8) permits coordinated precoding (that subsumes coordinated muting) and frequency non-selective DPS, which are the most promising CoMP schemes. Notice here that for any user k we can choose to impose an additional constraint $R(\mathbf{X}_{(j)}, j, k, m, r) = 0,\ \forall\ j, r$ for each $m \in \mathcal{C}_k : m \neq a_k$ and doing so will enforce that such a user can only be served by its anchor TP when it is scheduled.

The problem in (8) is a discrete combinatorial optimization problem. Unfortunately, the problem even in the single TP per cluster ($M = 1$) case is NP-hard (cf. [15]). For general $M > 1$, it can be shown that the problem in (8) can be even hard to approximate. Then, to efficiently obtain a good sub-optimal solution, we exploit a useful observation that a related sub-problem can be approximately solved in a simple manner. Towards this end, suppose $\hat{\mathbf{X}}_{(j)} = [\hat{x}_{j,k,m,r}]_{k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}},\ \forall\ j \in \mathcal{N}$ denotes a tentative feasible scheduling decision. Our aim is to improve the tentative decision at hand and obtain another feasible decision $\tilde{\mathbf{X}}_{(j)} = [\tilde{x}_{j,k,m,r}]_{k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}},\ \forall\ j \in \mathcal{N}$. For tractability in obtaining such an improvement, we impose the restriction that no assignment made in the decision $\hat{\mathbf{X}}_{(j)}\ \forall\ j \in \mathcal{N}$ can be changed and that on each RBG at-most one more assignment can be made while retaining feasibility. To summarize, we consider the sub-problem in (9).

A. Scheduling Algorithm

Our algorithm (referred to as the Iterative Submodular Algorithm (ISA)) obtains a solution to (8) by repeatedly solving the sub-problem in (9). In particular, each outer iteration of the ISA algorithm (i.e., all the steps within the While-do loop) enhances the tentative scheduling decision, $\hat{\mathbf{X}}_{(j)},\ \forall\ j \in \mathcal{N}$ at hand. The enhancement is done by solving the sub-problem in (9) via a *greedy* sub-routine implemented in the inner iterations (comprising of all the steps within the Repeat-until loop). Notice here that Θ_j denotes the weighted sum rate achieved on RBG j forall $j \in \mathcal{N}$, i.e., $\Theta_j = \sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} R(\hat{\mathbf{X}}_{(j)}, j, k, m, r)\ \forall\ j$. Each $g(j, k, m, r)$ denotes the weighted sum rate that can be obtained on RBG j by making an additional assignment (that of scheduling user k to receive data from TP m on RBG j with rank r), and where we set $g(j, k, m, r) = 0$ if the additional assignment is either infeasible (given the ones already made) or is not a new assignment (i.e., such an assignment has already been made). The outer iterations are terminated once no improvement can be made or if a maximum number of iterations has been reached. We can show that the enhancement done in each outer iteration to the solution at hand has the following optimality.

Theorem 1: At each outer-iteration, given a feasible solution $\hat{\mathbf{X}}_{(j)},\ \forall\ j \in \mathcal{N}$, the enhanced solution $\tilde{\mathbf{X}}_{(j)},\ \forall\ j \in \mathcal{N}$ obtained via the greedy sub-routine is within $1/2$ of best permissible enhancement, i.e., the weighted sum rate yielded by $\tilde{\mathbf{X}}_{(j)},\ \forall\ j \in \mathcal{N}$ is no less than half of that yielded by the optimal solution to (9).

Proof 1: Consider the problem in (9) for a given feasible solution $\hat{\mathbf{X}}_{(j)},\ \forall\ j \in \mathcal{N}$. Recalling the definition of $g(j, k, m, r)$ (12) and (13) for each choice of $j \in \mathcal{N}, k \in \mathcal{U}, m \in \mathcal{C}_k$ and $r \in \mathcal{R}_{k,m}$, we can re-formulate (9) as (10). Next, we define the set $\breve{\mathcal{U}} = \{k \in \mathcal{U}^{\text{non.comp.}} : |\mathcal{R}_{k,a_k}| = 1\}$ that includes all non CoMP users for whom only one rank can be assigned. Then, let $\tilde{\mathcal{U}} = \mathcal{U} \setminus \breve{\mathcal{U}}$ denote the remaining set of users. The key observation is that each user in set $\breve{\mathcal{U}}$ can be served by only its respective anchor TP under one rank when scheduled on any RBG. Consequently, we can first determine the best possible assignment involving only such users for each RBG j. After that we can consider the remaining users to obtain further improvements. Put differently, defining $\hat{\Theta}_j = \max\{\Theta_j, \max_{k\in\breve{\mathcal{U}},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}\{g(j, k, m, r)\}\}$ we can obtain an equivalent version of (10) given by (11). Notice that in the optimization problem in (11) we can regard the pair (m, r) as one format so that the second and third constraints enforce that each scheduled user must be assigned one format. Further, without loss of optimality, we can set $\tilde{z}_{j,k,m,r} = 0$ for each $j, k, m, r :\ g(j, k, m, r) < \hat{\Theta}_j$. Then, the resulting problem is in-fact equivalent to a single-cell downlink scheduling problem considered in [16] in which at-most one user can be scheduled on each sub-carrier (the first constraint in (11)) and that each scheduled user must be assigned one format across all its assigned subcarriers. Invoking the results developed in [16], we can then deduce that (11) is equivalent to the maximization of a normalized non-negative non-decreasing submodular set-function over a partition matroid and thus can be solved to within $1/2$ optimality by a greedy algorithm [17]. The greedy sub-routine of the ISA is indeed that greedy algorithm. Consequently, we can assert that the theorem is true.

Remark 1: Notice that an alternate design of ISA would be to use the greedy method directly on the problem in (10). Compared to that, our design implements the greedy method after a *hot start* on the one in (11) instead. The intuition is that since the best possible assignments over users in $\breve{\mathcal{U}}$ can be readily determined (in a decoupled manner across the RBGs), we can apply the greedy method on a relatively smaller part of the original problem in (10), thereby reducing the impact of its sub-optimality. Thus, while both approaches provide an identical worst-case guarantee of $1/2$, we observed that the method detailed here yields a better average case performance.

$$\max_{\substack{x_{j,k,m,r},y_{k,m,r}\in\{0,1\}\\ j\in\mathcal{N},k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} R(\mathbf{X}_{(j)},j,k,m,r)\right\}$$
$$\text{s.t.} \sum_{\substack{k\in\mathcal{U}:\\ m\in\mathcal{C}_k}}\sum_{r\in\mathcal{R}_{k,m}} x_{j,k,m,r}\le 1,\ \forall\, j,m;\ \ y_{k,m,r}\ge \max_{j\in\mathcal{N}}\{x_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} y_{k,m,r}\le 1,\ \forall\, k; \tag{8}$$

$$\max_{\substack{\tilde{x}_{j,k,m,r},\tilde{y}_{k,m,r}\in\{0,1\}\\ j\in\mathcal{N},k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} R(\tilde{\mathbf{X}}_{(j)},j,k,m,r)\right\}$$
$$\text{s.t.} \sum_{\substack{k\in\mathcal{U}:\\ m\in\mathcal{C}_k}}\sum_{r\in\mathcal{R}_{k,m}} \tilde{x}_{j,k,m,r}\le 1,\ \forall\, j,m;\ \ \tilde{x}_{j,k,m,r}\ge \hat{x}_{j,k,m,r},\ \forall\, j,k,m,r; \tag{9}$$
$$\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{x}_{j,k,m,r}\le 1+\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \hat{x}_{j,k,m,r},\ \ \forall\, j$$
$$\tilde{y}_{k,m,r}\ge \max_{j\in\mathcal{N}}\{\tilde{x}_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{y}_{k,m,r}\le 1,\ \forall\, k;$$

$$\sum_{j\in\mathcal{N}}\Theta_j+\max_{\substack{\tilde{z}_{j,k,m,r},\tilde{y}_{k,m,r}\in\{0,1\}\\ j\in\mathcal{N},k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{z}_{j,k,m,r}(g(j,k,m,r)-\Theta_j)\right\}$$
$$\text{s.t.} \sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{z}_{j,k,m,r}\le 1,\ \ \forall\, j;\ \ \tilde{y}_{k,m,r}\ge \max_{j\in\mathcal{N}}\{\tilde{z}_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{y}_{k,m,r}\le 1,\ \forall\, k; \tag{10}$$

$$\sum_{j\in\mathcal{N}}\hat{\Theta}_j+\max_{\substack{\tilde{z}_{j,k,m,r},\tilde{y}_{k,m,r}\in\{0,1\}\\ j\in\mathcal{N},k\in\tilde{\mathcal{U}},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\tilde{\mathcal{U}}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{z}_{j,k,m,r}(g(j,k,m,r)-\hat{\Theta}_j)\right\}$$
$$\text{s.t.} \sum_{k\in\tilde{\mathcal{U}}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{z}_{j,k,m,r}\le 1,\ \ \forall\, j;\ \ \tilde{y}_{k,m,r}\ge \max_{j\in\mathcal{N}}\{\tilde{z}_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} \tilde{y}_{k,m,r}\le 1,\ \forall\, k; \tag{11}$$

B. Post-processing

One drawback of the ISA due to its successive nature is that it can get stuck on poor local optimal on one or more RBGs, which makes post-processing essential. We adopt a simple post-processing stage. For each user k that is scheduled (i.e., assigned at-least one RBG) by the ISA we let $(\hat{r}_k,\hat{m}_k)$ denote the assigned rank and serving TP. Notice that for each scheduled user there will be only one assigned rank and serving TP and moreover, for each scheduled non-CoMP user (i.e., each scheduled user lying in $\mathcal{U}^{\text{non.comp}}$) the assigned serving TP will necessarily be its anchor TP. For all other users (i.e., those not assigned any RBG), let us fix $\hat{m}_k=a_k$, where we recall that a_k denotes the anchor TP of user k and set $\hat{r}_k=1$ (or some default choice such as a rank reported by user k for its anchor a_k). Then, considering the class-A approach, on each RBG j we check if scheduling the best set of non-CoMP users would yield a higher weighted sum rate, i.e., if

$$\sum_{m\in\mathcal{M}}\max_{k\in\mathcal{U}^{\text{non.comp}}:\hat{m}_k=m}\{R^{Int.}(j,k,\hat{m}_k,\hat{r}_k)\}>\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}} R(\hat{\mathbf{X}}_{(j)},j,k,m,r) \tag{15}$$

Notice that in (15) we have used the fact that even in the class-A approach we can express $R(\mathbf{X}_{(j)},j,k,m,r)=x_{j,k,m,r}R^{Int.}(j,k,m,r),\ \forall\ k\in\mathcal{U}^{\text{non.comp}}$ given that only non CoMP users are scheduled on RBG j under any tentative decision $\mathbf{X}_{(j)}$. On the other hand, for the class-B approach we check the condition in (15) but after replacing $\mathcal{U}^{\text{non.comp}}$ with $\mathcal{U}$. Then, if the respective conditions are satisfied, we accordingly change the decision $\hat{\mathbf{X}}_{(j)}$. Note also that this simple update can be done separately for each RBG and improves the system utility while preserving feasibility.

C. Benchmarking

Over simple networks we can benchmark the ISA against an upper bound obtained via brute-force search after ignoring the constraints of assigning each scheduled user one rank and/or one serving TP. Moreover, in the scenario where the class B approach is adopted, we can re-formulate (8) as the integer linear program (ILP) in (16). Upon relaxing the binary value constraints in (16), we obtain an LP. The latter LP can be optimally solved to obtain an upper-bound to (16) and benchmark the solution yielded by the ISA for this case.

D. User Pre-selection

In a practical HetNet the number of active users in a cluster can be quite large. In such scenarios the control channel constraints may limit the number of users that can be served. In particular, the feedforward signalling (containing the scheduling grant) to each scheduled user is sent from its

$$\max_{\substack{u_{j,k,m,r},v_{j,k,m,r},y_{k,m,r},z_{j,m}\in\{0,1\}\\ j\in\mathcal{N},k\in\mathcal{U},m\in\mathcal{C}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}}\Big(u_{j,k,m,r}R^{No.Int.}(j,k,m,r)+v_{j,k,m,r}R^{Int.}(j,k,m,r)\Big)\right\}$$
$$\text{s.t. } z_{j,m}\geq\sum_{\substack{k\in\mathcal{U}:\\ m\in\mathcal{C}_k}}\sum_{r\in\mathcal{R}_{k,m}}(u_{j,k,m,r}+v_{j,k,m,r}),\ \forall\, j,m;\ \max_{r\in\mathcal{R}_{k,m}}\{u_{j,k,m,r}\}+\max_{m'\in\mathcal{C}_k:m'\neq m}\{z_{j,m'}\}\leq 1,\ \forall\, j,k,m; \tag{16}$$
$$y_{k,m,r}\geq\max_{j\in\mathcal{N}}\{u_{j,k,m,r}+v_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{C}_k}\sum_{r\in\mathcal{R}_{k,m}}y_{k,m,r}\leq 1,\ \forall\, k;$$

$$\max_{\substack{v_{j,k,m,r},y_{k,m,r}\in\{0,1\}\\ j\in\mathcal{N},k\in\mathcal{U},m\in\mathcal{M}_k,r\in\mathcal{R}_{k,m}}} \left\{\sum_{j\in\mathcal{N}}\sum_{k\in\mathcal{U}}\sum_{m\in\mathcal{M}_k}\sum_{r\in\mathcal{R}_{k,m}}\Big(v_{j,k,m,r}R^{Int.}(j,k,m,r)\Big)\right\}$$
$$\text{s.t. } y_{k,m,r}\geq\max_{j\in\mathcal{N}}\{v_{j,k,m,r}\},\ \forall\, k,m,r;\ \sum_{m\in\mathcal{M}_k}\sum_{r\in\mathcal{R}_{k,m}}y_{k,m,r}\leq 1,\ \forall\, k; \tag{17}$$
$$\sum_{k\in\mathcal{U}:a_k=m'}\sum_{m\in\mathcal{M}_k}\sum_{r\in\mathcal{R}_{k,m}}y_{k,m,r}\leq J,\ \forall\, m'\in\mathcal{M}.$$

anchor TP, over the control channel that has a limited capacity. It thus makes sense from a complexity stand-point to pre-select a pool of *good* users in a simple manner and then use the ISA on the selected pool of users. Here we propose one such user pre-selection method which assumes that the class-B approach is adopted. We consider the problem in (17) (which assumes that all TPs are active on all RBGs) where J is a specified user limit for each TP. $\mathcal{M}_k$ is the set of TPs that can serve user $k\in\mathcal{U}$ and we set $\mathcal{M}_k=\mathcal{C}_k\ \forall\, k$. We offer the following result whose proof is skipped due to space constraints.

Proposition 1: The problem in (17) is equivalent to the problem of maximizing a normalized non-negative non-decreasing submodular set function subject to one matroid constraint.

We also consider a reference second method where we set $\mathcal{M}_k=\{a_k\},\ \forall\, k$, upon which (17) reduces to a problem considered in [16] for which Proposition 1 readily holds. Consequently, we employ the classical greedy algorithm [17] on the equivalent reformulations (which in either case offers a $1/2$ approximation) to obtain the respective pre-selection algorithms. From the output of the pre-selection algorithm, the set of users that have each been assigned at-least one RBG is retrieved and used (instead of $\mathcal{U}$) as the input user set to the ISA (the RBG or rank or serving TP assignments made by the pre-selection algorithm are ignored). Notice that to further reduce the complexity of pre-selection, we can appropriately restrict each set $\mathcal{R}_{k,m}$ to be a singleton set.

V. System Simulations

We first evaluated ISA with the class-A approach over a simple network (comprising of one cluster of $M=3$ TPs, $K=15$ users and $N=4$ RBGs). We suppose that each user k is a CoMP user with $|\mathcal{C}_k|=3$ but which can only be served by its anchor TP. Each TP is the anchor TP of 5 users. Each TP has three transmit antennas and each user has three receive antennas. The channel seen by each user from each TP over each RBG is modeled as i.i.d. Rayleigh fading. Each user if scheduled must be served using one of two ranks across all its assigned RBGs. We evaluate the class-A approach where for simplicity we assume ideal CSI feedback from each user. Further we assume a precoding rule in which the rank-1 beamforming vector is always chosen as the first column of the 3×3 DFT matrix, while the rank-2 precoder is always formed by the first two columns of the 3×3 DFT matrix. In Fig. 1 we plot the average system sum rate, where the sum rate is computed across all TPs, all users and all RBGs and the average is computed over five hundred independent fading realizations. We plot the average sum rate obtained upon scheduling using the ISA with $\mathrm{IterMax}=1$ (denoted as ISA-1 in the legend). We also plot the average sum rate obtained upon scheduling under ISA with $\mathrm{IterMax}=2$ and $\mathrm{IterMax}=4$, respectively. Further, we consider the case when each TP employs maximum rank (rank-2) transmission on each RBG. Since this restriction fixes the choice of precoders in each RBG (recall that only one rank-2 precoder is available), on each RBG we can independently determine the best user for each TP. Finally, we compute and plot an upper bound which is obtained by ignoring the at-most one rank per user constraint and then doing a separate exhaustive search for each RBG. Thus, the obtained upper bound sum-rate need not be achievable. Notice from the plot that ISA with $\mathrm{IterMax}=2$ and $\mathrm{IterMax}=4$ settings achieves more than 84% and 88% of the upper bound, respectively. Further, a very large gain is obtained by moving from $\mathrm{IterMax}=1$ to $\mathrm{IterMax}=2$ and two iterations capture almost all of the possible gain, particularly at moderate to high SNRs. Moreover, the maximum rank based scheduling performs well at low SNRs but saturates at high SNRs. Indeed the system will become interference limited under such maximum rank transmission.

We next proceeded to emulate a realistic HetNet defined in [2] for which the main simulation parameters are summarized in Table II. The system simulation results were obtained using a state-of-the-art calibrated simulator that was developed in house. As another sanity check, we also considered the class-B approach and for several instances (with each instance corresponding to a set of user weights and their CSI reports) we compared the weighted sum rate yielded by ISA to that yielded by the LP relaxation of (16). In each instance we found that ISA achieved more than 80% of the upper bound.

Algorithm 1: Iterative Submodular Algorithm (ISA)

1) **Initialize** $\Theta_j = 0,\ \forall\ j,\ \{\hat{x}_{j,k,m,r} = 0\}\ \forall\ j,k,m,r,$ Outer = false and Iter = 1.
2) **While** (Outer = false) and (Iter $\leq$ IterMax) **Do**
3) **For** each $j \in \mathcal{N}, k \in \mathcal{U}, m \in \mathcal{C}_k, r \in \mathcal{R}_{k,m}$, Do
4) **If** $\max_{k',r'}\{\hat{x}_{j,k',m,r'}\} > 0$ or $\max_{m',r'}\{\hat{x}_{j,k,m',r'}\} > 0$ or $\max_{m',j':j'\neq j,r':r'\neq r}\{\hat{x}_{j',k,m',r'}\} > 0$ or $\max_{m':m'\neq m,j':j'\neq j,r'}\{\hat{x}_{j',k,m',r'}\} > 0$ Then set
$$g(j,k,m,r) = 0, \tag{12}$$
5) **Else** update $\hat{x}_{j,k,m,r} = 1$ to compute $g(j,k,m,r)$ as
$$\sum_{k'\in\mathcal{U}}\sum_{m'\in\mathcal{C}_k}\sum_{r'\in\mathcal{R}_{k',m'}} R(\hat{\mathbf{X}}_{(j)}, j, k', m', r') \tag{13}$$
and revert $\hat{x}_{j,k,m,r} = 0$.
6) **EndIf**
7) **EndFor**
8) Set $\tilde{z}_{j,k,m,r} = 0,\ \forall\ j,k,m,r,$ Outer = true and Inner = false.
9) **Repeat**
10) **For** each $j \in \mathcal{N}$, Do
11) Determine
$$(k^*, m^*, r^*) = \arg \max_{k\in\breve{\mathcal{U}}, m\in\mathcal{C}_k, r\in\mathcal{R}_{k,m}} \{g(j,k,m,r)\}$$
12) **If** $g(j,k^*,m^*,r^*) > \Theta_j$ Then
13) Set $\tilde{z}_{j,k^*,m^*,r^*} = 1$, Outer = false and $\Theta_j = g(j,k^*,m^*,r^*)$
14) **EndIf**
15) **EndFor**
16) **For** each $k \in \tilde{\mathcal{U}}, m \in \mathcal{C}_k, r \in \mathcal{R}_{k,m}$, Do
17) Determine
$$h(k,m,r) = \sum_j \max\{0, g(j,k,m,r) - \Theta_j\}$$
18) **EndFor**
19) Determine
$$(k^*, m^*, r^*) = \arg \max_{k,m,r} \{h(k,m,r)\} \tag{14}$$
20) **If** $h(k^*,m^*,r^*) > 0$ Then
21) Set Outer = false
22) **For** each $j \in \mathcal{N}$
23) **If** $g(j,k^*,m^*,r^*) > \Theta_j$ Then
24) Set $\tilde{z}_{j,k^*,m^*,r^*} = 1$, $\Theta_j = g(j,k^*,m^*,r^*)$ and $\tilde{z}_{j,k',m',r'} = 0,\ \forall\ (k',m',r') \neq (k^*,m^*,r^*)$
25) **EndIf**
26) **EndFor**
27) **Else** set Inner = true
28) **EndIf**
29) **Until** Inner = true
30) Update $\hat{\mathbf{X}}_{(j)},\ \forall\ j$ by accepting assignments made in $\{\tilde{z}_{j,k,m,r}\}$
31) **EndWhile**
32) Output $\hat{\mathbf{X}}_{(j)},\ \forall\ j$

We are now ready to offer our system evaluation results. For either class-A or class-B approach, we found that *using OLLA based on user ACK/NACK feedback* as well as *injecting isotropic interference on the IMRs* are absolutely essential to realize CoMP gains over the baseline scheme described in the Introduction. *In the absence of any one of these two features, CoMP results are in-fact uniformly worse (both in average and 5-percentile throughputs) than the baseline scheme.* We remark here that scheduling schemes that entail coordination across a number of TPs to determine their scheduling decisions

TABLE I
SYMBOL DEFINITIONS

$\mathcal{U}$	Set of users with cardinality $\lvert\mathcal{U}\rvert = K$
$\mathcal{N}$	Set of N RBs
$\mathcal{M}$	Set of M TPs
$\mathcal{U}^{\text{comp}}$	Set of CoMP users
$\mathcal{U}^{\text{non-comp}}$	Set of non CoMP users
$\mathcal{C}_k$	CoMP set of user k
$\mathcal{R}_{k,m}$	Set of ranks that can be assigned to user k when served by TP m
$\breve{\mathcal{U}}$	Set of all non CoMP users for whom only one rank can be assigned
$\tilde{\mathcal{U}}$	Set of remaining users obtained as $\mathcal{U} \setminus \breve{\mathcal{U}}$
$x_{j,k,m,r}$	Indicator variable which is 1 when RBG j is assigned to user k under TP m and rank r
$\mathbf{X}_{(j)}$	Vector containing all indicator variables pertaining to RBG j
$y_{k,m,r}$	Indicator variable which is 1 when at-least one RBG is assigned to user k under TP m and rank r
$R(\mathbf{X}_{(j)}, j, k, m, r)$	Estimate of weighted rate achieved for user k on RBG j when served by TP m and rank r under assignment $\mathbf{X}_{(j)}$
$R^{No.Int.}(j,k,m,r)$	Direct feedback based estimate of weighted rate of user k on RBG j when served by TP m and rank r without interference
$R^{Int.}(j,k,m,r)$	Direct feedback based estimate of weighted rate of user k on RBG j when served by TP m and rank r with interference

TABLE II
SIMULATION PARAMETERS

Parameter	Value used for evaluation
Network and cell layout	HetNet with multiple clusters, wrap around; There is one macro TP and 4 RRHs uniformly distributed within each cluster.
Carrier frequency	2 GHz
Transmission bandwidth	10M Hz
Channel model	Macro: ITU UMA with user (UE) speed 3 km/Hr; RRH: ITU UMi
Antenna configuration	ULA; Macro/RRH TX antennas=2; UE RX antennas=2
Traffic model	Full buffer
Feedback interval	5 ms
UE Receiver	MMSE-IRC
CSI feedback scheme	Class A and Class B
Transmission scheme	DPS/SU-MIMO
CoMP set threshold	9 dB
UE distribution	Configuration 4b with $\frac{1}{3}$ macro TP UEs and $\frac{2}{3}$ RRH UEs;
Number of UEs	30 UEs per cluster
Scheduling metric	Proportional fair

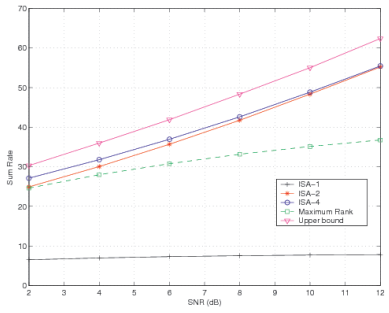

Fig. 1. Sum Rate versus SNR (dB)

TABLE III
THROUGHPUT (KBPS) OF COMP WITH ISA AND CLASS A.

Scheduling scheme	ISA	Baseline
Average	3934 (-1.13%)	3984.2
5−percentile	1090 (35.52%)	804.3

are much more sensitive to the quality of CSI feedback. In contrast to the large gains promised by the ideal CSI results of Fig. 1, in realistic deployments significant optimization in terms of configuring the CSI feedback (while being standards compliant) is need to achieve the 5-percentile throughput gains without sacrificing the average throughput performance. We evaluated the class-A and class-B approaches for several different settings to determine the best ones. The results yielded by the latter settings are presented below for a representative *drop*, denoting a particular realization of user and RRH locations. To assess the variation across drops we also obtained the performance across 20 different drops, where each such drop was generated following the 3GPP evaluation methodology.

A. Class A results

In Table III we present the results yielded by the best setting that we found. In that setting we restrict the rank that can be assigned to any scheduled CoMP user to be one, i.e., we set $\mathcal{R}_{k,m} = \{1\},\ \forall\ m \in \mathcal{C}_k, k \in \mathcal{U}^{\text{comp}}$. No process is defined to be a reference rank process and no rank restriction is imposed on any CSI process. Transmit precoding based on the SLNR criterion is employed. Upon evaluating this setting over twenty different drops we found that the gain over the baseline in terms of the cell average throughput varied from -2.3% to 0.5% with a mean value of -1%. The 5-percentile throughput gain varied from 20.5% to 37.4% with a mean value of 29.7%.

B. Class B results

In Table IV we present the results yielded by the best setting that we found. In that setting, for each CoMP user we define a particular CSI process to be the reference rank process for all other ones, so that each CoMP user also reports only one distinct rank. The process chosen to be the reference is the second CSI process configured for the anchor TP, i.e., the process in which the pilots are sent by the anchor TP and the IMRs include interference from all the other TPs in the CoMP set. Further, we also restrict the rank that can be assigned to any scheduled user to be one it has reported. Upon evaluating this setting over twenty different drops we found that the gain over the baseline in terms of the cell average throughput varied from 1.1% to 3.5% with a mean value of 2.3%. The 5-percentile throughput gain varied from 23.9% to 37.3% with a mean value of 30.6%.

TABLE IV
THROUGHPUT (KBPS) OF COMP WITH ISA AND CLASS B.

Scheduling scheme	ISA	Baseline
Average	4082.8 (2.47%)	3984.2
5−percentile	1079 (34.15%)	804.3

Upon comparing the results presented for these two approaches, we see that the class-B approach has a clear advantage. For the class-B approach we also simulated another example considering an average of 100 users per cluster with $J = 20$ and found that the user pre-selection method 1 performs well and provides an 8% gain in the $5-$percentile SE over pre-selection method 2, while yielding an almost identical average SE.

VI. CONCLUSIONS

We presented a detailed analysis and evaluation of a practical CoMP scheduling algorithm under two standards compliant approaches for obtaining and utilizing channel feedback from the users. We demonstrated that significant CoMP gains can indeed be realized over realistic HetNets.

REFERENCES

[1] A. Ghosh et.al., "Heterogeneous cellular networks: From theory to practice," *IEEE Commun. Mag.*, vol. 50, pp. 54–64, June 2012.
[2] 3GPP, "Coordinated multi-point operation for LTE physical layer aspects (Release 11)," *TR36.819 V11.1.0*, Dec 2011.
[3] J. Lee et.al., "Coordinated multipoint transmission and reception in lte-advanced systems," *IEEE Commun. Mag.*, Nov. 2012.
[4] M. Karakayali et.al., "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Comm. Magazine*, vol. 13, pp. 56–61, Aug. 2006.
[5] D. Gesbert et.al., "Adaptation, coordination, and distributed resource allocation in interference-limited wireless networks," *Proceedings of the IEEE*, vol. 95, pp. 2393–2409, Dec. 2007.
[6] L. Venturino et.al., "Coordinated linear beamforming in downlink multi-cell wireless networks," *IEEE Trans. Wireless Commun.*, 2010.
[7] D. Gesbert et.al., "Multi-cell MIMO cooperative networks: A new look at interference," *IEEE J. Select. Area Commun.*, vol. 28, pp. 1380 –1408, Dec. 2010.
[8] W. Yu et.al., "Multicell coordination via joint scheduling, beamforming and power spectrum adaptation," in *Proc. IEEE INFOCOM*, pp. 2570–2578, Apr. 2011.
[9] T. Bogale et.al., "Weighted sum rate optimization for MIMO coordinated base station systems: Centralized and distributed algorithms," *IEEE Trans. Sig. Proc.*, Dec. 2011.
[10] M. Sanjabi et.al., "Optimal joint base station assignment and beamforming for heterogeneous networks," *IEEE Trans. on Sig. Proc.*, Apr. 2014.
[11] N. Prasad et.al., "Coordinated resource allocation over heterogenous wireless networks:," in *Proc. 2013 IEEE Globecom*, Dec. 2013.
[12] B. Clerckx et.al., "A practical cooperative multicell MIMO-OFDMA network based on rank coordination," *IEEE Trans. on Wireless Comm.*, Apr. 2013.
[13] M. Sadek et.al., "A leakage-based precoding scheme for downlink multi-user MIMO channels," *IEEE Trans. Wireless Commun.*, vol. 6, pp. 1711–1721, May 2007.
[14] C. M. Yetis et.al., "On feasibility of interference alignment in MIMO interference networks," *IEEE Trans. Signal Processing*, vol. 58, pp. 4771–4782, sept. 2010.
[15] S. Lee et.al., "Downlink MIMO with frequency-domain packet scheduling for 3GPP LTE," *Proc. IEEE INFOCOM*, 2009.
[16] H. Zhang et.al., "MIMO downlink scheduling in LTE and LTE-advanced systems," in *IEEE INFOCOM*, 2012.
[17] G. L. Nemhauser et.al., "Best algorithms for approximating the maximum of a submodular set function," *Math. Operations Research*, 1978.