Utility Optimization in Heterogeneous Networks via CSMA-Based Algorithms

Matthew Andrews
Alcatel-Lucent Bell Labs, Murray Hill, NJ
andrews@research.bell-labs.com

Lisa Zhang
Alcatel-Lucent Bell Labs, Murray Hill, NJ
ylz@research.bell-labs.com

Abstract—We study algorithms for carrier and rate allocation in cellular systems with distributed components such as a heterogeneous LTE system with macrocells and femtocells. Existing work on LTE systems often involves centralized techniques or requires significant signaling, and is therefore not always applicable in the presence of femtocells. More distributed CSMA-based algorithms (carrier-sense multiple access) were developed in the context of 802.11 systems and have been proven to be utility optimal. However, the proof typically assumes a single transmission rate on each carrier. Further, it relies on the CSMA collision detection mechanisms to know whether a transmission is feasible.

In this paper we present a framework for LTE scheduling that is based on CSMA techniques. In particular we first prove that CSMA-based algorithms can be generalized to handle multiple transmission rates in a multi-carrier setting while maintaining utility optimality. We then show how such an algorithm can be implemented in a heterogeneous LTE system where the existing Channel Quality Indication (CQI) mechanism is used to decide transmission feasibility.

I. INTRODUCTION

Interference mitigation is a fundamental problem in wireless networks. The exact method for handling interference depends on the nature of the network, e.g. whether it is a centrally controlled cellular network or a more unstructured ad-hoc network. For cellular networks interference can be mitigated via techniques such as power-control, frequency reuse, and fine-grained rate control based on channel-quality measurements together with some aspect of central planning. On the other hand, for ad-hoc networks, especially those running the 802.11 protocol, interference is typically mitigated by a distributed collision-based random access scheme, perhaps coupled with a fairly coarse-grained rate-adaptation procedure.

In this paper we are concerned with interference mitigation in cellular systems with distributed components such as heterogeneous 4G LTE systems that include small cells. Small cells are basestations that aim to provide high data rate coverage over a small high-traffic area. For example, *picocells* are owned by a cellular provider and placed on public locations such as lamp posts. Alternatively, *femtocells* are owned by an end-user with the aim of improving coverage in a private home or business. An important property of femtocells is that they can operate in Closed Subscriber Group (CSG) mode in which the basestation restricts the set of mobile terminals that can connect to it. Another effect introduced by femtocells is that interference to macrocell users can now come from a femtocell

in the interior of the macrocell itself, not just from neighboring macrocells.

LTE networks with small cells represent a hybrid of traditional cellular networks and traditional ad-hoc networks. On the one hand, basestations are running the full LTE protocols which allows for the many interference mitigation schemes that these protocols provide. On the other hand, the placement of picocells and femtocells in an LTE network is likely to be unstructured and so the interference configurations are likely to resemble a typical ad-hoc configuration. As a consequence, there is no hope for any centralized planning, which is a possibility for cellular network interference mitigation techniques such as frequency reuse. We therefore need distributed algorithms.

We are interested in scheduling algorithms for a heterogeneous system that consists of a mixture of macrocells and small cells. We wish to determine the channels, or carriers, used by each basestation as well as the transmission rate on each channel. This should be done in order to maximize a utility function associated with the system. Although a number of scheduling algorithms have been proposed in the LTE context, many of them require a non-trivial amount of signaling among the transmitters. For example, in some algorithms a scheduling decision is preceded by a calculation of how the decision would affect the overall system utility, e.g. by exchanging partial derivative information between neighboring transmitters. This is difficult to support in heterogeneous networks with small cells due to the complexity of setting up the necessary communication channels.

Our main result is to show that LTE scheduling in heterogeneous networks can be performed using techniques developed in the context of 802.11 networks. These networks utilize a Carrier-Sense Multiple Access (CSMA) protocol, often enhanced with a Request-to-Send/Clear-to-Send (RTS/CTS) mechanism [11]. In this setup, transmitters sense the channel before transmitting and proceed only if no conflicting transmissions are active.

The attractiveness of this framework is that CSMA scheduling algorithms can achieve optimal throughputs without any explicit signaling. Coordination is implicit in the "collision" mechanism defined by the CSMA mechanism. In particular, Jiang and Walrand showed in [10] that such mechanisms can be used to achieve any set of feasible throughputs (i.e. they are *throughput-optimal*). Among the sequence of papers that

followed, Liu et al.[14] presented utility-optimal algorithms for CSMA networks with a single carrier and single transmission rate. Two subsequent papers discussed implementation issues associated with these algorithms [13] and presented an extension to the case of a single transmission rate on multiple carriers [15]. However, to the best of our knowledge no previous work has looked at how CSMA-based techniques can be applied in the LTE context.

A number of issues arise if we are to use CSMA-based scheduling algorithms for LTE. First, LTE networks utilize an OFDM physical layer which consists of multiple transmission rates over multiple carriers realized by multiple adjustable power levels. Hence, in addition to deciding when to transmit on each carrier as in [10], [14], [15], the scheduler now also chooses the power level used and the transmission rate. Second, LTE networks do not have an explicit carrier sense mechanism to detect conflicting transmissions. We need to build this capability via the existing Channel Quality Indication (CQI) mechanism. Lastly, each basestation in an LTE network typically has its own local scheduler, such as Proportional Fair, that governs the transmissions to the users within the cell. We need a mechanism that allows the existing local scheduler to work with the CSMA-based algorithm.

The main result of our paper is a scheduling algorithm for utility maximization in heterogeneous LTE networks. Our methodology is motivated by the CSMA analysis of [14]. We believe our main contribution is in showing that the proof of convergence in [14] can be adapted for the case of multiple per-carrier transmission rates and powers, and (perhaps more importantly) describing how the algorithms can be implemented using the CQI mechanism present in LTE.

Another related paper [16] examines multi-hop MIMO networks under SINR constraints. The model of [16] is similar to ours in that it handles multiple transmission rates, but differs in two ways. First, it does not directly fit into the LTE framework as the focus is on local distributed algorithms for finding feasible sets of transmissions under SINR constraints, which require the ability to broadcast control messages to neighbors. Second, [16] focuses on throughput-optimality and does not directly address whether utility-optimality can be achieved for the case of multiple transmission rates, especially in the setting of [14] where the random access parameters for the system can be updated before the previous system has had a chance to converge.

We begin with a more abstract version of our algorithm in which details of the interference are abstracted away into a *feasibility region* for the transmissions. Many of the existing CSMA algorithms implicitly work with this notion. We initially assume an "oracle" that informs a transmitter whether a potential transmission would disrupt existing transmissions. The later sections of the paper discuss how to realize such a scheme in a heterogeneous LTE system in practice. We structure the paper as follows.

- In Section II we present an abstract model of a multi-carrier system that allows for multiple transmission rates on each carrier. This model assumes that each transmitter knows whether a potential transmission is feasible.
- In Sections III-IV we adapt the Liu et al. single-carrier single-rate utility maximization algorithm for CSMA to the abstract framework of Section II to address the general case of multiple transmission rates on multiple carriers.
- In Sections V and VI we give a concrete model for heterogeneous LTE networks with small cells, in which we address power level and interference directly instead of via the notion of a feasibility region for transmission. We also discuss practical issues such as CQI-based collision detection and incorporating a local scheduler.
- In Section VII we present simulation results.
- In Section VIII we give an overview of past work on scheduling and resource allocation in cellular and 802.11 networks.

II. ABSTRACT MODEL

We begin by describing an abstract model that captures the notion of multiple transmission rates on multiple carriers and variable-power scheduling. We consider a system in which a set of transmitters communicate to a set of receivers via a set of links $\mathcal{L}$ on a set of carriers $\mathcal{C}$ at transmission rates from a set $\mathcal{R}$ of positive numbers. Each link is associated with a transmitter and a receiver, where multiple links may share a common transmitter but each link corresponds one-to-one with a receiver.

We solve a scheduling problem, i.e. at each time instant we specify for every carrier the links that are transmitting on that carrier together with the associated transmission rates. More precisely, we represent a *schedule* on a carrier $c \in \mathcal{C}$ by a vector $(r_0, \ldots, r_{L-1}) \in \mathcal{R}_0^L$ where $l = |\mathcal{L}|$ and $\mathcal{R}_0 = \mathcal{R} \cup \{0\}$. Such a schedule is *feasible* if it can be realized by an appropriate power allocation so that every link $\ell \in \mathcal{L}$ can transmit from its transmitter to its receiver at rate r_ℓ on carrier c simultaneously. A scheduling algorithm describes a schedule for every carrier $c \in \mathcal{C}$ at every time instant. Note that a link is allowed to simultaneously transmit on multiple carriers.

This abstract model captures interference and power assignments by the notion of a *feasibility region*, which consists of all valid schedules. In the next two sections we describe our basic algorithmic framework in this abstract model for which we do not concern ourselves with how the system knows whether a schedule is feasible but simply assume an *oracle* that indicates whether a potential transmission leads to feasibility. As mentioned before, for 802.11 networks this can be approximately realized by CSMA techniques coupled with RTS/CTS messages. In later sections we describe how the oracle can be realized in an LTE heterogeneous network.

We consider the problem of system utility maximization. In particular let $\gamma_{\ell,c}$ indicate the transmission rate on link ℓ on carrier c. For a given concave utility function $U(\cdot)$ we wish to maximize the aggregate utility over all links, i.e. to maximize $\sum_\ell U(\sum_c \gamma_{\ell,c})$. Note that for each link the utility function is applied to the total transmission rate on the link over all

carriers. This coupling between the carriers implies that we cannot simply treat each carrier as an isolated system.

A formal version of the optimization problem is given below. Consider a schedule $m \in \mathcal{N}_c$ where $\mathcal{N}_c$ is the set of feasible schedules on carrier c. If $\pi_m \in [0,1]$ indicates the fraction of time that m occurs and $r_{\ell,m}$ indicates the transmission rate on ℓ under schedule m, then $\gamma_{\ell,c}$ can be viewed as a weighted sum of $r_{\ell,m}$ where π_m serves as the weight. Throughout the paper, notations such as $\vec{\gamma}$ indicate a vector $(\gamma_{\ell,c})_{\ell \in \mathcal{L}, c \in \mathcal{C}}$. We wish to solve:

$$\max \quad f_1(\vec{\gamma}) = \sum_{\ell \in \mathcal{L}} U(\sum_c \gamma_{\ell,c}) \tag{1}$$

$$\text{s.t.} \quad \gamma_{\ell,c} \leq \sum_{m \in \mathcal{N}_c} r_{\ell,m} \pi_m \quad \forall \ell, c$$

$$\sum_{m \in \mathcal{N}_c} \pi_m = 1 \quad \forall c.$$

We can view this as an extension of the formulation of [15] that allows for multirate transmissions (and implicitly variable transmission powers).

In reality the sets $\mathcal{L}$ and $\mathcal{N}_c$ can change over time due to mobility. However, as in common in the literature we assume that this happens on a slow enough timescale that it makes sense to solve the utility maximization problem for the current network configuration.

III. MULTI-CARRIER, MULTI-RATE SCHEDULING ALGORITHM

We first describe a routine RANDACC (in Figure Algorithm III.1), a continuous-time random access algorithm that determines when a transmission will take place. We use $\langle \ell, r, c \rangle$ to denote the transmission on link $\ell \in \mathcal{L}$ on carrier $c \in \mathcal{C}$ at rate $r \in \mathcal{R}$. Each transmission $\langle \ell, r, c \rangle$ is associated with two parameters, $\lambda_{\ell,r,c}$ the channel access rate and $\mu_{\ell,r,c}$ which represents the expected transmission duration. After an exponentially distributed waiting period with mean $1/\lambda_{\ell,r,c}$, RANDACC checks whether $\langle \ell, r, c \rangle$ leads to a valid schedule in $\mathcal{N}_c$ at that time instant. If yes, the transmission starts immediately and lasts for an exponentially distributed time period with mean $\mu_{\ell,r,c}$. Note that an invalid schedule includes the situation in which ℓ conflicts with itself, namely ℓ already transmits on c, or ℓ conflicts with another link on c, both of which are captured by $\mathcal{N}_c$. Note also that since RANDACC operates in a continuous manner, two links make a scheduling decision simultaneously with zero probability and therefore they make conflicting decisions with zero probability.

The algorithm MMUO (multi-carrier multi-rate utility optimization) approximates a solution to (1) as follows (see Algorithm III.2). Time is divided into frames of fixed duration. During each frame f, each potential transmission $\langle \ell, r, c \rangle$ calls the RANDACC routine with parameters $\lambda_{\ell,r,c}[f]$ and $\mu_{\ell,r,c}[f]$.

At the end of each frame, link ℓ calculates the service received during the frame (denoted by $S_\ell[f]$), and updates a virtual queue size parameter (denoted by $q_\ell[f]$) as follows.

$$q_\ell[f+1] = \left[q_\ell[f] + b[f] \cdot \left(U'^{-1}(q_\ell[f]/V) - S_\ell[f] \right) \right]_{q_{\min}}^{q_{\max}} \tag{2}$$

Algorithm III.1 RANDACC($\langle \ell, r, c \rangle$, λ, μ, T)
- ℓ : link
- r : rate
- c : carrier
- λ : channel access rate
- μ : expected transmission duration
- T : time frame

$t \leftarrow$ beginning of T
while $t \in T$ **do**
   $x \leftarrow$ randomly drawn from $Exp(\lambda)$
   $t \leftarrow t + x$
   if $\langle \ell, r, c \rangle$ leads to config. in $\mathcal{N}_c$ **then**
      $x \leftarrow$ randomly drawn from $Exp(1/\mu)$
      ℓ transmits on c at r during $(t, t+x]$
      $t \leftarrow t + x$

Algorithm III.2 MMUO
for each time frame f **do**
   for each link ℓ, carrier c and rate r **do**
      RANDACC($\langle \ell, r, c \rangle$, $\lambda_{\ell,r,c}[f]$, $\mu_{\ell,r,c}[f]$, f)
   for each ℓ **do**
      update service received $S_\ell[f]$
      update virtual queue $q_\ell[f+1]$ according to (2)
   for each ℓ, c and r **do**
      update $\lambda_{\ell,r,c}[f+1]$ and $\mu_{\ell,r,c}[f+1]$ according to (3)

In the above equation, b is a step size function that satisfies property $(A1)$ (defined later), $q_{\max}$ and $q_{\min}$ are bounds on the virtual queue size, and $[x]_{q_{\min}}^{q_{\max}} = \min(q_{\max}, \max(q_{\min}, x))$. The positive parameter V controls the accuracy of the algorithm.

The values of $\lambda_{\ell,r,c}[f]$ and $\mu_{\ell,r,c}[f]$ stay unchanged during each frame f, and are updated to $\lambda_{\ell,r,c}[f+1]$ and $\mu_{\ell,r,c}[f+1]$ at the end of frame f, so that

$$\lambda_{\ell,r,c}[f+1] \cdot \mu_{\ell,r,c}[f+1] = \exp(r \cdot q_\ell[f+1]). \tag{3}$$

As we shall see in (4) the performance of RANDACC depends on the product of λ and μ. The choice of this product is explained in the proof of Theorem 2.

IV. ANALYSIS

In this section we sketch a proof that the MMUO algorithm leads to an optimal solution to (1). (For reasons of space some of the details are deferred to the full version of the paper [3].) Using a similar argument to [10] we have,

Lemma 1. *The schedule sequence $m^{\vec{\lambda}_c, \vec{\mu}_c}(t)$ for $t \geq 0$ is a continuous time reversible Markov Chain with the following stationary distribution.*

$$\pi_m^{\vec{\lambda}_c, \vec{\mu}_c} = \frac{\Pi_{\ell:r_{\ell,m}>0} \lambda_{\ell,r_{\ell,m},c} \cdot \mu_{\ell,r_{\ell,m},c}}{\sum_{n \in \mathcal{N}_c} \Pi_{\ell:r_{\ell,n}>0} \lambda_{\ell,r_{\ell,n},c} \cdot \mu_{\ell,r_{\ell,n},c}} \quad \forall m \in \mathcal{N}_c \tag{4}$$

Combined with (4) this implies,

$$\pi_m^{\vec{q}} = \frac{\exp\left(\sum_{\ell:r_{\ell,m}>0} r_{\ell,m} \cdot q_\ell\right)}{\sum_{m'\in\mathcal{N}_c} \exp\left(\sum_{\ell:r_{\ell,m'}>0} r_{\ell,m'} \cdot q_\ell\right)} \quad \forall m \in \mathcal{N}_c \quad (5)$$

The resulting link throughput is therefore,

$$\gamma_{\ell,c}^{\vec{q}} = \sum_{m\in\mathcal{N}_c} \pi_m^{\vec{q}} r_{\ell,m} \quad \forall \ell, c \quad (6)$$

For utility optimization, our goal is to show that the virtual queues under MMUO converge to a vector $\vec{q}^*$ so that the above link throughput under $\vec{q}^*$ maximizes the utility as defined in (1). Note that for this problem we cannot treat each carrier in isolation since the throughput of a link is aggregated over all carriers. Note also that the optimization problem (7) in Theorem 2 differs from that of (1) in its objective function, but shares the same constraints. The motivation of this reformulation is to obtain a more useful set of KKT conditions. Further, Theorem 2 will state that the optimal values of the two objective functions can be arbitrarily close. The main result of this section is captured in the following theorem. It relies on two standard but technical assumptions $(A1)$ and $(A2)$ which we describe below.

Theorem 2. *Under assumptions $(A1)$ and $(A2)$, for any initial condition $\vec{q}[0]$, MMUO converges in the following sense.*

$$\lim_{f\to\infty} \vec{q}[f] = \vec{q}^*$$

where $\vec{q}^*$ and $\vec{\gamma}^*$ are such that $(\vec{\gamma}^*, \vec{\pi}^{\vec{q}^*})$ is the solution to the following convex optimization problem over $\vec{\gamma}$ and $\vec{\pi}$.

$$\begin{aligned}
\max \quad & f_2(\vec{\gamma}, \vec{\pi}) = \\
& V \sum_{\ell\in\mathcal{L}} U\left(\sum_c \gamma_{\ell,c}\right) - \sum_c \sum_{m\in\mathcal{N}_c} \pi_m \log \pi_m \quad (7) \\
\text{s.t.} \quad & \gamma_{\ell,c} \leq \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m \quad \forall \ell, c \\
& \sum_{m\in\mathcal{N}_c} \pi_m = 1 \quad \forall c.
\end{aligned}$$

Further, if $\vec{\gamma}^\dagger$ is the optimal solution to (1), then

$$|f_1(\vec{\gamma}^*) - f_1(\vec{\gamma}^\dagger)| \leq |\mathcal{C}| \log |\cup_c \mathcal{N}_c|/V \quad (8)$$

The assumptions of Theorem 2 are:
$(A1)$ $\sum_{f=0}^{\infty} b[f] = \infty$ and $\sum_{f=0}^{\infty} b^2[f] < \infty$.
$(A2)$ If $\vec{p}^o \in \Re_+^L$ is a solution to

$$U'^{-1}(p_\ell/V) - \sum_c \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m^{\vec{p}} = 0 \quad \forall \ell \in \mathcal{L}$$

then $q_{\min} \leq p_\ell^o \leq q_{\max}$ for all $\ell \in \mathcal{L}$.
The parameters $b[\cdot]$ will only be used in the analysis, not in the algorithm itself. In addition, the parameters $q_{\min}$ and $q_{\max}$ are under our control. Hence for any problem instance we can make sure that Assumptions (A1) and (A2) hold.

Proof: (Sketch) The KKT conditions for (7) are:

$$VU'\left(\sum_c \gamma_{\ell,c}\right) = \nu_{\ell,c} \quad \forall \ell, c \quad (9)$$

$$-1 - \log \pi_m + \sum_\ell r_{\ell,m}\nu_{\ell,c} - \eta_c = 0 \quad \forall m \in \mathcal{N}_c, \forall c (10)$$

$$\gamma_{\ell,c} \leq \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m \quad \forall \ell, c \quad (11)$$

$$\nu_{\ell,c} \times \left(\gamma_{\ell,c} - \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m\right) = 0 \quad \forall \ell, c \quad (12)$$

$$\nu_{\ell,c} \geq 0 \quad \forall \ell, c \quad (13)$$

$$\sum_{m\in\mathcal{N}_c} \pi_m - 1 = 0 \quad \forall c \quad (14)$$

These equations can be solved via the standard subgradient method by updating a set of variables p_ℓ according to the following system of differential equations,

$$\dot{p}_\ell = U'^{-1}(p_\ell/V) - \sum_c \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m^{\vec{p}}, \quad (15)$$

and then setting $\nu_{\ell,c} = p_\ell$ for all c and the remaining variables according to Equations (9), (10) and (14). Due to the convexity of the problem (7) the system (15) will eventually converge to a fixed point, $\vec{p}^*$. In the full version of the paper [3] we adapt the argument of [14] to show that the update rules (2) and Assumption $(A2)$ imply:

Lemma 3. *There exists a continuous interpolation of $\vec{q}[f]$ whose limit is a fixed point of,*

$$\dot{p}_\ell = U'^{-1}(p_\ell/V) - \sum_c \sum_{m\in\mathcal{N}_c} r_{\ell,m}\pi_m^{\vec{p}}, \quad (16)$$

Since (15) is identical to (16) the limit point of the continuous interpolation is therefore equal to a vector $\vec{p}^*$ that defines an optimal dual solution of (7) (via the KKT conditions). Hence via (3) the optimal virtual queue sizes determine channel access parameters $\lambda_{\ell,r,c}$ and $\mu_{\ell,r,c}$ for which the corresponding link throughputs provide an optimal primal solution to problem (7).

Finally, since

$$\begin{aligned}
f_2(\vec{\gamma}^*, \vec{\pi}^*) &\geq f_2(\vec{\gamma}^\dagger, \vec{\pi}^\dagger) \\
f_1(\vec{\gamma}^*) &\leq f_1(\vec{\gamma}^\dagger)
\end{aligned}$$

and the entropy $\sum_m \pi_m \log \pi_m \leq \log |\cup_c \mathcal{N}_c|$, the proof of Theorem 2 is complete. ∎

V. A MORE CONCRETE MODEL: HETEROGENEOUS LTE SYSTEM

In this section we present a more concrete model for the scheduling problem so that it more closely matches resource allocation in LTE heterogeneous networks.

We begin with a brief system description. We consider downlink transmissions from a set of basestations to a set of mobile users in a time-slotted system. We assume an OFDM-based air interface in which the spectrum is divided into a set of carriers called resource blocks (RBs), each of which can

be scheduled separately. For example a 20MHz LTE system is typically divided into 100 resource blocks. In the time dimension a time slot corresponds to a Transmission Time Interval (TTI) which has a typical duration of 1ms in an LTE system.

We consider a heterogeneous network in which the bases-tations are divided into two classes, namely macrocells and femtocells. (For ease of description we use the terms "macros" and "femtos". However, our discussions also apply directly to networks with picocells.) Macros typically have a much higher max transmit power than femtos, since macrocells provide wide-area coverage, whereas femtocells (which may be privately owned) provide focused coverage in one specific location, e.g. a house or apartment. At any time instant each mobile user associates with one basestation. Each macro accepts an association with any mobile user. A femto however may be in "Closed Subscriber Group" mode (CSG) and only accept an association with a small subset of users. We remark that femtos have two notable effects that are departures from traditional cellular networks. First, they may create strong interference to a macrocell from within the cell itself, whereas in a macro-only network interference to a cell mostly comes from outside that cell. Second, a mobile user may not be able to associate with the basestation with the strongest signal if the basestation is a femto in CSG mode and cannot associate with the user. We assume that each mobile user associates with the basestation for which the received signal is strongest, among those that the user is able to associate with.

Unlike in the abstract model, we address basestation power allocation and interference directly instead of via the notion of feasibility regions. For basestation i, let U_i be the set of associated users. To abuse notation, we also use U_i to denote the set of links that are incident to i as there is a one-to-one correspondence between the users and links.

The maximum transmit power p_i for basestation i is given and fixed. The scheduling problem is how to distribute p_i among the resource blocks $c \in \mathcal{C}$ and among the users in U_i. Let $p_{i,c}(t)$ be the power allocation of p_i on resource block c at time t; let $p_{i,c,j}(t)$ be the allocation of $p_{i,c}(t)$ on user $j \in U_i$. Note $\sum_c p_{i,c}(t) \leq p_i$, and
$$\begin{cases} p_{i,c,j}(t) &= p_{i,c}(t) \quad \text{for one } j \in U_i \\ p_{i,c,j'}(t) &= 0 \quad\quad\; \text{for } j' \neq j \end{cases}$$
. That is, $p_{i,c}$ is allocated entirely to *one* chosen user $j \in U_i$.

Power settings and transmission rates are related through the *channel quality information* (CQI). CQI values are defined on pairs of links and resource blocks. During every time slot t, the values of $\text{CQI}_{c,\ell}(t)$ for all $c \in \mathcal{C}$ and $\ell \in U_i$ are reported to basestation i. We assume that each basestation has perfect CQI reporting.

Let $r_{c,\ell}(t)$ be the transmission rate along link ℓ on carrier c during time slot t. Specifically, for link $\ell = ij$ between the basestation i and the associated user j, we define

$$r_{c,\ell}(t) = w_c \cdot F(p_{i,c,j}(t) \cdot \text{CQI}_{c,\ell}(t)) \quad (17)$$

$$\text{CQI}_{c,\ell}(t) = \frac{g_{ijc}(t)}{N_c + \sum_{i' \neq i} p_{i'c} g_{i'jc}(t)} \quad (18)$$

In (18), g_{ijc} represents the path loss between i and j on resource block c, and N_c is the background noise on c. Both g_{ijc} and N_c depend on c since radio propagation conditions and background interference may be different on different frequencies. The product of $p_{i,c,j}$ and $\text{CQI}_{c,\ell}$ is commonly referred to as *signal-to-interference-plus-noise ratio*, SINR. (We can therefore think of $\text{CQI}_{c,\ell}$ as the SINR for a unit power transmission.) In (17), w_c is the bandwidth of resource block c and $F(\cdot)$ represents spectral efficiency as a function of SINR. For example $F(\cdot)$ could be a suitably discretized version of the Shannon function $\log(1 + x)$. We assume that $F(\cdot)$ is such that $r_{c,\ell}$ is always a member of a discrete set $\mathcal{R} \cup \{0\}$.

The primary scheduling decision is to determine the power levels $p_{i,c}$. In the literature this problem is sometimes known as inter-cell interference coordination (ICIC). The secondary scheduling decision is to allocate $p_{i,c}$ to the user-level power $p_{i,c,j}$. Typically, each basestation in an LTE network has its own local scheduler for user-level allocation, in which case the scheduling freedom is at the inter-cell level. For concreteness we assume the local scheduler uses the Proportional Fair (PF) algorithm. In the following section we give details on how to allocate $p_{i,c}$ and $p_{i,c,j}$. For user-level allocation $p_{i,c,j}$, we consider two cases depending on whether a local scheduler exists.

VI. IMPLEMENTATION

In Section III we described the utility-optimal MMUO algorithm for the abstract model. In this section we present MMUO-based heuristics for the LTE resource allocation problem in heterogeneous networks. We address a number of issues. First, scheduling decisions need to be made in slotted time rather than in continuous time as in MMUO. Second, scheduling decisions are about setting power levels rather than transmission rates as in MMUO. Third, we discuss how to incorporate a local scheduler such as Proportional Fair. Fourth, perhaps most significantly, we show CQI-based methods for feasibility detection. This replaces the feasibility oracle and the CSMA collision detection mechanism. Lastly, since the basestations are divided into two classes, macro and femto, interference can be reduced by not having every basestation compete on every resource block.

We begin with a basic heuristic that bypasses the last two issues. We then describe three methods through which feasibility can be detected in practice. We conclude with a modified heuristic in which macros and femtos have priority on different sets of resource blocks.

A. Basic Heuristic

Our basic heuristic works very much in the spirit of MMUO. To address the first issue regarding slotted time each frame now consists of an integral number of time slots. When the subroutine RANDACC is called with parameters λ and μ, the time between transmission attempts (resp. the transmission period) is drawn from a geometric distribution with mean $1/\lambda$ (resp. mean μ). One problem is that two links may make

decisions during the same time slot. We can set the $1/\lambda$ values large enough so that this rarely happens. If this does happen we assume that both conflicting transmissions cease. A detailed explanation of how rare collisions affect the performance of utility-optimal CSMA was given in [14] and we can apply a similar analysis to MMUO.

The output of MMUO, as described above, specifies the transmission rate $r_{c,\ell}(t)$ for every transmission $\langle \ell, r, c \rangle$ that takes place during time slot t. To obtain power settings, equation (17) provides the direct translation from transmission rates to power levels.

$$p_{i,c,j}(t) = \frac{F^{-1}(r_{c,\ell}(t)/w_c)}{\mathrm{CQI}_{c,\ell}(t)}, \qquad \text{for link } \ell = ij. \quad (19)$$

For a given basestation i and resource block c, the feasibility oracle guarantees that one user $j \in U_i$ has positive power allocation $p_{i,c,j}(t)$. Let $p_{i,c}(t) = p_{i,c,j}(t)$ for this user j. The transmission now takes place as long as $\sum_c p_{i,c}(t) \le p_i$. This addresses the second issue.

If it is the case that we can specify both user-level as well as inter-cell power allocations, we are done. However, as discussed in Section V, in many instances we only have the freedom for specifying $p_{i,c}$ since the user-level power is determined by a local scheduler such as the Proportional Fair (PF) algorithm. In this case we run the MMUO algorithm "in the background" to compute the $p_{i,c}$ values and then determine which user in U_i receives the transmission power $p_{i,c}$ using the PF algorithm. (We remark that once the power levels are set then which user is chosen by PF does not affect the interference experienced in other cells.)

The PF algorithm works as follows. For each link $\ell \in U_i$ basestation i maintains an estimate R_ℓ of the recent average transmit rate on link ℓ, and allocates power $p_{i,c}$ exclusively to the link ℓ that maximizes the ratio $\tilde{r}_{c,\ell}(t)/R_\ell$, where $\tilde{r}_{c,\ell}(t)$ is the nominal rate if user j has power allocation $p_{i,c,j}(t) = p_{i,c}(t)$. Again from (17), we have $\tilde{r}_{c,\ell}(t) = w_c \cdot F\left(p_{i,c,j}(t) \cdot \mathrm{CQI}_{c,\ell}(t)\right)$ for link $\ell = ij$. After each scheduling decision R_ℓ is updated for each link according to an exponential filter. This addresses the third issue.

B. Methods for Implementing the Feasibility Oracle

We now examine options for the only part of the algorithm that requires coordination among basestations, namely feasibility detection. Since resource block power assignment is typically done on a slower timescale than individual time slots, we are interested in determining whether a set of transmissions will be feasible over multiple timeslots. In particular, we do not want to declare a transmission feasible if this is only true for a single timeslot due to fast fading.

We discuss multiple mechanisms which all use techniques that have been proposed in the standardization process for heterogeneous networks (e.g. [1]). Our initial mechanisms use the existing CQI channel with one extra piece of information which we call the *activity indicator*. We also allow for a basestation to "overhear" a link to which it is not associated. Our later mechanisms show how the algorithm could be

implemented if we indeed have a channel for exchanging information between basestations (such as the X2 channel that is defined in LTE). The bit-rate of such channels is typically limited and so we stress that all we need to exchange are short messages such as the activity values. No detailed exchange of channel state is required.

Method 1: In this method, an activity indicator is reported along with the CQI. Specifically, let $y_{\ell,r,c}(t)$ be the binary activity indicator that is set to one if and only if MMUO makes a transmission $\langle \ell, r, c \rangle$ during time slot t. When $\mathrm{CQI}_{c,\ell}(t)$ is reported to basestation i for which $\ell \in U_i$, $y_{\ell,r,c}(t)$ is also reported if it is set to 1. Each basestation i listens to all CQI that it can decode, not just the CQI for links in U_i. If i hears $y_{\ell',r',c} = 1$ for some ℓ' on resource block c, then every potential transmission $\langle \ell, r, c \rangle$, for $\ell \in U_i$ and $r \in \mathcal{R}$, is declared infeasible. Note that this method is similar in spirit to the Clear-to-Send (CTS) mechanism for 802.11.

Method 2: This method is less stringent than Method 1 in declaring infeasibility. For each activity indicator $y_{\ell,r,c}(t) = 1$ we define the *safety margin* to be the ratio between the currently achievable transmission rate $\tilde{r}_{c,\ell}(t)$ and the actual rate $r_{c,\ell}(t)$ that is used by MMUO. This achievable rate can be computed from the $\mathrm{CQI}_{c,\ell}(t)$ values together with the current power levels. We assume that the safety margins are transmitted on the CQI channel along with the activity indicators. For some threshold $\upsilon > 1$ we say that the activity indicator is *safe* if the safety margin is above υ, *vulnerable* if the margin is between 1 and υ, and *in outage* if the margin is below 1. (Note that if we are in outage then user ℓ could not receive data at rate r for the current CQI values.) Method 2 is the same as method 1 except that basestation i does not refrain from declaring a potential transmission on $\ell \in U_i$ on block c feasible, even if it overhears an activity indicator $y_{\ell',r',c}(t) = 1$ as long as this indicator is currently safe.

The exact value of υ could be a network-wide parameter. Alternatively each basestation could gradually lower a local estimate of υ until it observes links going into outage.

Method 3: This method applies probing to feasibility detection. Whenever a basestation i needs to decide if MMUO could transmit on $\langle \ell, r, c \rangle$, for $\ell \in U_i$, it briefly sets power level $p_{i,c}$ on resource block c and observes the effects on other users. Here $p_{i,c}$ is the power necessary to carry out the transmission $\langle \ell, r, c \rangle$ and can be calculated as in (19). If basestation i overhears that any activity indicator moves into outage then it sets $p_{i,c}$ back to 0 and declares $\langle \ell, r, c \rangle$ infeasible. This method has the drawback that it could send neighboring users into outage for short periods (and this would need to rectified by more robust channel coding on the data channels). However, it has the advantage that basestation i gets a much better sense of the "damage" that might be caused by setting a particular power level $p_{i,c}$ on resource block c.

Methods 4-6: The next three methods are essentially the same as Methods 1-3. However, instead of basestations overhearing activity indicators and their associated safety status, each basestation would directly communicate their own activity indicators and safety margins to all their neighboring

basestations. This can be done using a channel such as the X2 channel in LTE that provides communication between neighboring basestations. Note that this is a lightweight communication since the activity indicator only has 2 possible values and the safety status has only 3 possible values. In particular the basestations would not be exchanging any detailed channel state information.

Relationship to current LTE proposals: We now briefly discuss how the above methods could fit with mechanisms that have been proposed in LTE standards for interference coordination. In the document [1] on RF requirements for femtos, three options are proposed for communication between macros and femtos. The first is direct over-the-air communication. The second is over-the-air via "victim" users. This would correspond to the overhearing methods 1-3 proposed above in that the victim user broadcasts channel quality information that indicates to an interferer whether it is safe to transmit. The third option is via an existing backhaul which corresponds to methods 4-6 above.

Enhanced heuristic: Resource block prioritization:: Note that throughout this section we have assumed that attempts by a link to access resource block c are governed by $\lambda_{\ell,r,c}$ and $\mu_{\ell,r,c}$ and are performed independently across resource blocks. However, since in heterogeneous networks we have two classes of basestations (macros and femtos) there is potential to reduce interference if each class has priority on a different set of resource blocks. We now describe a heuristic to achieve this. In particular if basestation i is a macro we bias it towards low numbered resource blocks by only letting one of its users be active on a resource block if it also has active users on all lower numbered resource blocks. More formally, if $\ell \in U_i$ then $\langle \ell, r, c \rangle$ is feasible if for all $c' < c$ there exists $\ell' \in U_i$ and $r' \in \mathcal{R}$ such that $y_{\ell',r',c'} = 1$. Similarly, if basestation i is a femto we bias it towards high numbered resource blocks by only letting one of its users be active on a resource block if it also has active users on all higher numbered resource blocks. More formally, if $\ell \in U_i$ then $\langle \ell, r, c \rangle$ is feasible if for all $c' > c$ there exists $\ell' \in U_i$ and $r' \in \mathcal{R}$ such that $y_{\ell',r',c'} = 1$.

VII. SIMULATION RESULTS

We now provide an example to show how the algorithms work. We consider a simple toy example since it allows us to compute the optimal schedule. We consider three omnidirectional basestations, 1 macro and 2 femtos, together with six users, two for each basestation. The two macro users are at distance 100m (user 0) and 780m (user 1) respectively. Each femto has two users (users 2-5) at distance 10m. The exact configuration together with the user numbering is shown in Figure 1. We assume that user 1 has to be associated with the macro since the nearby femto is in CSG mode. The transmit power of the macro is 46dBm and for the femtos it is 8dBm. The system bandwidth is 5MHz and the noise density is -165dBm/Hz. We split the system bandwidth into three resource blocks. The pathloss is represented by a COST-231 Hata model. In particular the path loss at distance d meters is
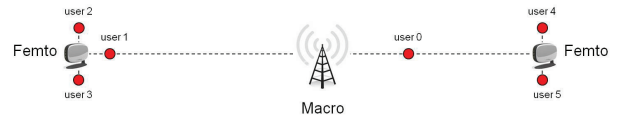

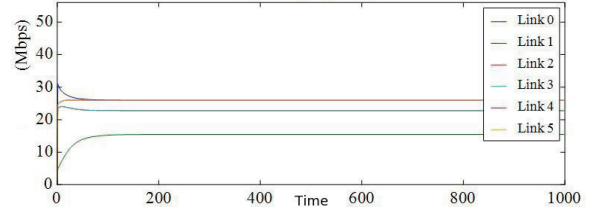Fig. 1.   Network configuration: 1 macro and 2 femtos each with 2 users.

Fig. 2.   Link throughputs under the optimal solution. Top curves for users/links 0, 4 and 5; middle curves for users 2 and 3; bottom curve for user 1. (Note that some curves coincide.)

assumed to be $0.525 * d^{-3.523}$. For simplicity we consider two instantaneous transmission rates, a "low" rate of 8bits/sec/Hz and a "high" rate of 16bits/sec/Hz.

For this configuration we aim to achieve user throughputs that solve problem (7). We can compute the following optimal solution offline via a standard subgradient algorithm. For brevity we use notation of the form $0h2\ell4\ell$ to represent a schedule, which means user 0 is receiving data at the high rate and users 2 and 4 are receiving data at the low rate. In the optimal solution for schedules $m \in \{0h2\ell4\ell, 0h3\ell4\ell, 0h2\ell5\ell, 0h3\ell5\ell\}$ we have $\pi_m = 8.4\%$, for schedules $m \in \{1h4\ell, 1h5\ell\}$ we have $\pi_m = 10.7\%$, and for schedules $m \in \{2h4h, 2h5h, 3h4h, 3h5h\}$ we have $\pi_m = 11.2\%$. Figure 2 shows the corresponding optimal link throughputs.

The behavior of MMUO and MMUO with Proportional Fair are similar and so we present the plot for MMUO with PF only (in Figure 3). (For this plot we assume that the feasibility of a configuration is known perfectly.) The link throughputs closely approximate the optimal rates in Figure 2. User 0 (the closer user to the macro basestation) has the highest throughput while user 1 has the lowest throughput. Among the femto users, users 4 and 5 have higher throughputs since they create less interference to a macro user than users 2 and 3.

We now illustrate the algorithms for testing feasibility. In particular suppose that we are in configuration $1h4\ell$ and suppose that user 3 wants to transmit at the high rate. This is infeasible. User 3 may discover this by either a) overhearing the CQI reported by user 1 and realizing it is sufficiently close to the minimum acceptable CQI or b) briefly probing the channel at the high rate and then discovering that user 1 can no longer support its current rate. In both cases user 3 decides not to transmit.

VIII. PREVIOUS WORK

We now describe how our work relates to existing techniques. Prior work mainly falls into two categories, resource
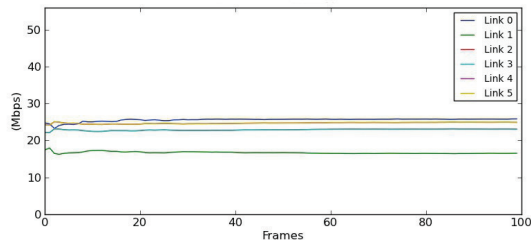
Fig. 3. Link throughputs under MMUO in combination with PF. Top curve for user 0; second curve for users 4 and 5; third curve for users 2 and 3; bottom curve for user 1.

allocation in OFDM systems and CSMA-based algorithms for 802.11 networks. As we have seen, our proposal has been to derive an algorithm based on CSMA techniques for the case of OFDM resource allocation.

OFDM resource allocation: LTE uses an OFDM physical layer. Resource allocation in OFDM systems addresses problems such as channel selection, local scheduling, power control and user association, i.e. which basestation serves which user. One popular technique, e.g. used in [6], [4], [9], is a *Gibbs sampler* approach based on Interacting Particle Systems. The main idea here is that for a given network configuration each node has a local energy based on the interference that it both causes and receives. Nodes then pick new states based on their local energy. Gibbs sampler techniques have also been used to motive greedy algorithms for LTE resource block selection, e.g. [2]. Another popular technique, e.g. used in [20], [19], is to set power levels according to a gradient ascent approach. In particular each transmitter adjusts power levels so as to improve network utility in its neighborhood. Both the Gibbs sampler and the gradient ascent based methods require information exchange on how much interference each transmitter causes to each receiver. For the Gibbs sampler methods interference information needs to be exchanged in order to calculate local energy levels. For the gradient ascent methods nodes need to exchange "partial derivative" information to indicate how the interference they experience would be affected by a change in a neighbor's power levels. We remark that MMUO does not require such detailed information exchange. It bases its calculations on CQI messages that are already included in LTE, augmented with the activity indicators (and possibly safety margins).

CSMA-based Algorithms: In the classic CSMA setup all links wish to access a single channel. Jiang and Walrand [10] showed that CSMA can achieve any set of feasible throughputs. Since this result, a number of papers have looked at how to make channel access rates dependent on local queue sizes in order to keep the system stable, e.g. [7], [8], [5], [17], [18], [16]. As already discussed, we have based our analysis on the work [14] (later extended in [15], [13]) that analyzed utility maximization in a CSMA setting.

IX. CONCLUSION

In this paper we have presented a CSMA-based scheduling algorithm for heterogenous LTE networks with both macro and small cells. Our main contribution is twofold. Mathematically, our algorithm handles the general multiple transmission rates on multiple carriers and achieves utility optimality. For the practical setting, the communication among the basestations utilizes the existing CQI-based technology and hence the additional signaling is minimal.

REFERENCES

[1] 3GPP TR 36.921. Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis.

[2] M. Andrews, V. Capdevielle, A. Feki, and P. Gupta. Self-organizing and self-optimizing networks. *Bell Labs Technical Journal*, 15(3):85 − 97, December 2010.

[3] M. Andrews and L. Zhang. Multi-carrier multi-rate utility optimization in heterogeneous networks via csma-based algorithms. *CoRR*, abs/1211.4258, 2012.

[4] S. Borst, M. Markakis, and I. Saniee. Distributed power allocation and user assignment in OFDMA cellular networks. In *Proceedings of the 49th Annual Allerton Conference on Communication, Control, and Computing*, Monticello, Illinois, 2011.

[5] S.C. Borst, J. Ghaderi, and P.A. Whiting. Backlog-based random access in wireless networks: fluid limits and instability issues (invited paper). In *Proceedings of CISS 2012 and WiOpt 2012*, 2012.

[6] C. S. Chen, F. Baccelli, and L. Roullet. Joint optimization of radio resources in small and macro cell networks. In *Proceedings of IEEE Vehicular Technology Conference*, pages 1 − 5, May 2011.

[7] J. Ghaderi and R. Srikant. On the design of efficient CSMA algorithms for wireless networks. In *Proceedings of CDC 2010*, 2010.

[8] J. Ghaderi and R. Srikant. Effect of access probabilities on the delay performance of Q-CSMA algorithms. In *Proceedings of IEEE INFOCOM 2012*, pages 2068 − 2076, 2012.

[9] I. Hou and P. Gupta. Distributed resource allocation for proportional fairness in multi-band wireless systems. In *Proceedings of the IEEE International Symposium on Information Theory*, 2011.

[10] L. Jiang and J. Walrand. A CSMA distributed algorithm for throughput and utility maximization in wireless networks. In *Proceedings of the Conference on Communication, Control and Computing*, Monticello, IL, 2008.

[11] P. Karn. MACA - A new channel access method for packet radio. In *Proceedings of the 9th ARRL Computer Networking Conference*, London, Ontario, Canada, 1990.

[12] F. Kelly. *Reversibility and Stochastic Networks*. Wiley: Chichester, 1979.

[13] J. Lee, Y. Yi, S. Chong, B. Nardelli, E. Knightly, and M. Chiang. Making 802.11 dcf optimal: Design, implementation, and evaluation. *CoRR*, abs/1207.3740, 2012.

[14] J. Liu, Y. Yi, A. Proutiere, M. Chiang, and H. V. Poor. Towards utility-optimal random access without message passing. *Wireless Communications and Mobile Computing*, 10:115 − 128, 2010.

[15] A. Proutiere, Y. Yi, T. Lan, and M. Chiang. Resource allocation over network dynamics without timescale separation. In *INFOCOM*, pages 406–410, 2010.

[16] D. Qian, D. Zheng, J. Zhang, and N. Shroff. CSMA-based distributed scheduling in multi-hop MIMO networks under SINR model. In *INFOCOM*, 2010.

[17] S. Rajagopalan, D. Shah, and J. Shin. Network adiabatic theorem: an efficient randomized protocol for contention resolution. In *Proceedings of SIGMETRICS/Performance 2009*, pages 133 − 144, 2009.

[18] D. Shah and J. Shin. Delay optimal queue-based CSMA. In *Proceedings of SIGMETRICS 2010*, pages 373 − 374, 2010.

[19] K. Son, S. Lee, Y. Yi, and S. Chong. Refim: A practical interference management in heterogeneous wireless access networks. *CoRR*, abs/1105.0738, 2011.

[20] A. L. Stolyar and H. Viswanathan. Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination. In *Proceedings of IEEE INFOCOM 2009*, 2009.