Utility optimization in heterogeneous networks via CSMA-based algorithms

Matthew Andrews · Lisa Zhang

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 mechanism is used to decide transmission feasibility.

Keywords Random access algorithm · Utility optimization

1 Introduction

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.

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 [10]. 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 [9] 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. [13] 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 [12] and presented an extension to the case of a single transmission rate on multiple carriers [14]. The best of our knowledge no previous work has looked at how CSMA-based techniques can be applied in the LTE context. In particular, work on CSMA has typically assumed the presence of an interference graph that captures pairwise interference relationships between links (i.e. a link can transmit if and only if none of its neighbors are transmitting). A key question we wish to address is how to use algorithms similar to [9] and [13] in a context the interference relationships are too complex to be captured by a simple graph.

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 [9, 13, 14], 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 [13]. We believe our main contribution is in showing that the proof of convergence in [13] 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.

We begin in Sect. 2 with an abstract model in which details of the interference are abstracted away into a feasibility region for the transmissions. Many of the existing CSMA algorithms work with an interference graph in which all interference relationships are pairwise between links. We instead work with a more general setting in which an “oracle” informs a transmitter whether a potential transmission would disrupt existing transmissions. This allows us to capture interference relationships that are more general than can be captured by an interference graph.

In Sects. 3 and 4 we adapt the Liu et al. single-carrier single-rate utility maximization algorithm for CSMA to the abstract framework of Sect. 2 to address the general case of multiple transmission rates on multiple carriers. The main issue we must overcome with multiple carriers is that the utility for a user depends on the sum-rate of its service across all carriers. This in turn leads to a number of changes to the Karush–Kuhn–Tucker (KKT) conditions that form the underlying basis of the algorithm. In Sects. 5 and 6 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.

2 Abstract model

We begin with an abstract model 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 \( r_\ell \) 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 $\gamma_{t,c}$ indicate the transmission rate on link $\ell$ on carrier $c$. For a given concave utility function $U()$ we wish to maximize the aggregate utility over all links, i.e. to maximize $\sum_{c} U(\sum_{\ell} \gamma_{t,c})$. 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 $r_{t,m}$ indicates the transmission rate on $\ell$ under schedule $m$, then $\gamma_{t,c}$ can be viewed as a weighted sum of $r_{t,m}$ where $\pi_m$ serves as the weight. Throughout the paper, notations such as $\langle \ell, r \rangle$ indicate that we cannot simply treat each carrier as an isolated system.

We can view this as an extension of the formulation of [14] 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 is 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.

### 3 Multi-carrier, multi-rate scheduling

We first describe a routine RANDACC (in Algorithm 3.1), 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 $r \in \mathbb{R}$. 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 $\ell$ conflicts with itself, namely $\ell$ already transmits on $c$, or $\ell$ 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.

---

**Algorithm 3.1 RANDACC($\langle \ell, r, c \rangle$, $\lambda$, $\mu$, $T$)**

$t \leftarrow$ beginning of time frame $T$

while $t \in T$ do

$x \leftarrow$ randomly drawn from $\text{Exp}(\lambda)$

$t \leftarrow t + x$

If $\langle \ell, r, c \rangle$ leads to a configuration in $\mathcal{N}_c$ then

$x \leftarrow$ randomly drawn from $\text{Exp}(1/\mu)$

link $\ell$ transmits on carrier $c$ at rate $r$ during $(t, t + x)$

$t \leftarrow t + x$

---

$$
\max_{\gamma_{t,c}} f_1(\gamma) \leq \sum_{\ell \in \mathcal{L}} U \left( \sum_{c} \gamma_{t,c} \right)
$$

s.t. $\gamma_{t,c} \leq \sum_{m \in \mathcal{N}_c} r_{t,m} \pi_m \forall \ell, c$ and $\sum_{m \in \mathcal{N}_c} \pi_m = 1 \forall c$.

The algorithm MMUO (multi-carrier multi-rate utility optimization) approximates a solution to (1) as follows (see Algorithm 3.2). 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 $\mu_{\ell,r,c}[f]$. 

© Springer
At the end of each frame, link $\ell$ calculates the service received during the frame (denoted by $S[\ell]$), and updates a virtual queue size parameter (denoted by $q[\ell]$) as follows.

$$q[\ell] + 1 = \left[q[\ell] + b[\ell] \cdot \left(U^{\ell-1}(q[\ell]/V) - S[\ell]\right)\right]^{q_{\text{max}}}_{q_{\text{min}}}$$

(2)

In the above equation, $b$ is a step size function that satisfies property (A1) (defined later), $q_{\text{max}}$ and $q_{\text{min}}$ are bounds on the virtual queue size, and $[x]_{q_{\text{max}}} = \min(q_{\text{max}}, \max(q_{\text{min}}, x))$. 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

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

(3)

As we shall see in (4) the performance of RANDAcc depends on the product of $\lambda$ and $\mu$. The choice of this product is explained in the proof of Theorem 4.

Algorithm 3.2 MMUO

for each time frame $f$ do
  for each link $\ell$, carrier $c$ and rate $r$ do
    RANDAcc($\ell$, $r$, $c$), $\lambda_{\ell,r,c}[f]$, $\mu_{\ell,r,c}[f]$, $f$
  for each $\ell$ do
    update service received $S[\ell]$ for $\ell$
    update virtual queue $q[\ell] + 1$ according to (2)
  for each $\ell$, $c$ and $r$ do
    update $\lambda_{\ell,r,c}[f + 1]$ and $\mu_{\ell,r,c}[f + 1]$ according to (3)

4 Analysis

We now show that the MMUO algorithm leads to an optimal solution to (1). The proof builds upon the following optimality properties of RANDAcc that were shown in [9] for the special case of one carrier and 0/1 transmission rates.

4.1 RANDAcc for 0/1 rates

Suppose each link $\ell \in \mathcal{L}$ calls RANDAcc with parameters $(\lambda_{\ell}, \mu_{\ell})$ over a sufficiently long time frame. Let $\lambda = (\lambda_{\ell})_{\ell \in \mathcal{L}}$ and let $\mu = (\mu_{\ell})_{\ell \in \mathcal{L}}$. Let $m^\lambda\mu(t)$ be the schedule at time $t$ under the RANDAcc routine. Recall $r_{\ell,m}$ is the transmission rate on link $\ell$ under schedule $m$. In this case $r_{\ell,m} \in \{0, 1\}$.

Lemma 1 [11] The sequence of schedules $m^\lambda\mu(t)$ for $t \geq 0$ form a continuous time reversible Markov Chain with the following stationary distribution.

$$\pi_m^\lambda\mu = \frac{\Pi_{\ell,r,m=1}^{\lambda_{\ell}, \mu_{\ell}}}{\sum_m \Pi_{\ell,r,m=1}^{\lambda_{\ell}, \mu_{\ell}}} \quad \forall m,$$

where $\Pi_{\ell \in \mathcal{L}}(\cdot) = 1$.

Further, the resulting link throughput $\gamma_{\ell}^\lambda\mu = \sum_m \pi_m^\lambda\mu r_{\ell,m}$ is optimal for every link $\ell \in \mathcal{L}$ in the following sense. We say that a link throughput vector $\gamma^\lambda\mu$ is feasible if there exist $\pi_m^\lambda\mu$ such that $\gamma_{\ell}^\lambda\mu \leq \gamma_{\ell}^\lambda\mu$.

Lemma 2 [9] and Lemma 1 in [13] For any feasible link throughput vector $\gamma_{\ell}^\lambda\mu$, there exists $\lambda$ and $\mu$ such that $\gamma_{\ell}^\lambda\mu \leq \gamma_{\ell}^\lambda\mu$.

4.2 MMUO

For multiple transmission rates on multiple carriers, the generalization from Lemmas 1–3 is straight-forward. This is because RANDAcc that runs on one carrier does not interfere with that on a different carrier. Further, each rate-link pair can be treated as a distinct link. For a particular carrier $c$, let vectors $\gamma_{c}^\lambda\mu = (\lambda_{\ell,r}, \mu_{\ell,r,c})_{\ell \in \mathcal{L}, r \in \mathcal{R}}$ and $\mu_{c} = (\mu_{\ell,r,c})_{\ell \in \mathcal{L}, r \in \mathcal{R}}$. Recall $\mathcal{N}_c$ is the set of feasible schedules on carrier $c$.

Lemma 3 For multiple transmission rates on multiple carriers, the generalization from Lemmas 1–3 is straight-forward. This is because RANDAcc that runs on one carrier does not interfere with that on a different carrier. Further, each rate-link pair can be treated as a distinct link. For a particular carrier $c$, let vectors $\gamma_{c}^\lambda\mu = (\lambda_{\ell,r}, \mu_{\ell,r,c})_{\ell \in \mathcal{L}, r \in \mathcal{R}}$ and $\mu_{c} = (\mu_{\ell,r,c})_{\ell \in \mathcal{L}, r \in \mathcal{R}}$. Recall $\mathcal{N}_c$ is the set of feasible schedules on carrier $c$.
where the virtual queue vector \( \tilde{q} \) 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 formulation (3) can be written in the following convex optimization problem over \( N_c \). Further, if \( \tilde{q}^* \) is the optimal solution to (1), then
\[
|f_1(\tilde{q}^*) - f_1(\tilde{q}^*)| \leq |C| \log |\cup_c N_e|/V
\]

The assumptions of Theorem 4 are:

(A1) \( \sum_{t=0}^{\infty} b[f] = \infty \) and \( \sum_{t=0}^{\infty} b^2[f] < \infty \).

(A2) If \( p^0 \in \mathbb{R}_+^L \) is a solution to
\[
U^{r-1}(p_r/V) - \sum_{c} \sum_{m \in N_e} r_{c,m} \hat{p}^0_m = 0 \quad \forall \ell \in \mathcal{L}
\]
then \( q_{\text{min}} \leq \hat{p}^0 \leq q_{\text{max}} \) for all \( \ell \in \mathcal{L} \).

The parameters \( b[.] \) will only be used in the analysis, not in the algorithm itself. In addition, the parameters \( q_{\text{min}} \) and \( q_{\text{max}} \) are under our control and hence can be chosen to encompass any potential solution to Eq. 9. Hence for any problem instance we can make sure that Assumptions (A1) and (A2) hold.

**Proof** There are several steps in the proof. In the first two steps we follow the framework of [13] to show that the dynamics of MMUO can be captured by a system of differential equations. In the third step we must deviate from the approach of [13] in order to handle multiple transmission rates on multiple carriers.

We begin by replacing the discrete time frames of MMUO with a more convenient continuous interpolation. For notation we use square brackets \( [\cdot] \) indexed with integers \( f \) for discrete sequences defined on frames \( f \), and we use round brackets \( (\cdot) \) indexed with real numbers \( t \) for a continuous *scaled* version of time. For the discrete time sequence of virtual queue vectors \( \tilde{q}[f] \) for integral frames \( f = 1, 2, \ldots \), we define as follows its continuous interpolation \( \tilde{q}(t) \) for all real positive numbers \( t \). We also define a continuous version \( S_t(t) \) for the discrete service sequence \( S_t[f] \). For \( f = 1, 2, \ldots \), let \( t_f = \sum_{i=1}^f b[i] \). For \( t \in [t_f, t_{f+1}) \), let
\[
q_{ij}(t) = q_i(t) \cdot \frac{t_f + 1 - t}{t_{f+1} - t_f} + q_i[0] \cdot \frac{t - t_f}{t_{f+1} - t_f}
\]
\[
S_t(t) = S_t[f]
\]
(In other words the continuous time process \( t \) is created from the discrete time frames scaled by the intervals \( [t_f, t_{f+1}) \).)

The following lemma is shown in [13] and says that the continuous sequence \( \tilde{q}(t) \) converges to the solution of the system of stochastic differential equations (10) which can be viewed as a continuous version of (2). The equations are
stochastic due to the term $S_t(t)$ which is the result of the stochastic process of MMUO.

**Lemma 5** (Lemma 2 in [13]) Let $\hat{\bar{p}}^*$ be the solution to the following system of differential equations with variable $\bar{p} = (\bar{p}_t)_{t \in \mathbb{C}}$.

$$\dot{\bar{p}}_t = \left[U^{-1}(p_t/V) - S_t(t)\right] \cdot 1_{p_t \in [q_{\min}, q_{\max}]}$$

(10)

where $1_{p_t \in [q_{\min}, q_{\max}]}$ is an indicator variable for whether $p_t$ is in the range of $[q_{\min}, q_{\max}]$. Fix any time instant $\tau$. If $\bar{p}^*(\tau) = q(\tau)$, then $\lim_{\tau \to \infty} \sup_{t \in [\tau, T]} \|\bar{p}^*(t) - q(t)\| = 0$.

The second step of the proof says that any fixed point of the stochastic system (10) will also be a fixed point of an associated deterministic system of equations. Note that the RANDAcc routine may not converge to the stationary distribution of (4), or equivalently (5), with the given $\lambda$ and $\mu$ parameters within each frame before the parameters are updated to their respective new values for the next frame. Hence the service $S_t(t)$ in the above differential equations is a stochastic quantity. In other words, when MMUO invokes the RANDAcc routine, if each frame is sufficiently long, $S_t(t)$ would converge to its long-term average $\sum_c \cdot \sum_{m \in \mathcal{N}} \cdot \pi_m^0$ and we could replace the stochastic term $S_t(t)$ in (10) by its long term average. This would result in the following system (11). Unfortunately, in reality, the frame $f$ may not be long enough. However, the following result still allows us to say that systems (10) and (11) are closely related, regardless of the convergence of RANDAcc. As we shall see, (11) also provides a connection to solving (7) via the KKT conditions.

**Lemma 6** (Lemma 2 in [13]) Every limit point of system (10) is always a fixed point of the following system.

$$\dot{\bar{p}}_t = \left[U^{-1}(p_t/V) - \sum_c \sum_{m \in \mathcal{N}} r_{t,m} \cdot \pi_m^0\right] \cdot 1_{p_t \in [q_{\min}, q_{\max}]}$$

(11)

for all $\ell \in \mathcal{L}$ where $\pi_m^0$ is defined in (5).

In the third step of the proof we show that the system (11) leads to a solution of the optimization problem (7). It effectively solves the problem via the subgradient method. For this step we must deviate from the analysis of [13] in order to handle the multi-carrier and multi-rate aspects of (7). In particular, the Lagrangian of (7) is given by:

$$L(\bar{y}, \bar{\pi} : \bar{v}, \bar{\eta})$$

$$= \sum_{\ell \in \mathcal{L}} \left(V \cdot U \left(\sum_c \gamma_{t,c} \right) - \sum_c v_{t,c} \gamma_{t,c}\right)$$

$$+ \sum_c \left(\sum_{\ell \in \mathcal{L}} v_{t,c} \cdot \sum_{m \in \mathcal{N}} r_{t,m} \pi_m - \sum_{m \in \mathcal{N}} \pi_m \log \pi_m\right)$$

$$- \sum_c \eta_c \left(\sum_{m \in \mathcal{N}} \pi_m - 1\right).$$

Hence, the KKT conditions for (7) are:

$$V_U\left(\sum_c \gamma_{t,c}\right) = v_{t,c} \quad \forall \ell, c$$

(12)

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

(13)

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

(14)

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

(15)

$$\pi_m \geq 0 \quad \forall \ell, c$$

(16)

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

(17)

Equations (12) and (13) state the gradient of the Lagrangian $L(\bar{y}, \bar{\pi} : \bar{v}, \bar{\eta})$ is zero with respect to the variables $\bar{y}$ and $\bar{\pi}$. Note that Equality (12) implies that for each $\ell$ the Lagrange multipliers $v_{t,c}$ are all equal at optimality. This reflects the coupling between the carriers, i.e. the fact that the utility is a function of the sum-rate across all carriers. Inequalities (14), (15) and (16) state that the first set of constraints of (7) hold and has zero duality gap. The last equality (17) states the second set of constraints of (7) hold. Note that the incorporation of multiple rates is reflected in the presence of $r_{t,m}$ in Inequalities (12)–(15).

We introduce a new variable $\bar{p} = (p_t)_{t \in \mathbb{C}}$ and set the primary variables $\pi_m$ and dual variables $v_{t,c}, \eta_c$ as follows.

$$\pi_m = \pi_m^0 \quad \text{as in } (5) \quad \forall m$$

$$v_{t,c} = p_t \quad \forall \ell, c$$

$$\eta_c = \log \left(\sum_{m \in \mathcal{N}_c} \exp\left(\sum_{\ell} r_{t,m} q_{t}\right)\right) - 1 \quad \forall c.$$

We can see that the KKT conditions (13), (16) and (17) are easily satisfied due to the definition of $\pi_m^0$ in (5) and as long as $\bar{p} \in \mathbb{R}_+$.

We next aim to satisfy the remaining KKT conditions (12), (14) and (15) via the subgradient method. From (12), (15) and the fact that $v_{t,c}$ is set to $p_t$, we wish to have,

$$p_t \times \left(U^{-1}(p_t/V) - \sum_{m \in \mathcal{N}_c} \sum_{t,m} p_t \pi_m^0\right) = 0 \quad \forall \ell$$

If this does not hold (and hence we are not yet at optimality), the standard subgradient method updates the $p_t$ according to the following system of differential equations.

$$\dot{p}_t = U^{-1}(p_t/V) - \sum_{m \in \mathcal{N}_c} \sum_{t,m} p_t \pi_m^0.$$  

(18)
Due to the convexity of the problem \( (7) \) the system \( (18) \) will eventually converge to a fixed point, \( \overline{\pi}^* \). Note that the system \( (18) \) is identical to \( (11) \) as long as \( p_t \in [q_{\min}, q_{\max}] \) for all \( \ell \in \mathcal{L} \). Due to assumption \( (A2) \) and the definition of \( \gamma_{\ell,c} \), the solution to the dual of \( (7) \) without the constraint \( p_t \in [q_{\min}, q_{\max}] \) falls in the range of \( [q_{\min}, q_{\max}] \) and is therefore equivalent to the fixed point \( \overline{\pi}^* \) of the system \( (18) \). With \( \overline{\pi}^* \) chosen, KKT condition \( (12) \) is satisfied. We now set
\[
\gamma_{\ell,c} = \sum_{m \in \mathcal{N}_c} \pi_{m}^\overline{\pi} \rho_{t,m} \quad \forall \ell,c,
\]
which satisfy the remaining conditions of \( (14) \) and \( (15) \).

Finally, since \( f_2(\overline{\pi}^*, \overline{\pi}^*) \geq f_2(\overline{\pi}, \overline{\pi}) \), \( f_1(\overline{\pi}) \leq f_1(\overline{\pi}) \) and the entropy \( \sum_m \pi_m \log \pi_m \leq \log |\cup_c \mathcal{N}_c| \), the proof of Theorem 4 is complete.

We conclude by briefly summarizing in what sense we have shown that MMUO is optimal. We have shown that an appropriate continuous interpolation of the virtual queue dynamics approaches in the limit a vector \( \overline{\pi}^* \) that defines an optimal dual solution of \( (7) \) (via the KKT conditions). Via (3) the optimal virtual queue sizes determine channel access parameters \( \lambda_{\ell,c} \) and \( \mu_{\ell,c} \) for which the corresponding link throughputs provide an optimal primal solution to problem \( (7) \).

### 5 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. 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 20 MHz 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 1 ms in an LTE system.

We consider a heterogeneous network in which the basestations are divided into two classes, namely macro-cells and femtocells. (Note that 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 accesses 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 \( p_t \) for basestation \( i \) is given and fixed. The scheduling problem is how to distribute \( p_t \) 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 \( p_t \) 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 \( \sum_c p_{i,c}(t) \leq p_t \), and
\[
\begin{cases}
  p_{i,c,j}(t) = p_{i,c}(t) & \text{for } j \in U_i, \\
  p_{i,c,j}(t) = 0 & \text{for } j' \neq j
\end{cases}
\]
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 CQI \( c, \ell \) for all \( c \in \mathcal{C} \) and \( \ell \in \mathcal{L} \) 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 \( \ell \) on carrier \( c \) during time slot \( t \). Specifically, for link \( \ell = ij \) between the basestation \( i \) and the associated user \( j \), we define
\[
r_{c,\ell}(t) = w_c \cdot F(p_{i,c,j}(t)) \cdot \text{CQI}_{c,\ell}(t) \tag{19}
\]
\[
\text{CQI}_{c,\ell}(t) = \frac{g_{ijc}(t)}{N_c + \sum_{c' \neq c} g_{i,c',j}(t)} \tag{20}
\]
In (20), \( g_{ijc} \) 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 \( \text{CQI}_{c,\ell}(t) \) is commonly referred to as signal-to-interference-plus-noise ratio, SINR. (We can therefore think of \( \text{CQI}_{c,\ell}(t) \) as the SINR for a unit power transmission.) In (19), \( w_c \) 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 $\log(1 + \lambda)$. We assume that $F(\cdot)$ is such that $r_{\cdot,\cdot}$ is always a member of a discrete set $\mathcal{R} \cup \{0\}$.

The primary scheduling decision is to determine the power levels $p_{t,\cdot}$. In the literature this problem is sometimes known as inter-cell interference coordination (ICIC). The secondary scheduling decision is to allocate $p_{t,\cdot}$ to the user-level power $p_{t,\cdot}$. 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 $p_{t,\cdot}$ and $p_{t,\cdot}$. For user-level allocation $p_{t,\cdot}$, we consider two cases depending on whether a local scheduler exists.

6 Implementation

In Sect. 3 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.

6.1 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 $\lambda$ and $\mu$, the time between transmission attempts (resp. the transmission period) is drawn from a geometric distribution with mean $1/\lambda$ (resp. mean $\mu$). 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 [13] and we can apply a similar analysis to MMUO.

The output of MMUO, as described above, specifies the transmission rate $r_{\ell,\cdot}(t)$ for every transmission $(\ell, r, c)$ that takes place during time slot $t$. To obtain power settings, Eq. (19) provides the direct translation from transmission rates to power levels. In particular we set $p_{t,\cdot}(t)$ to be the minimum possible power level such that:

$$r_{\ell,\cdot}(t) \leq w_{c} \cdot F(p_{t,\cdot}(t)) \cdot \text{CQI}_{c,\cdot}(t), \quad \text{for link } \ell = ij.$$  (21)

The reason that the above might not be an equality is that $r_{\ell,\cdot}(t)$ could be restricted to a finite set of possible powers.

For a given basestation $i$ and resource block $c$, the feasibility oracle guarantees that at most one user $j \in U_{i}$ has positive power allocation $p_{t,\cdot}(t)$. Let $p_{t,\cdot}(t) = p_{t,\cdot}(t)$ for this user $j$. The transmission now takes place as long as $\sum_{c} p_{t,\cdot}(t) \leq p_{t}$. 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 Sect. 5, in many instances we only have the freedom for specifying $p_{t,\cdot}$ 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 $p_{t,\cdot}$ values and then determine which user in $U_{i}$ receives the transmission power $p_{t,\cdot}$ 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 $R_{i}$ of the recent average transmit rate on link $\ell$, and allocates power $p_{t,\cdot}$ exclusively to the link $\ell$ that maximizes the ratio $\tilde{r}_{\ell,\cdot}(t)/R_{i}$, where $\tilde{r}_{\ell,\cdot}(t)$ is the nominal rate if user $j$ has power allocation $p_{t,\cdot}(t) = p_{t,\cdot}(t)$. Again from (19), we have $\tilde{r}_{\ell,\cdot}(t) = w_{c} \cdot F(p_{t,\cdot}(t)) \cdot \text{CQI}_{c,\cdot}(t)$. For link $\ell = ij$. After each scheduling decision $R_{i}$ is updated for each link according to an exponential filter. This addresses the third issue.

6.2 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 the need to exchange 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 only if MMUO makes a transmission \( \langle \ell, r, c \rangle \) during time slot \( t \). When CQL\(_c\rangle(t) \) 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 CQL\(_c\rangle(t) \) that it can decode, not just the CQI for links in \( U_i \). If \( i \) hears \( y_{\ell,r,c}(t) = 1 \) for some \( \ell' \) on resource block \( c \), then every potential transmission \( \langle \ell', r, c \rangle \), for \( \ell \in U_i \) and \( r \in R_c \), 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 \( s_{\ell,r,c}(t) \) to be the ratio between the currently achievable transmission rate \( \bar{r}_{\ell,c}(t) \) and the actual rate \( r_{\ell,c}(t) \) that is used by MMUO. This achievable rate can be computed from the CQL\(_c\rangle(t) \) values together with the current power levels. (There are two reasons why the achievable transmission rate might be different from the actual rate. First, both the transmission powers and transmission rates may be restricted to finite sets and so the transmission power used may be higher than is strictly necessary for the current transmission rate. Second, some interfering basestations might have reduced their power since the transmission started and hence there is less interference.)

We assume that the safety margins are transmitted on the CQI channel along with the activity indicators. For some threshold \( v > 1 \) we say that the activity indicator is safe if the safety margin is above \( v \), vulnerable if the margin is between 1 and \( v \), and in outage if the margin is below 1. (Note that if we are in outage then user \( \ell \) 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 \( v \) could be a network-wide parameter. Alternatively each basestation could gradually lower a local estimate of \( v \) until it observes links going into outage. Note that Method 2 may not calculate the feasible sets exactly since the safety margin does not depend on the interference that basestation \( i \) may cause. The next method addresses this shortcoming.

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 \( p_{\ell,c} \) on resource block \( c \) and observes the effects on other users. Here \( p_{\ell,c} \) is the power necessary to carry out the transmission \( \langle \ell, r, c \rangle \) and can be calculated as in (21). If basestation \( i \) overhears that any activity indicator moves into outage then it sets \( p_{\ell,c} \) 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 \( p_{\ell,c} \) 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 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 $(\ell, r, c)$ 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 $(\ell, r, c)$ is feasible if for all $c' > c$ there exists $\ell' \in U_i$ and $r' \in \mathcal{R}$ such that $y_{\ell', r', c'} = 1$.

7 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 100 m (user 0) and 780 m (user 1) respectively. Each femto has two users (users 2–5) at distance 10 m. The exact configuration together with the user numbering is shown in Fig. 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 46 dBm and for the femtos it is 8 dBm. The system bandwidth is 5 MHz and the noise density is $-165$ dBm/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 assumed to be $0.525 + d^{-3.523}$. For simplicity we consider two instantaneous transmission rates, a “low” rate of 8 bits/s/Hz and a “high” rate of 16 bits/s/Hz.

We aim to achieve user throughputs that solve problem (7). We wish to compare the performance of MMUO with the optimal solution and so we purposely choose an example for which the optimal solution is computable (but non-obvious). Note that there are 6 links and 2 transmission rates but only 3 links can be transmitting at any given time. Hence there are at most $\binom{12}{6} + \cdots + \binom{12}{2} = 299$ feasible schedules on each carrier, i.e. $|\mathcal{S}_c| \leq 299$. This number is small enough that we can use Eq. (4) to compute the stationary distribution of the Markov Chain for any values of $\tilde{\lambda}_c$ and $\tilde{\mu}_c$. This in turn allows us to use a standard subgradient algorithm on the $\tilde{\lambda}_c$ and $\tilde{\mu}_c$ values to compute the overall optimal solution. Note that this direct optimization would be more challenging for a larger example since the number of feasible states grows exponentially with the number of users. (This would not affect the implementability of MMUO since it does not store the full set of feasible schedules. It would however mean that we could not directly compare the performance of MMUO against the optimal solution.)

For the example shown in Fig. 1 the optimal solution is as follows. For brevity we use notation of the form $0h2f4\ell$ 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 $m \in \{0h2f4\ell, 0h3f4\ell, 0h2f5\ell, 0h3f5\ell\}$ we have $\pi_m = 8.4\%$, for schedules $m \in \{1h4f\ell, 1h5f\ell\}$ we have $\pi_m = 10.7\%$, and for schedules $m \in \{2h4f, 2h5f, 3h4f, 3h5f\}$ we have $\pi_m = 11.2\%$. Figure 2 shows the corresponding optimal link throughputs. The utility for a link is defined by $U(X) = \log_2(X)$ where $X$ is throughput measured in Mbps. The optimal utility for the system is 19.252.

The behavior of MMUO and MMUO with Proportional Fair are similar and so we present the plot for MMUO with PF only (in Fig. 3). (In this figure we assume that the feasibility of a configuration is known perfectly and in order to smooth out short term fluctuations in the curves we plot cumulative average service up to a given time $t$.) Throughputs closely approximate the optimal rates in Fig. 2. (Even though the relative ordering of user 0 and users 4 and 5 has switched, all three of them have similar throughput.) Among macro users, user 0 (the closer user to the macro basestation) has the higher throughput while user 1 has the lower 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. The utility achieved by the system is 18.756. Almost all the gap between this value and the optimal value of 19.252 can be
attributed to our choice of $V$, which we take to be 1.0. We ascertained this by also running the idealized subgradient algorithm with $V = 1.0$ and obtaining a utility of 18.759 which is only a little higher than that obtained by MMUO. The reason we do not use a higher value of $V$ within MMUO is simply to ensure that the system converges on a reasonable timescale.

We now illustrate the algorithms for testing feasibility. We focus on Methods 1–3 since Methods 4–6 are essentially the same from a performance perspective. They simply use a different mechanism for communicating among the basestations. We assume that the macro basestation can overhear the CQI channel for all users while each femto can only overhear the CQI channel for its own users and the macro user that is closest to it. For example, the left hand femto in Fig. 1 can overhear the CQI channel for users 1, 2, and 3. For Method 1 a basestation will only start transmitting if it does not overhear the CQI channel for a transmission that is already in progress. One implication of this is that the order in which transmissions start affects whether or not a schedule is declared feasible. For example, if $4h$ is in progress then the macro will not start a transmission $1h$. However, if $1h$ is in progress then the right hand femto would start a transmission $4h$ since it cannot overhear the CQI channel for user 1.

The link throughputs for Method 1 are shown in Fig. 4 and the aggregate utility is 16.7. The main problem with this approach is that the macro users suffer starvation since they the macro hears the CQI channel for all of the other users. This is similar to the well-known problem with collision-based systems such as 802.11 in which if three links are side-by-side and the outer links do not hear each other but the middle link hears both of the others, the middle link will suffer starvation.

For Method 2 we need to know the safety margin for each feasible configuration. In the following table we provide a set of examples. For example, the row with
0h2ℓ4ℓ in the left column means that whenever that schedule is active, user 0 has a safety margin of 1.38, and users 2 and 4 have safety margins of 1.21.

| Schedule | Safety Margin |
|----------|---------------|
| 0h       | 4.45          |
| 0h2ℓ     | 1.49          |
| 0h4ℓ     | 1.41          |
| 0h2ℓ4ℓ   | 1.38          |
| 1h       | 3.80          |
| 1h4ℓ     | 1.01          |
| 2h       | 4.39          |
| 2h4h     | 1.61          |
| 2ℓ       | 8.78          |
| 2ℓ4ℓ     | 3.22          |

Suppose we set the safety threshold \( v = 2.0 \). Then, (unlike Method 1), Method 2 would allow user 0 to start transmitting when the current schedule is 2ℓ4ℓ since the safety margins for both existing transmissions are greater than \( v \). However, Method 2 may sometimes be too aggressive. Suppose the current schedule is 1h in which case the safety margin for user 1 is 3.80. In this case the left-hand femto may start to transmit even though this will then push user 1 into outage (at which point MMUO will act as if the transmission to user 1 had terminated). The link throughputs for Method 2 are shown in Fig. 5 and the aggregate utility is 17.715. In comparison to Method 1, the throughputs for the macro users has improved at the expense of the femto users.

In contrast to Methods 1 and 2, Method 3 is able to determine exactly whether or not a schedule is feasible. For example, suppose that we are in configuration 1h4ℓ and user 3 wants to transmit at the high rate. This is infeasible. User 3 discovers this by briefly probing the channel at the high rate and then discovering (by overhearing user 1’s CQI channel) that user 1 can no longer support its current rate. In this case user 3 decides not to transmit. Although user 1 would briefly be in outage during the probing, this fraction of data lost can be made arbitrarily small by increasing the ratio of the transmission period to the probing period. Hence the link throughputs will be equivalent to those shown in Fig. 3.

8 Previous work

We now describe how our work relates to existing techniques. Prior work mainly falls into two categories, resource 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 [3, 5, 8], 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 motivate greedy algorithms for LTE resource block selection, e.g. [2]. Another popular technique, e.g. used in [18, 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 [9] 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. [4, 6, 7, 16, 17]. As already discussed, we have based our analysis on the work [13] (later extended in [12, 14]) that analyzed utility maximization in a CSMA setting.

Another related paper [15] handles multiple transmission rates in a different model. It has the objective of throughputs optimization in multi-hop MIMO networks under SINR constraints. The paper focuses on local distributed algorithms for finding feasible sets of 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. In our work, we assume that channel conditions vary on a slow time scale enough so that the optimization can be carried out on a static scenario. An interesting open problem is to perform the optimization for fast moving mobile users who could move to completely different locations before the algorithm has fully converged.

9 Conclusion

In this paper we have presented a CSMA-based scheduling algorithm for heterogeneous 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. In our work, we assume that channel conditions vary on a slow time scale enough so that the optimization can be carried out on a static scenario. An interesting open problem is to perform the optimization for fast moving mobile users who could move to completely different locations before the algorithm has fully converged.

References

1. 3GPP TR 36.921. Evolved Universal Terrestrial Radio Access (E-UTRA); FDD Home eNode B (HeNB) Radio Frequency (RF) requirements analysis.
2. Andrews, M., Capdevielle, V., Feki, A., & Gupta, P. (2010). Self-organizing and self-optimizing networks. Bell Labs Technical Journal, 15(3), 85–97.
3. Borst, S., Markakis, M., & Saniei, I. (2011). 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.
4. Borst, S. C., Ghaderi, J., & Whiting, P. A. (2012). Backlog-based random access in wireless networks: Fluid limits and instability issues (invited paper). In Proceedings of CISS 2012 and WiOpt 2012.
5. Chen, C.S., Baccelli, F., & Roullet, L. (2011). Joint optimization of radio resources in small and macro cell networks. In Proceedings of IEEE Vehicular Technology Conference, pp. 1–5.
6. Ghaderi, J., & Srikant, R. (2010). On the design of efficient CSMA algorithms for wireless networks. In Proceedings of CDC.
7. Ghaderi, J., & Srikant, R. (2012). Effect of access probabilities on the delay performance of Q-CSMA algorithms. In INFOCOM.
8. Hou, I., & Gupta, P. (2011). Distributed resource allocation for proportional fairness in multi-band wireless systems. In Proceedings of the IEEE international symposium on information theory.
9. Jiang, L., & Walrand, J. (2008). A CSMA distributed algorithm for throughput and utility maximization in wireless networks. In Proceedings of the conference on communication, control and computing.
10. Karn, P. (1990). MACA—A new channel access method for packet radio. In Proceedings of the 9th ARRL computer networking conference, London, ON, Canada.
11. Kelly, F. (1979). Reversibility and stochastic networks. Chichester: Wiley.
12. Lee, J., Yi, Y., Chong, S., Nardelli, B., Knightly, E., & Chiang, M. (2012). Making 802.11 dcf optimal: Design, implementation, and evaluation. CoRR, abs/1207.5370.
13. Liu, J., Yi, Y., Proutiere, A., Chiang, M., & Poor, H. V. (2010). Towards utility-optimal random access without message passing. Wireless Communications and Mobile Computing, 10, 115–128.
14. Proutiere, A., Yi, Y., Lan, T., & Chiang, M. (2010). Resource allocation over network dynamics without timescale separation. In INFOCOM.
15. Qian, D., Zheng, D., Zhang, J., & Shroff, N. (2010). CSMA-based distributed scheduling in multi-hop MIMO networks under SINR model. In INFOCOM.
16. Rajagopalan, S., Shah, D., & Shin, J. (2009). Network adiabatic theorem: an efficient randomized protocol for contention resolution. In Proceedings of SIGMETRICS/Performance 2009, pp. 133–144.
17. Shah, D., & Shin, J. (2010). Delay optimal queue-based CSMA. In Proceedings of SIGMETRICS 2010, pp. 373–374.
18. Son, K., Lee, S., Yi, Y., & Chong, S. (2011). Refim: A practical interference management in heterogeneous wireless access networks. CoRR, abs/1105.0738.
19. Stolyar, A. L., & Viswanathan, H. (2009). Self-organizing dynamic fractional frequency reuse for best-effort traffic through distributed inter-cell coordination. In Proceedings of IEEE INFOCOM 2009.

Matthew Andrews is an IEEE Fellow and a distinguished member of technical staff in the Math of Network Systems Department at Bell Labs in Murray Hill, NJ. He received his B.A., First Class Honors, in Mathematics from Oxford University in 1993 and his Ph.D. in theoretical computer science from the Massachusetts Institute of Technology in 1997. His research interests include network economics and pricing, wireless resource allocation, packet scheduling and approximation algorithms for network design.

He has served on numerous TPCs for networking and theoretical computer science conferences and he holds a range of patents in the area of telecommunications. In 2010 he received the best paper award at the 2010 IEEE Symposium on Foundations of Computer Science.
Lisa Zhang is a Distinguished Member of Technical Staff Bell Labs in Murray Hill, New Jersey. Her research broadly concerns algorithmic and complexity issues of networking, with a focus on design and optimization, routing and scheduling protocols, and stability and Quality-of-Service analyses. She holds multiple patents on methods and devices for efficient communications. She has also served on technical and organizing committees of numerous IEEE and ACM conferences. Dr. Zhang received her B.A., summa cum laude, in mathematics from Wellesley College in 1993 and her Ph.D. in Theory of Computing from the Massachusetts Institute of Technology in 1997. She twice won the Bell Labs President’s Gold Award and the Lucent Chairman’s Award for innovation and technical excellence.