Learning in Markov Game for Femtocell Power Allocation with Limited Coordination

Wenbo Wang*, Pengda Huang[†], Peizhao Hu* and Jing Na[‡]
*Department of Computer Science, Rochester Institute of Technology, NY, USA
[†]University of Electronic Science and Technology of China, Chengdu, China
[‡]Faculty of Engineering, University of Bristol, Bristol, UK

***Abstract*—In this paper, we study the power allocation problem for the downlink transmission in a set of closed-access femtocells which overlay a number of macrocells. We introduce a mutli-step pricing mechanism for the macrocells to control the cross-tier interference by femtocell transmissions without explicit coordination. We model the cross-tier joint power allocation process in the heterogeneous network as a non-cooperative, average-reward Markov game. By investigating the structure of the instantaneous payoff functions in the game, we propose a self-organized strategy learning scheme based on learning automata for both the macrocell base stations and the femtocell access points to adapt their transmit power simultaneously. We prove that the proposed learning scheme is able to find a pure-strategy Nash equilibrium of the game without the need for the femtocell access points to share any local information. Simulation results show the efficiency of the proposed learning scheme.**

I. Introduction

Recent years have seen a surge of mobile data traffic demand with the explosive growth of smart mobile terminals. According to Cisco's most recent forecast on mobile data usage [1], the global mobile data traffic will increase nearly eightfold between 2015 and 2020. However, mobile network connection speeds will increase only threefold by 2020. To cope with the problem of data explosion, deploying additional low-power, short-range femtocell Base Stations (BSs) in the networks becomes an appropriate solution for improving spatial spectrum reuse and delivering higher link throughput.

Femtocell BSs, also known as Femtocell Access Points (FAPs), are low-cost, plug-and-play BSs deployed by terminal consumers with backhaul connections. By overlaying the traditional macrocell in a small area, femtocells are expected to be able to off-load traffic for the users who are far from the Macro BS (MBS) or experiencing significant indoor penetration losses. However, the random and dense co-channel deployment of FAPs could induce significant cross-tier interference from femtocells to macrocells, or inter-cell interference between femtocells, hence undermining the capacity of the network [2]. Moreover, due to unplanned FAP deployment by end users, it will be difficult to mitigate interference through traditional network planning and optimization techniques. As a result, self-organization of FAPs becomes a primary consideration for network designers to control the cross-tier interference as well as the inter-femtocell interference. In [3], a distributed downlink power adaptation mechanism is proposed based on the analysis of outage probabilities of the Orthogonal Frequency Division Multiple Access (OFDMA)-based macro and femto cells. In [4], a teaching (docition) process is introduced based on decentralized reinforcement learning for FAPs to control their downlink power levels in a non-stationary wireless environment. In [5]–[7], the cross-tier and inter-cell interactions are formulated as non-cooperative [5], [6] or coalition games [7]. Accordingly, decentralized, iterative solutions are proposed for the networks to reach the equilibrium operation point with respect to different game-based models.

In this paper, we consider the downlink operation of a closed-access femtocell set which overlays a number of macrocells. Compared with the existing studies in the literature, we introduce monetary incentive for the macrocells to adaptively enforce the interference level caused by the links from each FAPs to their subscribed Femto User Equipments (FUEs) according to the macrocell traffic load. By only allowing limited information exchange from macrocells to femtocells, we formulate the power allocation process in femtocells as a stochastic game. In order for the FAPs to properly adjust their transmit power in a self-organized manner without the need of co-tier information exchange, we introduce a distributed strategy learning mechanism based on Learning Automata (LA) [8]. Theoretical analysis shows that the proposed LA-based power allocation scheme is able to reach a pure-strategy Nash Equilibrium (NE). Numerical simulation results show that the proposed scheme is able to provides a better link throughput than the potential game-based power allocation algorithm without a pricing mechanism.

II. Problem Formulation

A. Network Model

We consider the downlink transmission of a two-tier heterogeneous network containing M MBS and N FAPs ($N > M$). The FAPs operate in closed-access manner and underlay the macrocell band with bandwidth W. For analytical tractability, we assume that the channel is block-fading to all the links and remains constant during each transmission block. The inter-cell interference among the macrocells is not negligible and the MBSs are able to coordinate for joint power allocation. Due to random FAP deployment, One FAP does not have access to the other FAPs' local strategy information, and information exchange only happens cross-tier between MBSs and FAPs. In order to maintain the QoS of links to Macrocell User Equipments (MUEs), the MBSs expect the cross-tier interference from the FAPs to be kept below an acceptable level. MBSs and FAPs are able to adapt their transmit power by choosing power levels from a discrete power level set. We denote the power level set for MBSs by $\mathcal{P}^M = \{\bar{p}_1^M, \ldots, \bar{p}_{|\mathcal{P}^M|}^M\}$ and the power level set for FAPs by $\mathcal{P}^F = \{0, \bar{p}_1^F, \ldots, \bar{p}_{|\mathcal{P}^F|}^F\}$.

Let m denote the index of a transmitter-receiver pair in MBS-MUE pair set $\mathcal{M}$ and n denote the index of a transmitter-receiver pair in FAP-FUE pair set $\mathcal{N}$. We denote the channel power gain between the MBS of link i and the MUE of link m ($i, m \in \mathcal{M}$) by $h_{i,m}^{MM}$, the channel power gain between FAP j and FUE n ($j, n \in \mathcal{N}$) by $h_{j,n}^{FF}$, the channel power gain between MBS m and FUE n by $h_{m,n}^{MF}$ and the channel power gain between FAP n and MUE m by $h_{n,m}^{FM}$. We also denote the power level used by MBS m as p_m^M ($p_m^M \in \mathcal{P}^M$) and the power level used by FAP n as p_n^F ($p_n^F \in \mathcal{P}^F$). We assume that a link i, $i \in \mathcal{M} \cup \mathcal{N}$, experiences an additive white Gaussian noise with variance σ_i^2. Then, the Signal-to-Interference-plus-Noise-Ratio (SINR) of FAP-FUE link $n \in \mathcal{N}$ can be measured as follows at time interval t:

$$\gamma_n^F(\mathbf{p}^M(t), \mathbf{p}^F(t)) = \frac{h_{n,n}^{FF}(t) p_n^F(t)}{\sigma_n^2 + \sum\limits_{m \in \mathcal{M}} h_{m,m}^{MF}(t) p_m^M(t) + \sum\limits_{j \in \mathcal{N} \setminus \{n\}} h_{j,m}^{FF}(t) p_j^F(t)}, \tag{1}$$

where $\mathbf{p}^M(t) = [p_1^M(t), \ldots, p_{|\mathcal{M}|}^M(t)]^T$ is the vector of MBS transmit powers and $\mathbf{p}^F(t) = [p_1^F(t), \ldots, p_{|\mathcal{N}|}^F(t)]^T$ is the vector of FAP transmit powers. Similarly, the SINR of MBS-MUE link $m \in \mathcal{M}$ can be measured as follows at time interval t:

$$\gamma_m^M(\mathbf{p}^M(t), \mathbf{p}^F(t)) = \frac{h_{m,m}^{MM}(t) p_m^M(t)}{\sigma_m^2 + \sum\limits_{i \in \mathcal{M} \setminus \{m\}} h_{i,m}^{MM}(t) p_i^M(t) + \sum\limits_{n \in \mathcal{N}} h_{n,m}^{FM}(t) p_n^F(t)}. \tag{2}$$

We consider that the FAPs are self-centric and are interested only in maximizing their individual payoffs. Based on Shannon's capacity, the throughput of FAP-FUE link n can be expressed as:

$$r_n^F(\mathbf{p}^M(t), \mathbf{p}^F(t)) = W \log\left(1 + \gamma_n^F(\mathbf{p}^M(t), \mathbf{p}^F(t))\right). \tag{3}$$

On the other hand, each MBS monitors its individual traffic load, which is jointly determined by its transmit rate and the incoming traffic rate. Similar to (3), the throughput of a MBS-MUE link can be measured by

$$r_m^M(\mathbf{p}^M(t), \mathbf{p}^F(t)) = W \log\left(1 + \gamma_m^M(\mathbf{p}^M(t), \mathbf{p}^F(t))\right). \tag{4}$$

Without loss of generality, we consider that the packet arrival at each MBS m is independent and can be modeled as a Poisson process with an unknown arrival rate d_m. Let L denote the packet buffer capacity at each MBS, $l_m(t)$ denote the packet buffer length of MBS m at time interval t, and B denote the packet length in bits. Then, we can model the buffer state evolution at MBS m as follows:

$$l_m(t+1) = \min\left(L, \left(l_m(t) + D_m - \frac{T}{B} r_m^M(\mathbf{p}^M(t), \mathbf{p}^F(t))\right)^+\right), \tag{5}$$

where T is the duration of one time interval, D_m is the number of arrived packets during time interval t, and $(x)^+ = \max(x, 0)$.

In order to compensate for the performance loss due to cross-tier interference caused by FAP transmissions, we introduce a biased pricing mechanism for each MBS to charge an FAP according to the interference that it causes. At each MBS, multi-step pricing is adopted, and the interference price is set by the MBS according to its current buffer usage level. We divide the buffer length of each MBS into a number of ranges to indicate the critical level of local buffer usage:

$$c_m(t) = \begin{cases} 0, & \text{if } 0 \le l_m(t) < \bar{l}_1, \\ 1, & \text{if } \bar{l}_1 \le l_m(t) < \bar{l}_2, \\ \ldots & \\ C, & \text{if } \bar{l}_{C-1} \le l_m(t) < \bar{l}_C, \end{cases} \tag{6}$$

where $\bar{l}_i$ ($1 \le i \le C$) is the threshold for buffer critical level i. Let $\lambda_m(c_m(t))$ represent the unit interference price associated with buffer usage level $c_m(t)$. Then, the net revenue of FAP n at time interval t is

$$u_n(t) = r_n^F(\mathbf{p}^M(t), \mathbf{p}^F(t)) - \sum_{m \in \mathcal{M}} \lambda_m(c_m(t)) h_{n,m}^{FM} p_n^F. \tag{7}$$

We note that the price incentive may not necessarily be virtual, and the monetary transaction can be easily implemented by charging the subscriber FUEs of each FAP according to the data service they receive.

B. Power Allocation as a Markov Game

Based on our discussion on the individual link payoffs and MBS buffer queue states, we are ready to formulate the cross-tier power allocation process as a discrete-time Markov game. Mathematically, a discrete-time Markov game is defined as a 5-tuple Multi-Agent Markov Decision Process (MAMDP) [9]:

Definition 1. *A general Markov game is defined by a 5-tuple:* $G = \langle \mathcal{K}, \mathcal{A}, \mathcal{S}, \mathbf{u}, \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a}) \rangle$*, in which*

1) $\mathcal{K}$ *is the set of agent participating in the game.*
2) $\mathcal{A}$ *is the space of joint action* $\mathbf{a}$, $\mathbf{a} = [a_1, \ldots, a_{|\mathcal{K}|}]^T$ *and is the composition of each agent's local action* a_k, $k \in \mathcal{K}$.
3) $\mathcal{S}$ *is the space of system state* $\mathbf{s}$ *as a random state variable vector, whose transition is determined by an underlying controlled Markov chain defined by* $\Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a})$.
4) $\mathbf{u} = [u_1, \ldots, u_{|\mathcal{K}|}]^T$ *is the vector of the agents' instantaneous payoff,* $u_k(t) = u_k(\mathbf{s}(t), \mathbf{a}(t))$ *at time interval* t.
5) $\Pr(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ *is the state transition probability function controlled by joint action* $\mathbf{a}$.

Following the MAMDP-based definition of Markov games, we formulate the cross-tier power allocation process in the heterogeneous network as a 5-tuple $G_p = \langle \mathcal{K}, \mathcal{P}, \mathcal{S}, \mathbf{u}, \Pr(\mathbf{s}'|\mathbf{s}, \mathbf{p}) \rangle$, in which

- $\mathcal{K}$ is the union of all MBS-MUE links and FAP-FUE links, $\mathcal{K} = \mathcal{M} \cup \mathcal{N}$.
- $\mathcal{P}$ is the space of joint power level vector $\mathbf{p} = (\mathbf{p}^M, \mathbf{p}^F)$.

- $\mathcal{S} = \times\mathcal{S}_m$, $m \in \mathcal{M}$, is the Cartesian product of each MBS's buffer state space and $s_m = l_m \in \mathcal{S}_m = \{0, 1, \dots, L\}$ corresponds to a feasible buffer length.
- For macrocell link $i \in \mathcal{M}$, the local instantaneous payoff is defined as $u_i(\mathbf{s}, \mathbf{p}^M, \mathbf{p}^F) = r_i^M(\mathbf{p}^M(t), \mathbf{p}^F(t))$. For femtocell link $i \in \mathcal{N}$, the local instantaneous payoff is given by (7).
- $\Pr(\mathbf{s}'|\mathbf{s}, \mathbf{p})$ is the state transition map controlled by the joint power level $\mathbf{p}$.

According to Definition 1, G_p is a Markov game if the state transition map $\Pr(\mathbf{s}'|\mathbf{s}, \mathbf{p})$ is a well-defined probability function. For two buffer queue states s_m and s'_m of MBS-MUE link m, let $\delta l = s'_m - s_m + T/Br_m^M(\mathbf{p}^M, \mathbf{p}^F)$ given a fixed power allocation $\mathbf{p} = (\mathbf{p}^M, \mathbf{p}^F)$. With the packet arrival at each MBS following independent Poisson process with arrival rate d_m, we can express the transition probability function between each state of link m based on (5) as follows:

- if $\dfrac{T}{B} r_m^M(\mathbf{p}) - s_m + L > 0$ and $0 < s'_m < L$,
$$\Pr(s'_m|s_m, \mathbf{p}) = \frac{(d_m)^{\delta l} \exp(d_m)}{\delta l!}, \tag{8}$$
- if $s'_m = L$,
$$\Pr(s'_m|s_m, \mathbf{p}) = \sum_{k=\delta l}^{\infty} \frac{(d_m)^k \exp(d_m)}{k!}, \tag{9}$$
- if $s_m \le \dfrac{T}{B} r_m^M(\mathbf{p})$ and $s'_m = 0$,
$$\Pr(s'_m|s_m, \mathbf{p}) = \sum_{k=0}^{\delta l} \frac{(d_m)^k \exp(d_m)}{k!}, \tag{10}$$
- otherwise $\Pr(s'_m|s_m, \mathbf{p}) = 0$.

With (8)-(10) we can easily check that $\Pr(s'_m|s_m, \mathbf{p})$ is a well-defined probability function with any fixed power allocation $\mathbf{p}$: $0 \le \Pr(s'_m|s_m, \mathbf{p}) \le 1$ and $\sum_{s'_m} \Pr(s'_m|s_m, \mathbf{p}) = 1$. Therefore, G_p is a well-defined Markov game.

Observing the instantaneous payoff functions in game G_p, we note that the payoff function for MBS-MUE link m given by (4), $u_m(\mathbf{s}, \mathbf{p}^M, \mathbf{p}^F) = r_m^M(\mathbf{p}^M, \mathbf{p}^F)$, is independent of system state $\mathbf{s}_m$. On the other hand, the payoff function for FAP n can be written in the form of
$$u_n(\mathbf{s}, \mathbf{p}^M, \mathbf{p}^F) = r_n^F(\mathbf{p}^M, \mathbf{p}^F) + g_n(\mathbf{s}, p_n^F), \tag{11}$$
where $g_n(\mathbf{s}, p_n^F) = \sum_{m\in\mathcal{M}} \lambda_m(s_m) h_{n,m}^{FM} p_n^F$. With such observations, we obtain the following property of game G_p:

Theorem 1. *The formulated Markov game G_p is an exact potential game, if (a) each link (player) admits the expected average payoff as its objective, and (b) the underlying Markov chain for any fixed power allocation profile $\mathbf{p}$ is ergodic.*

Proof. If all links admit expected average payoffs, the goal of link i during play is to maximize its local objective [10]:
$$J_i(\boldsymbol{\pi}(\mathbf{p})) = \lim_{\tau\to\infty} \frac{1}{\tau} E_{\boldsymbol{\pi}} \left[\sum_{t=0}^{\tau-1} u_i(\mathbf{s}(t), \mathbf{p}) \right], \tag{12}$$
where $\boldsymbol{\pi}(\mathbf{p})$ is the vector of probabilities for adopting any feasible power allocation profile $\mathbf{p}$. Under the assumption that the underlying Markov chain corresponding to any power allocation profile $\mathbf{p}$ is ergodic, we can show that the state distribution of the Markov chain corresponding to a (mixed) strategy $\boldsymbol{\pi}$ converges to a limiting distribution $\boldsymbol{\alpha} = (\alpha_1(\boldsymbol{\pi}), \dots, \boldsymbol{\alpha}_{|\mathcal{S}|}(\boldsymbol{\pi}))$, $\forall i$, $\alpha_i(\boldsymbol{\pi}) > 0$ [10]. Then, with starting state $\mathbf{s}$, (12) can be rewritten as:
$$J_i(\boldsymbol{\pi}(\mathbf{p})) = \sum_{\mathbf{s}\in\mathcal{S}} \alpha_{\mathbf{s}}(\boldsymbol{\pi}) \sum_{\mathbf{s}'\in\mathcal{S}} \Pr(\mathbf{s}'|\mathbf{s}, \boldsymbol{\pi}) E_{\boldsymbol{\pi}} \left[u_i(\mathbf{s}, \mathbf{p}) \right]. \tag{13}$$

We check the property of $J_i(\boldsymbol{\pi}(\mathbf{p}))$ with a deterministic strategy $\mathbf{p}$. According to the definition of a potential game [9], Markov game G_p in the matrix form with payoff $J_i(p_i, p_{-i})$ given by (13) is an exact potential game if there exists an exact potential function $\Phi(\mathbf{p})$ such that $\forall i \in \mathcal{K}$, $\forall p_i, p'_i$, $\forall \mathbf{p} \in \mathcal{P}$,
$$\Phi(p_i, p_{-i}) - \Phi(p'_i, p_{-i}) = J_i(p_i, p_{-i}) - J_i(p'_i, p_{-i}), \tag{14}$$
where p_{-i} is the joint power allocation profile of the adversary links of link i. By examining (4) and (7), we can define the instantaneous potential function for MBSs and for FAPs as (15) and (16), respectively. In what follows, we omit the superscripts in p_m^M ($m \in \mathcal{M}$) and p_n^F ($n \in \mathcal{N}$) for conciseness:
$$\begin{aligned} &\phi_i(\mathbf{s}, p_i, p_{-i}) = \\ &W \log \left(\sum_{m\in\mathcal{M}} h_{mi}^{MM} p_m + \sum_{n\in\mathcal{N}} h_{ni}^{FM} p_n + \sigma_i^2 \right), \forall i \in \mathcal{M}, \end{aligned} \tag{15}$$
$$\begin{aligned} \phi_i(\mathbf{s}, p_i, p_{-i}) &= f_i(p_i, p_{-i}) - g_i(\mathbf{s}, p_i) \\ &= W \log \left(\sum_{m\in\mathcal{M}} h_{mi}^{MF} p_m + \sum_{n\in\mathcal{N}} h_{ni}^{FF} p_n + \sigma_i^2 \right) \\ &- \sum_{m\in\mathcal{M}} \lambda_m(c_m) h_{i,m}^{FM} p_i, \forall i \in \mathcal{N}. \end{aligned} \tag{16}$$

Since $\forall m \in \mathcal{M}$, local instantaneous payoff u_m is independent of system state $\mathbf{s}$, then for MBS m we can rewrite (13) as
$$\begin{aligned} J_m(\boldsymbol{\pi}(\mathbf{p})) &= \sum_{\mathbf{s}\in\mathcal{S}} \alpha_{\mathbf{s}}(\boldsymbol{\pi}) E_{\boldsymbol{\pi}} \left[u_m(\mathbf{p}) \right] \sum_{\mathbf{s}'\in\mathcal{S}} \Pr(\mathbf{s}'|\mathbf{s}, \boldsymbol{\pi}) \\ &= E_{\boldsymbol{\pi}} \left[u_m(\mathbf{p}) \right]. \end{aligned} \tag{17}$$

Similarly, if we replace the instantaneous payoff function u_i in (13) by the instantaneous potential function ϕ_i, we can obtain the potential function for the matrix game as
$$\Phi_m(\boldsymbol{\pi}(\mathbf{p})) = E_{\boldsymbol{\pi}} \left[\phi_m(\mathbf{p}) \right]. \tag{18}$$

For a deterministic policy $\boldsymbol{\pi}$ with $\pi_i(p_i) = 1$, where p_i is an element of $\mathbf{p}$ ($p_i \in \mathcal{P}^M$ for MBSs and $p_i \in \mathcal{P}^F$ for FAPs), we have $J_m(\boldsymbol{\pi}(\mathbf{p})) = u_m(\mathbf{p})$. Then,
$$\begin{aligned} &J_m(\pi_m, \pi_{-m}) - J_m(\pi'_m, \pi_{-m}) = u_m(p_m, p_{-m}) - u_m(p'_m, p_{-m}) \\ &= W \log(1 + \gamma_m^M(p_m, p_{-m})) - W \log(1 + \gamma_m^M(p'_m, p_{-m})) \\ &= W \log(\frac{I_m^M + h_{m,m}^{MM} p_m}{I_m^M + h_{m,m}^{MM} p'_m}) = \phi_m(p_m, p_{-m}) - \phi_m(p'_m, p_{-m}) \\ &= \Phi_m(p_m, p_{-m}) - \Phi_m(p'_m, p_{-m}), \end{aligned} \tag{19}$$

where according to (2) I_m^M is the total interference to link m:

$$I_m^M = \sigma_m^2 + \sum_{i\in\mathcal{M}\setminus\{m\}} h_{i,m}^{MM} p_i + \sum_{n\in\mathcal{N}} h_{n,m}^{FM} p_n. \quad (20)$$

For FAP-FUE link $n \in \mathcal{N}$, we note that both the original instantaneous payoff in (11) and the proposed corresponding function in (16) are comprised by two parts. Then, the expected payoff of link n in the matrix form game can be written as:

$$J_n(\boldsymbol{\pi}(\mathbf{p})) = \sum_{\mathbf{s}\in\mathcal{S}} \alpha_{\mathbf{s}}(\boldsymbol{\pi}) \sum_{\mathbf{s}'\in\mathcal{S}} \Pr(\mathbf{s}'|\mathbf{s},\boldsymbol{\pi}) E_{\boldsymbol{\pi}}\left[r_n^F(\mathbf{p}) + g_n(\mathbf{s},\mathbf{p})\right]. \quad (21)$$

Again, since $r_n^F(\mathbf{p})$ in (21) is independent of system state $\mathbf{s}$, we can use the same technique for proving the potential function of MBS-MAP links as in (17)-(19) and show that $f_n(p_i, p_{-i})$ given in (16) defines an exact potential function for the expected average payoff related to $r_n^F(\mathbf{p})$ in the matrix-form game. Since both (11) and (16) share the same term $g_n(\mathbf{s},\mathbf{p})$, we have:

$$\begin{aligned} &J_n(\pi_n, \pi_{-n}) - J_n(\pi_n', \pi_{-n}) = f_n(p_n, p_{-n}) - f_n(p_n', p_{-n}) \\ &+ \sum_{\mathbf{s}\in\mathcal{S}} \alpha_{\mathbf{s}}(p_i, p_{-i}) \sum_{\mathbf{s}'\in\mathcal{S}} \Pr(\mathbf{s}'|\mathbf{s}, p_i, p_{-i}) E_{\boldsymbol{\pi}(\mathbf{p})}\left[g_i(\mathbf{s}, p_i)\right] \\ &- \sum_{\mathbf{s}\in\mathcal{S}} \alpha_{\mathbf{s}}(p_i', p_{-i}) \sum_{\mathbf{s}'\in\mathcal{S}} \Pr(\mathbf{s}'|\mathbf{s}, p_i', p_{-i}) E_{\boldsymbol{\pi}(\mathbf{p}')}\left[g_i(\mathbf{s}, p_i')\right] \\ &= \Phi_n(p_n, p_{-n}) - \Phi_n(p_n', p_{-n}). \end{aligned} \quad (22)$$

Then, (16) defines an exact potential function for FAP-FUE link n in the matrix-game form of G_p. □

With Theorem 1, the following property holds for game G_p:

Corollary 1. *If the two conditions in Theorem 1 are satisfied, then the Markov power allocation game, G_p, has at least one pure-strategy Nash equilibrium.*

Proof. Corollary 1 immediately follows Theorem 1 and Corollary 3.1 of [9]. □

III. Self-Organized Learning for Cross-tier Power Allocation

Based on Theorem 1 and Corollary 1, we are ready to develop a self-organized strategy learning mechanism for both MBSs and FAPs in the network. In this section, we propose to apply learning automata for the BSs to simultaneously adapt their transmit power without the need of explicit coordination.

A. LA-based Equilibrium Learning with Limited Coordination

In Markov game G_p, link i aims at finding the best-response to the joint adversary power allocation strategy π_{-i} as:

$$\pi_i^*(p_i|\pi_{-i}) = \arg\max_{\pi_i} J_i(\pi_i, \pi_{-i}), \quad (23)$$

and if $\forall i \in \mathcal{K}$, $\pi_i(p_i|\pi_{-i}) = \pi_i^*$, $\boldsymbol{\pi}$ is an NE. We consider that each MBS is able to broadcast its buffer state information to the nearby FAPs in the network. Besides, the only information that an FAP can obtain is the achieved link throughput and the payment to be made to the macrocells. In order for FAPs and MBSs to learn the NE of the power allocation game without the need of any other information exchange, we introduce the LA-based Linear Reward Inaction (L_{R-I}) scheme [8] for power allocation strategy learning. With a generalized L_{R-I} scheme, Learning agent i updates its action-taking policy purely based on the local payoff that it observes:

$$\begin{cases} \pi_i^{t+1}(p_i) = \pi_i^t(p_i) + \theta\beta^t(\mathbf{p}(t))(1 - \pi_i^t(p_i)) & \text{if } p_i(t) = p_i, \\ \pi_i^{t+1}(p_i') = \pi_i^t(p_i') - \theta\beta^t(\mathbf{p}(t))\pi_i^t(p_i') & \text{if } p_i(t) \neq p_i', \end{cases} \quad (24)$$

where θ is the learning step size and $\beta(t)$ ($0 \leq \beta(t) \leq 1$) is the normalized payoff of LA agent i.

Instead of associating a single LA with one MBS-MUE or FAP-FUE link, we propose to associate an automaton with each system state $\mathbf{s} \in \mathcal{S}$ for each link in game G_p. In this case, for a network containing $|\mathcal{M}|$ MBSs with buffer capacity L, a number of $L^{|\mathcal{M}|}$ LA are to be created for each link. Based on (24), let $LA_i^{\mathbf{s}}$ denote the policy updating process associated with state $\mathbf{s}$ for link i. Then, at time interval t, only one automaton, $LA_i^{\mathbf{s}(t)}$, is activated on MBS/FAP i for policy updating. We note that with a finite discrete power level set for each link, the throughput of MBS-MUE link m is upper-bounded. Then for MBS m, the normalized payoff is defined as follows:

$$\beta_m^t(\mathbf{p}(t)) = \frac{\sum_{\tau=1}^t r_m^M(\mathbf{p}(\tau))}{t \cdot \overline{r}_m^M}, \quad (25)$$

where $\overline{r}_m^M = \max_{\mathbf{p}} r_m^M(\mathbf{p})$ is the maximum throughput that MBS-MUE link m can achieve. For FAP-FUE link n, considering that the link revenue may be negative when the nearby MBS-MUE links charges link n with a high interference price, we define its normalized payoff as follows:

$$\beta_n^t(\mathbf{p}(t)) = \frac{\sum_{\tau=1}^t (u_n^F(\mathbf{s}(\tau), \mathbf{p}(\tau)) - \underline{u}_n^F)}{t \cdot (\overline{u}_n^F - \underline{u}_n^F)}, \quad (26)$$

where $\overline{u}_n^F = \max_{\mathbf{s},\mathbf{p}} u_n^F(\mathbf{s},\mathbf{p})$ and $\underline{u}_n^F = \min_{\mathbf{s},\mathbf{p}} u_n^F(\mathbf{s},\mathbf{p})$.

For a general Markov game defined by Definition 1, a group of independent LA that are associated with the state-agent pairs guarantees to converge to a pure-strategy NE, if the conditions in Theorem 2 are satisfied:

Theorem 2 (Corollary 1 of [10])**.** *If an average-reward Markov game has a pure-strategy NE point and for any joint policy its underlying multi-agent Markov chain is ergodic, then by associating each state for each agent an LA, the L_{R-I} scheme given in (24) is guaranteed to find the pure-strategy NE point.*

According to Corollary 1 and Theorem 2, the learning scheme given by (24)-(26) is able to find a pure-strategy NE of game G_p as long as the underlying Markov chain defined by (8)-(10) is ergodic for any pure-strategy power allocation profile. By the definition of ergodic Markov chain, we only need to show that $\Pr(s_i'|s_i, \mathbf{p}) > 0, \forall i, \forall s_i, s_i \in \mathcal{S}_i$ and $\forall \mathbf{p}$. Thus, to derive the condition of always being an ergodic Markov chain, we only need to examine the boundary cases of smallest throughput for each MBS. Then, a sufficient condition for game G_p to be ergodic is given as follows:

$$\frac{h_{m,m}^{MM}\overline{p}_1^M}{\sigma_m^2 + \sum_{i\in\mathcal{M}\setminus\{m\}} h_{i,m}^{MM}\overline{p}_{|\mathcal{P}^M|}^M + \sum_{n\in\mathcal{N}} h_{n,m}^{FM}\overline{p}_{|\mathcal{P}^F|}^F} \geq 2^{\frac{L}{W}} - 1. \quad (27)$$

B. Approximate State Space Reduction

Although the learning scheme given by (24)-(26) does not need extra information exchange for reaching a pure-strategy NE, it still faces a problem of state space explosion as the number of MBSs in the network increases. A natural consideration for state space reduction is to partition the state space with a coarse granularity, e.g., by using the partition scheme given in (6). With such a state partition scheme, the space of the aggregated states, $\mathcal{C} = \{0, 1, \dots, C\}$, changes the original state transition map into the follows:

$$\hat{\Pr}(c'_m|c_m, \mathbf{p}) = \sum_{s'_m \in \psi_m^{-1}(c'_m)} \sum_{s_m \in \psi_m^{-1}(c_m)} \Pr(s'_m|s_m, \mathbf{p}), \quad (28)$$

where $\psi_m : \mathcal{S}_m \to \mathcal{C}_m$ is a mapping from the state space of s_m to the new state space of c_m. Since $\hat{\Pr}(c'_m|c_m, \mathbf{p})$ is ill-defined as a probability function, a weighting factor is needed to be imposed onto $\hat{\Pr}(c'_m|c_m, \mathbf{p})$ in order to obtain a well-defined probability function:

$$\Pr(c'_m|c_m, \mathbf{p}) = \frac{1}{\sum\limits_{s_m \in \psi^{-1}(c_m)} \alpha_{s_m}(\mathbf{p})} \hat{\Pr}(c'_m|c_m, \mathbf{p}). \quad (29)$$

Since the state aggregation does not really change the state transitions in the real world, we call the new Markov game with aggregated states $\mathbf{c} = [c_1, \dots, c_{|\mathcal{M}|}]^T$ the virtual game G'_p. We note that state aggregation preserves the ergodic property of the original MAMDP G_p. However, due to state aggregation, a value manipulation based on the original payoff in game G_p is needed for the virtual game to preserve the same expected average payoff of a link i with respect to the same joint policy as in G_p (see Lemma 1 in [11]). Let u'_i denote the new instantaneous payoff in the virtual game. For a deterministic policy $\mathbf{p}$, we have

$$u'_i(\mathbf{c}, \mathbf{p}) = \sum_{\mathbf{s}} \frac{\mathbb{1}(\psi(\mathbf{s}), \mathbf{c})\boldsymbol{\alpha}_{\mathbf{s}}(\mathbf{p})}{\sum\limits_{\mathbf{x} \in \psi^{-1}(\mathbf{c})} \boldsymbol{\alpha}_{\mathbf{x}}(\mathbf{p})} u_i(\mathbf{s}, \mathbf{p}), \quad (30)$$

where $\mathbb{1}(x, y)$ is the indicator function, $\mathbb{1}(x, y) = 1$ if $x = y$ and $\mathbb{1}(x, y) = 0$ if $x \neq y$. $\boldsymbol{\alpha}_{\mathbf{x}}$ is the joint distribution with respect to state vector $\mathbf{x}$.

Fortunately, the structure of the payoff functions for MBS-MUE links in (4) and FAP-FUE links in (7) guarantees that the exact potential function still exists for $u'_i(\mathbf{c}, \mathbf{p})$ in the virtual game[1]. Since the limiting distribution of joint state $\mathbf{s}$, $\boldsymbol{\alpha}_{\mathbf{s}}$, is unknown to the on-line learning scheme, we use the following scheme to estimate the value of u'_i:

$$\hat{u}'_i(\mathbf{c}, \mathbf{p}, t) = \frac{\sum_{\tau=1}^{t} \mathbb{1}(\mathbf{c}(\tau), \mathbf{c}) u_i(\tau)}{\sum_{\tau=1}^{t} \mathbb{1}(\mathbf{c}(\tau), \mathbf{c})}. \quad (31)$$

Then, $\lim\limits_{t \to \infty} \hat{u}'_i(\mathbf{c}, \mathbf{p}, t) = u'_i(\mathbf{c}, \mathbf{p})$. We can apply the same L_{R-I} scheme given by (24)-(26) to the virtual game G'_p with the value of u'_i estimated as (31). However, it is worth noting that although the new state value $u'_i(\mathbf{c}, \mathbf{p})$ for a fixed policy $\mathbf{p}$ preserves the value of link i's payoff for the same policy in the original game G_p, the NE of game G'_p is generally not the same as the NE of game G_p. Therefore, the NE learned based on the virtual game G'_p is an approximation of the NE of the original game G_p.

IV. Simulation Results

In our simulations, we consider a network where MBSs are deployed in a hexagonal grid with a node distance of 500m, and a number of FAPs are randomly deployed in the same region. We assume that the maximum interference distance from an FAP to an MBS is 250m. Then, an FAP only needs to keep track of at most two nearest MBSs' buffer states. The simulation parameters are given in Table I. The channel gains are generated by a lognormal shadowing pathloss model as a function of the node distance, $h_{i,j} = D_{i,j}^{-k}$, where k is the pathloss factor, $k = 2.7$ for cross-tier links and $k = 2.2$ for other links.

TABLE I
Main Parameters Used in the Femtocell Network Simulation

Parameter	Value
Shared Bandwidth W	1MHz
Feasible region for MBS transmit power p_m^M	$[10, 30]$dBm
Feasible region for FAP transmit power p_n^F	$[0, 24]$dBm
AWGN power σ^2	-40dBm
Buffer queue capacity	1×10^4
Bits per packet (bpp)	1000bpp
Packet arrival rate d_m	250

We first demonstrate in Figure 1 the converge property of the proposed strategy learning scheme for a network of 7 MBSs and 100 randomly deployed FAPs. We consider that both MBSs and FAPs adopt an action set of 6 power levels. The buffer usage is evenly divided into 4 levels (states), and their corresponding interference price are set to 100, 1000, 1×10^5 and 1×10^7, respectively. From Figure 1, we observe that when the LA associated with each state-link pair converges (see Figure 1c), the femto links are able to achieve a slightly higher throughput than the macro links (see Figure 1a). Such performance is achieved in the condition that the network is not too crowded, and the FAPs are able to transmit with a relatively high power level. From Figure 1b, we can also infer that for most of the time the buffer usage is kept at a medium level at each MBS.

In Figure 2, we compare the performance of the proposed learning scheme with that of LA-based learning without a pricing mechanism. We note from (15) and (16) that without pricing, the power allocation process is reduced to a single-state repeated game and each link is associated with only one automaton. In the simulation, buffer usage is divided into 6 levels (states). Figure 2a shows that with interference pricing, the throughput of femtocell links can be improved by at most 380%. Also, when FAPs are densely deployed in the network, learning with pricing is able to better prevent performance deterioration of the femtocell links than without pricing. Meanwhile, Figure 2b shows that with LA-based NE searching, MBSs may have a better throughput as the number of

[1] We omit the proof in this paper due to space limit. The proof can be derived in the same way as the proof of Theorem 1.

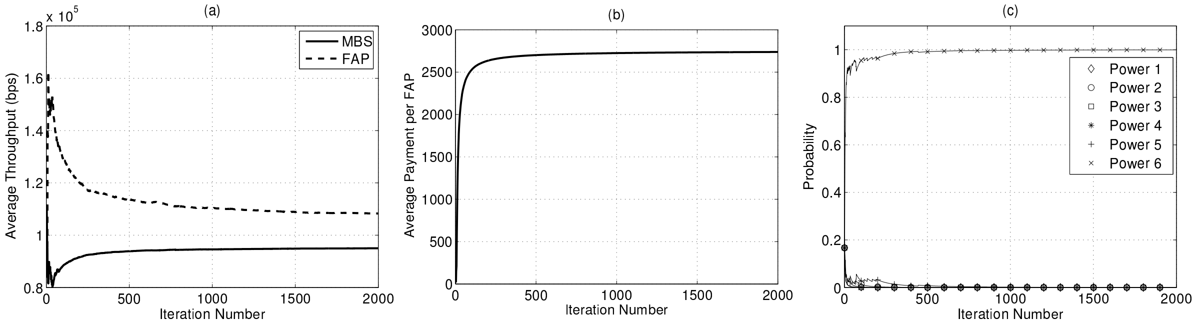

Fig. 1. (a) Evolution of average throughput. (b) Evolution of average payment of each FAP. (c) Policy evolution for FAP 1 at the most frequently visited state.

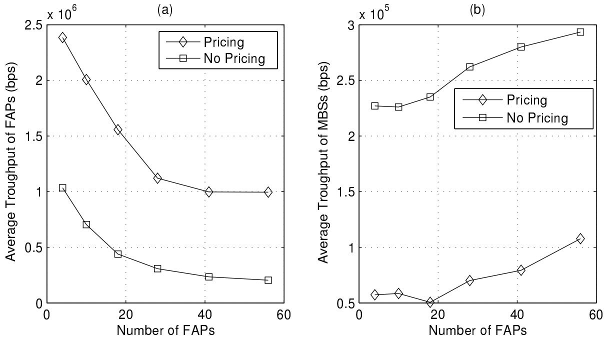

Fig. 2. (a) Average throughput of FAPs vs. number of FAPs in the network. (b) Average throughput of MBSs vs. number of FAPs in the network.

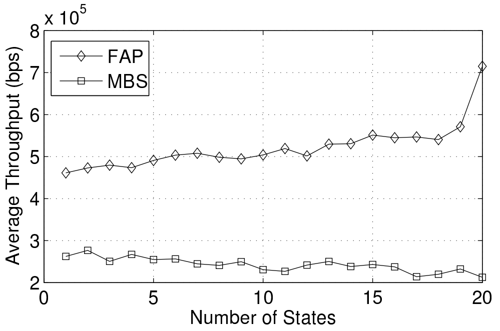

Fig. 3. FAP/MBS performance vs. state number at the approximate NE.

nearby FAPs increases. Intuitively, with a higher FAP density, the increasing inter-cell interference will drive the FAPs to greatly reduce their transmit power in order to reach the new NE. As a result, the received cross-tier interference of marcocell links will get smaller, hence a better throughput of the MBSs.

In Figure 3, we investigate the impact of state aggregation on the performance of the proposed learning scheme. The simulation is performed in a network with 1 MBS and 15 FAPs. The interference price is obtained from a range $[0, 1\times10^{13}]$. Figure 3 shows that as the buffer state size (equivalently, the number of price levels) increases, the social performance of the FAPs also increases. Figure 3 indicates that a finer price granularity will lead to a larger number of feasible options for FAPs' power allocation. Then, it is possible for the NE to move to a new point that leads to better FAP performance.

V. Conclusion

In this paper, we have studied the power allocation problem for the downlink transmission of a two-level, overlay heterogeneous network. We have introduced a multi-step pricing mechanism for macrocell base stations to implicitly control the cross-tier interference from femtocells. By considering the buffer load evolution at each macrocell base station as a stochastic process, we have formulated the cross-tier power allocation process as a Markov game. We have proposed an LA-based, distributed strategy learning mechanism for both the macrocell and the femtocell base stations to autonomously learn the pure-strategy equilibrium of the game. We have also proposed a state space aggregation scheme in order to address the problem of state space explosion in the proposed learning process. Simulation results have shown that LA-based learning with proper state aggregation is able to find a pure-strategy Nash Equilibrium and improve the femtocell performance by 130%-380% compared with the learning scheme without pricing.

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020," Tech. Rep., Feb. 2016.

[2] D. Lopez-Perez, A. Valcarce, G. de la Roche, and J. Zhang, "Ofdma femtocells: A roadmap on interference avoidance," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 41–48, Sep. 2009.

[3] X. Chu, Y. Wu, D. Lopez-Perez, and X. Tao, "On providing downlink services in collocated spectrum-sharing macro and femto networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 12, pp. 4306–4315, Dec. 2011.

[4] A. Galindo-Serrano, L. Giupponi, and M. Dohler, "Cognition and docition in ofdma-based femtocell networks," in *Proc. IEEE GLOBECOM'10*, Miami, FL, USA, 2010, pp. 1–6.

[5] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3202–3212, Jul. 2013.

[6] P. Semasinghe, E. Hossain, and K. Zhu, "An evolutionary game for distributed resource allocation in self-organizing small cells," *IEEE Trans. Mobile Comput.*, vol. 14, no. 2, pp. 274–287, Feb. 2015.

[7] R. Langar, S. Secci, R. Boutaba, and G. Pujolle, "An operations research game approach for resource and power allocation in cooperative femtocell networks," *IEEE Trans. Mobile Comput.*, vol. 14, no. 4, pp. 675–687, Apr. 2015.

[8] R. Wheeler and K. Narendra, "Decentralized learning in finite markov chains," *IEEE Trans. Autom. Control*, vol. 31, no. 6, pp. 519–526, Jun. 1986.

[9] Z. Han, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge University Press, 2012.

[10] P. Vrancx, K. Verbeeck, and A. Nowe, "Decentralized learning in markov games," *IEEE Trans. Syst., Man, Cybern.: Cybern.*, vol. 38, no. 4, pp. 976–981, Aug. 2008.

[11] Z. Ren and B. H. Krogh, "State aggregation in markov decision processes," in *Proc. IEEE CDC'02*, vol. 4, Las Vegas, NV, Dec. 2002, pp. 3819–3824 vol.4.