Power-Delay Tradeoff With Predictive Scheduling in Integrated Cellular and Wi-Fi Networks

Haoran Yu, *Student Member, IEEE*, Man Hon Cheung, Longbo Huang, *Member, IEEE*, and Jianwei Huang, *Fellow, IEEE*

***Abstract*—The explosive growth of global mobile traffic has led to rapid growth in the energy consumption in communication networks. In this paper, we focus on the energy-aware design of the network selection, subchannel, and power allocation in cellular and Wi-Fi networks, while taking into account the traffic delay of mobile users. Based on the two-timescale Lyapunov optimization technique, we first design an online Energy-Aware Network Selection and Resource Allocation (ENSRA) algorithm, which yields a power consumption within $O\left(\frac{1}{V}\right)$ bound of the optimal value, and guarantees an $O(V)$ traffic delay for any positive control parameter V. Motivated by the recent advancement in the accurate estimation and prediction of user mobility, channel conditions, and traffic demands, we further develop a novel predictive Lyapunov optimization technique to utilize the predictive information, and propose a Predictive Energy-Aware Network Selection and Resource Allocation (P-ENSRA) algorithm. We characterize the performance bounds of P-ENSRA in terms of the power-delay tradeoff theoretically. To reduce the computational complexity, we finally propose a Greedy Predictive Energy-Aware Network Selection and Resource Allocation (GP-ENSRA) algorithm, where the operator solves the problem in P-ENSRA approximately and iteratively. Numerical results show that GP-ENSRA significantly improves the power-delay performance over ENSRA in the large delay regime. For a wide range of system parameters, GP-ENSRA reduces the traffic delay over ENSRA by 20–30% under the same power consumption.**

***Index Terms*—Energy-aware communication, joint network selection and resource allocation, cellular and Wi-Fi integration, stochastic optimization.**

I. Introduction

With the explosive growth of global mobile data traffic, the energy consumption in communication networks has increased significantly. According to [2], the information and communications technology industry constituted 2% of global CO_2 emissions. In addition, the high energy consumption in communication networks accounts for a significant proportion of the operational expenditure (OPEX) to the mobile operators [3]. Therefore, mobile operators have the incentives to reduce the energy consumption, through innovations in several areas such as novel hardware design, efficient resource management, and dynamic base station activations [4], [5].

In this paper, we focus on the problem of energy-aware *network selection* and *resource allocation* (i.e., *subchannel and power allocation*). First, since Wi-Fi networks often consume less energy than the macrocell network due to their smaller coverages and shorter communication distances [6], the operator of an integrated cellular and Wi-Fi network can significantly reduce the system energy consumption by offloading part of the cellular traffic to the Wi-Fi networks. Second, within the cellular network, the operator can reduce the transmission power while maintaining the system throughput by allocating the subchannels and power to the cellular users with good channel conditions.

In the first part of this paper, we apply the two-timescale Lyapunov optimization technique [7] to design an online *Energy-Aware Network Selection and Resource Allocation* (ENSRA) algorithm. We show that ENSRA yields a power consumption that can be pushed arbitrarily close to the optimal value, at the expense of an increase in the average traffic delay.

In the second part of this paper, motivated by the recent advancement of accurate estimation of users' mobilities, traffic demands, and channel conditions, we improve the performance of ENSRA by incorporating the prediction of the system randomness into the algorithm design. We design a *Predictive Energy-Aware Network Selection and Resource Allocation* (P-ENSRA) algorithm through a novel predictive Lyapunov optimization technique. Different from the previous Lyapunov optimization techniques in [7], [8], we introduce a novel control parameter θ to optimize the operations within the entire information window. By properly adjusting θ, we can balance the variance of queue length within each information window, and significantly improve the delay performance.

To reduce the computational complexity of P-ENSRA, we further propose a *Greedy Predictive Energy-Aware Network Selection and Resource Allocation* (GP-ENSRA) algorithm,

Manuscript received April 15, 2015; revised September 7, 2015; accepted December 11, 2015. Date of publication March 21, 2016; date of current version May 11, 2016. The work of H. Yu, M. H. Cheung, and J. Huang was supported by the General Research Funds (Project Number CUHK 412713 and 14202814) established under the University Grant Committee of the Hong Kong Special Administrative Region, China. The work of L. Huang was supported in part by the National Basic Research Program of China under Grant 2011CBA00300 and Grant 2011CBA00301, in part by the National Natural Science Foundation of China under Grant 61033001, Grant 61361136003, and Grant 61303195, in part by the Tsinghua Initiative Research Grant, in part by the Microsoft Research Asia Collaborative Research Award, and in part by the China Youth 1000-talent Grant. This paper was presented in part at IEEE GLOBECOM, Austin, TX, USA, December 2014 [1] *(Corresponding author: Jianwei Huang.)*

H. Yu, M. H. Cheung, and J. Huang are with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: yh012@ie.cuhk.edu.hk; mhcheung@ie.cuhk.edu.hk; jwhuang@ie.cuhk.edu.hk).

L. Huang is with the Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China (e-mail: longbohuang@tsinghua.edu.cn).

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

Digital Object Identifier 10.1109/JSAC.2016.2544639

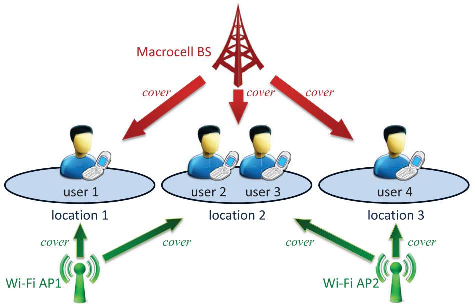


Fig. 1. An example of the system model, where user 1, 2, 3, and 4 are moving within the set of locations $\mathcal{S} = \{1, 2, 3\}$. The macrocell covers all locations. Each location is covered by a set of Wi-Fi networks, *e.g.*, $\mathcal{N}_1 = \{1\}$, $\mathcal{N}_2 = \{1, 2\}$, $\mathcal{N}_3 = \{2\}$.

where the operator solves the optimization problem in P-ENSRA approximately and iteratively. Our numerical results show that GP-ENSRA achieves a much better power-delay tradeoff than ENSRA in the large delay regime, and the improvement increases with the prediction window size.

There are many literatures studying either energy-aware network selection or energy-aware resource allocation problems. For example, Venturino *et al.* in [9] studied energy-efficient resource allocation and base station coordination in a static downlink cellular system. Xiong *et al.* in [10] investigated energy-efficient resource allocation under quality-of-service constraints, in a static cellular system with both downlink and uplink communications. However, these literatures did not consider joint energy-aware network selection and resource allocation in the stochastic cellular and Wi-Fi networks, which is the focus of our work.

II. System Model

We consider the downlink transmission in a slotted system, indexed by $t \in \{0, 1, \ldots\}$. We focus on the monopoly case, where the single operator serves users by its own macrocell and Wi-Fi networks. We introduce the following notations:

- $\mathcal{L} \triangleq \{1, 2, \ldots, L\}$: set of the users;
- $\mathcal{N} \triangleq \{1, 2, \ldots, N\}$: set of the Wi-Fi networks;
- $\mathcal{S} \triangleq \{1, 2, \ldots, S\}$: set of the locations.

We assume that the macrocell base station covers all S locations, and we use $\mathcal{N}_s \subseteq \mathcal{N}$ to denote the set of available Wi-Fi networks at location $s \in \mathcal{S}$. We illustrate the system model through an example in Figure 1.

A. Two-Timescale Operations

The operator aims at reducing the total power consumption through the network selection, subchannel allocation, and power allocation. We assume that the network selection is operated in a larger timescale than the subchannel and power allocation. This is because a frequent switch among different networks interrupts the data delivery and incurs a nonnegligible cost (*e.g.*, in the form of energy consumption, quality-of-service degradation, and delays).

We refer every T time slots as a *frame*, and define the k-th frame ($k \in \mathbb{N}$) as the time interval that contains a set $\mathcal{T}_k \triangleq \{kT, kT + 1, \ldots, kT + T - 1\}$ of time slots. We assume that:

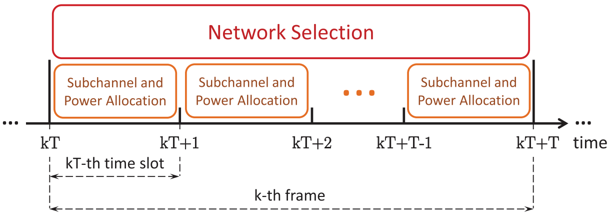

Network Selection
Subchannel and Power Allocation
Subchannel and Power Allocation
Subchannel and Power Allocation
kT
kT+1
kT+2
kT+T-1
kT+T
time
kT-th time slot
k-th frame

Fig. 2. Two-timescale operations: (a) at time $t = kT$, *e.g.*, the beginning of the k-th frame, the operator determines the network selection for the k-th frame; (b) at time $t \in \mathcal{T}_k$, the operator determines the subchannel and power allocation for time slot t.

- the operator determines network selection at the beginning of every frame (*large-timescale*);
- the operator determines subchannel and power allocation at the beginning of every time slot (*small-timescale*).

We illustrate such a two-timescale structure in Figure 2.

B. Frame-Based Network Selection

At time slot $t = kT$, *i.e.*, the beginning of the k-th frame, the operator determines the network selection for the k-th frame. We denote the network selection by $\boldsymbol{\alpha}(kT) = (\alpha_l(kT), \forall l \in \mathcal{L})$, where $\alpha_l(kT)$ indicates the network that user l is connected to during the k-th frame. Let the random variable $S_l(kT) \in \mathcal{S}$ be user l's location during the k-th frame, and define $\boldsymbol{S}(kT) = (S_l(kT), \forall l \in \mathcal{L})$.[1] Since the availabilities of Wi-Fi networks are location-dependent, we have the following constraint for $\boldsymbol{\alpha}(kT)$:

$$\alpha_l(kT) \in \mathcal{N}_{S_l(kT)} \cup \{0\}, \forall l \in \mathcal{L}, k = 0, 1, \ldots, \tag{1}$$

where selection $\alpha_l(kT) = 0$ indicates that user l is connected to the macrocell network.

C. Macrocell Network Model

We consider an Orthogonal Frequency Division Multiplexing (OFDM) system for the macrocell network, following the standard model as used in [11], [12].

1) Subchannel Allocation: Let $\mathcal{M} \triangleq \{1, 2, \ldots, M\}$ be the set of subchannels, and denote the subchannel allocation by $\boldsymbol{x}(t) = (x_{lm}(t), \forall l \in \mathcal{L}, m \in \mathcal{M})$. Variable $x_{lm}(t) \in \{0, 1\}$ for all l and m: if user l is allocated with subchannel m, $x_{lm}(t) = 1$; otherwise, $x_{lm}(t) = 0$. We assume that each subchannel can at most be allocated to one user:

$$\sum_{l=1}^{L} x_{lm}(t) \leq 1, \quad \forall m \in \mathcal{M}. \tag{2}$$

Different from the frame-based network selection $\boldsymbol{\alpha}(kT)$, the operator determines the subchannel allocation $\boldsymbol{x}(t)$ every time slot. Since the operator can only allocate subchannels to those users who are connected to the cellular network, we have the following constraint for $\boldsymbol{x}(t)$:

$$\alpha_l(t_T) x_{lm}(t) = 0, \quad \forall l \in \mathcal{L}, m \in \mathcal{M}, t \geq 0. \tag{3}$$

[1] User locations $\boldsymbol{S}(kT)$ do not change during the frame. The reason is that the user location usually changes much less frequently than the other types of randomness, *e.g.*, the channel condition in the macrocell network.

Here, $t_T \triangleq \left\lfloor \frac{t}{T} \right\rfloor T$ is the beginning of the frame that time slot t belongs to, and network selection $\alpha_l(t_T)$ indicates user l's associated network during the frame.

2) Power Allocation: We denote the power allocation by $\boldsymbol{p}(t) = (p_{lm}(t), \forall l \in \mathcal{L}, m \in \mathcal{M})$. Variable $p_{lm}(t) \geq 0$ denotes the power allocated to user l on subchannel m. We have the following power budget constraint:

$$\sum_{m=1}^{M} \sum_{l=1}^{L} p_{lm}(t) \leq P_{\max}^{C}, \quad \forall t \geq 0. \tag{4}$$

Similar as (3), the operator can only allocate the power to those users who are connected to the cellular network. We have the following constraint for $\boldsymbol{p}(t)$:

$$\alpha_l(t_T) p_{lm}(t) = 0, \quad \forall l \in \mathcal{L}, m \in \mathcal{M}, t \geq 0. \tag{5}$$

3) Macrocell Transmission Rate: We use $\boldsymbol{H}(t) = (H_{lm}(t), \forall l \in \mathcal{L}, m \in \mathcal{M})$ to denote the channel conditions, where $H_{lm}(t)$ is a random variable that represents the channel condition for user l on subchannel m at time slot t. Given the subchannel allocation $\boldsymbol{x}^l(t) = (x_{lm}(t), \forall m \in \mathcal{M})$ and power allocation $\boldsymbol{p}^l(t) = (p_{lm}(t), \forall m \in \mathcal{M})$, the transmission rate of a cellular user l (*i.e.*, $\alpha_l(t_T) = 0$) at time slot t is

$$r_l^C\left(\boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right) = \frac{B}{M} \sum_{m=1}^{M} x_{lm}(t) \log_2 \left(1 + \frac{p_{lm}(t) H_{lm}^2(t)}{N_0 \frac{B}{M}}\right), \tag{6}$$

where B is the total bandwidth and N_0 is the noise power spectral density.

4) Macrocell Power Consumption: According to [13], the power consumption of the macrocell base station contains two components: the first component is a fixed term that measures the radio frequency (RF) and baseband unit power consumptions; the second component corresponds to the transmission power. Since the first component is fixed, in our model, we focus on minimizing the time average of the second component, which is given by

$$P^C(\boldsymbol{p}(t)) = \kappa \sum_{m=1}^{M} \sum_{l=1}^{L} p_{lm}(t). \tag{7}$$

Here, parameter κ is the scale factor that depends on the power amplifier efficiency and the losses incurred by the antenna feeder, power supply, and cooling [13].

D. Wi-Fi Network Model

Let ρ_n be the number of users associated with Wi-Fi network n. We assume that Wi-Fi network n's total transmission rate and power consumption are functions of ρ_n, and we denote them by $R_n(\rho_n)$ and $P_n^W(\rho_n)$, respectively. We further assume that $R_n(\rho_n)$ and $P_n^W(\rho_n)$ are non-negative bounded functions, *i.e.*, there exist positive constants $R_{n,\max}$ and $P_{n,\max}^W$ such that

$$0 \leq R_n(\rho_n) \leq R_{n,\max} \text{ and } 0 \leq P_n^W(\rho_n) \leq P_{n,\max}^W \tag{8}$$

for all $\rho_n = 0, 1, 2, \ldots$.

We allow general functions of $R_n(\rho_n)$ and $P_n^W(\rho_n)$ that satisfy (8) in our algorithm design in Sections IV and V.

1) Wi-Fi Transmission Rate: Given function $R_n(\rho_n)$ and network selection $\boldsymbol{\alpha}(t_T)$, we can compute the transmission rate of a Wi-Fi user l (*i.e.*, $\alpha_l(t_T) > 0$) at time slot t by [14]:

$$r_l^W(\boldsymbol{\alpha}(t_T)) = \frac{R_{\alpha_l(t_T)}\left(\sum_{k=1}^{L} \mathbb{1}_{\{\alpha_k(t_T)=\alpha_l(t_T)\}}\right)}{\sum_{k=1}^{L} \mathbb{1}_{\{\alpha_k(t_T)=\alpha_l(t_T)\}}}. \tag{9}$$

Here, summation $\sum_{k=1}^{L} \mathbb{1}_{\alpha_k(t_T)=\alpha_l(t_T)}$ returns the number of users in the Wi-Fi network that user l is associated with.[2]

2) Wi-Fi Power Consumption: Given function $P_n^W(\rho_n)$ and network selection $\boldsymbol{\alpha}(t_T)$, we can compute the power consumption of all Wi-Fi networks as:

$$P^W(\boldsymbol{\alpha}(t_T)) = \sum_{n=1}^{N} P_n^W\left(\sum_{l=1}^{L} \mathbb{1}_{\{\alpha_l(t_T) = n\}}\right). \tag{10}$$

E. Users' Traffic Model

We assume that the users randomly generate traffic, and the traffic generation is not affected by the operator's operations. We use a random variable $A_l(t)$ to denote the traffic arrival rate of user $l \in \mathcal{L}$ at time slot t, and let $\boldsymbol{A}(t) = (A_l(t), l \in \mathcal{L})$. We assume that there exists a positive constant $A_{\max}$ such that

$$0 \leq A_l(t) \leq A_{\max}, \quad \forall l \in \mathcal{L}, t \geq 0. \tag{11}$$

F. Summary

1) Macrocell + Wi-Fi Transmission Rate: If a user is associated with the macrocell network, its transmission rate is given by $r_l^C\left(\boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right)$ in (6); if it is associated with Wi-Fi networks, its transmission rate is given by $r_l^W(\boldsymbol{\alpha}(t_T))$ in (9). To summarize, user l's transmission rate at time slot t is given by

$$r_l\left(\boldsymbol{\alpha}(t_T), \boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right) = \begin{cases} r_l^C\left(\boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right), & \text{if } \alpha_l(t_T) = 0, \\ r_l^W(\boldsymbol{\alpha}(t_T)), & \text{otherwise.} \end{cases} \tag{12}$$

Because of the power budget constraint (4) in the macrocell network, function $r_l^C\left(\boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right)$ is upper bounded. Furthermore, since Wi-Fi networks' total transmission rates are upper bounded as in (8), function $r_l^W(\boldsymbol{\alpha}(t_T))$ is also upper bounded. As a result, there exists a positive constant $r_{\max}$ such that

$$0 \leq r_l\left(\boldsymbol{\alpha}(t_T), \boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right) \leq r_{\max} \tag{13}$$

for all $l \in \mathcal{L}$ and $\boldsymbol{\alpha}(t_T), \boldsymbol{x}^l(t), \boldsymbol{p}^l(t)$ satisfying (1), (2), (3), (4), and (5).

2) Macrocell + Wi-Fi Power Consumption: The operator considers the power consumption in both the macrocell and Wi-Fi networks. The macrocell network's power consumption

[2] $\mathbb{1}_{\{\cdot\}}$ is the indicator function, which equals 1 if the event in the brace is true, and equals 0 if the event is false.

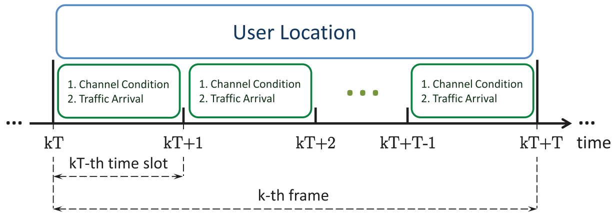


Fig. 3. Two-timescale randomness: (a) users' locations change every frame; (b) channel conditions and users' traffic arrivals change every time slot.

is given by $P^C(\boldsymbol{p}(t))$ in (7), and Wi-Fi networks' total power consumption is given by $P^W(\boldsymbol{\alpha}(t_T))$ in (10). Therefore, the operator's total power consumption at time slot t is given by

$$P(\boldsymbol{\alpha}(t_T), \boldsymbol{p}(t)) = P^C(\boldsymbol{p}(t)) + P^W(\boldsymbol{\alpha}(t_T)). \tag{14}$$

According to the cellular power budget constraint (4) and the bounded Wi-Fi power consumption condition (8), it is easy to find that $P(\boldsymbol{\alpha}(t_T), \boldsymbol{p}(t))$ is bounded:

$$0 \le P(\boldsymbol{\alpha}(t_T), \boldsymbol{p}(t)) \le P_{\max}, \quad \forall t \ge 0, \tag{15}$$

where $P_{\max} \triangleq \kappa P^C_{\max} + \sum_{n=1}^{N} P^W_{n,\max}$.

3) Randomness: There are three kinds of randomness in the system:

- Users' locations $\boldsymbol{S}(kT)$, introduced in Section II-B;
- The macrocell network's channel conditions $\boldsymbol{H}(t)$, introduced in Section II-C3;
- Users' traffic arrivals $\boldsymbol{A}(t)$, introduced in Section II-E.

As we assumed in Section II-B, $\boldsymbol{S}(kT)$ changes at the beginning of each frame, while $\boldsymbol{H}(t)$ and $\boldsymbol{A}(t)$ change every time slot. The two-timescale randomnesses is in Figure 3.

III. Problem Formulation

We assume that each user has a data queue, the length of which denotes the amount of unserved traffic. Let $\boldsymbol{Q}(t) = (Q_l(t), \forall l \in \mathcal{L})$ be the queue length vector, where $Q_l(t)$ is user l's queue length at time slot t. We assume that all queues are initially empty, *i.e.*,

$$Q_l(0) = 0, \forall l \in \mathcal{L}. \tag{16}$$

The queue length evolves according to the traffic arrival rate and transmission rate as

$$Q_l(t+1) = \left[Q_l(t) - r_l\left(\boldsymbol{\alpha}(t_T), \boldsymbol{x}^l(t), \boldsymbol{p}^l(t)\right)\right]^+ + A_l(t), \forall l \in \mathcal{L}, t \ge 0. \tag{17}$$

Algorithm 1. Energy-Aware Network Selection and Resource Allocation (ENSRA)

```
Set $t = 0$ and $\boldsymbol{Q}(0) = \boldsymbol{0}$;
while $t < t_{end}$ do
    if mod $(t, T) = 0$
        Set $k = \frac{t}{T}$ and solve problem (19) to determine $\boldsymbol{\alpha}(kT), \boldsymbol{x}(\tau), \boldsymbol{p}(\tau), \forall \tau \in \mathcal{T}_k$;
    end if
    Update $\boldsymbol{Q}(t+1)$, according to (17);
    $t \leftarrow t + 1$.
end while
```

Here $[x]^+ = \max\{x, 0\}$ is due to the fact that the actual amount of served packets cannot exceed the current queue size.

The objective of the operator is to design an online network selection and resource allocation algorithm that minimizes the expected time average power consumption, while keeping the network stable. This can be formulated as the following optimization problem:

$$\begin{aligned}
\min \quad & \overline{P} \triangleq \limsup_{K\to\infty} \frac{1}{KT} \sum_{t=0}^{KT-1} \mathbb{E}\{P(\boldsymbol{\alpha}(t_T)), \boldsymbol{p}(t)\} \\
\text{s.t.} \quad & \overline{Q_l} \triangleq \limsup_{K\to\infty} \frac{1}{KT} \sum_{t=0}^{KT-1} \mathbb{E}\{Q_l(t)\} < \infty, \forall l \in \mathcal{L}, \\
& \text{constraints}(1),\ (2),\ (3),\ (4),\ (5), \\
\text{var.} \quad & \boldsymbol{\alpha}(t_T), \boldsymbol{x}(t), \boldsymbol{p}(t), \quad \forall t \ge 0.
\end{aligned} \tag{18}$$

Here, $\overline{Q_l}$ is user l's time average queue length, and constraint $\overline{Q_l} < \infty$ for all $l \in \mathcal{L}$ ensures the stability of the network.

IV. Network Selection and Resource Allocation Without Prediction

A. Energy-Aware Network Selection and Resource Allocation (ENSRA) Algorithm

We assume that the operator has the complete information for the channel conditions within the current frame, *i.e.*, at time slot $t = kT$ (the beginning of the k-th frame), the operator has the information of $\boldsymbol{H}(\tau)$ for all $\tau \in \mathcal{T}_k$. We leave the algorithm design for the incomplete channel information in [15]. We present ENSRA in Algorithm 1. The detailed solution to (19), shown at the bottom of the page, is provided in [15]. The intuition behind ENSRA is that, by adjusting the control parameter $V > 0$, the operator can achieve a good tradeoff between the power consumption and the traffic delay.

$$\begin{aligned}
\min \quad & V \sum_{\tau=kT}^{kT+T-1} P(\boldsymbol{\alpha}(kT), \boldsymbol{p}(\tau)) - \sum_{l=1}^{L} Q_l(kT) \sum_{\tau=kT}^{kT+T-1} r_l\left(\boldsymbol{\alpha}(kT), \boldsymbol{x}^l(\tau), \boldsymbol{p}^l(\tau)\right) \\
\text{s.t.} \quad & \text{constraints } (1),\ (2),\ (3),\ (4),\ (5), \\
\text{var.} \quad & \boldsymbol{\alpha}(kT), \boldsymbol{x}(\tau), \boldsymbol{p}(\tau), \forall \tau \in \mathcal{T}_k.
\end{aligned} \tag{19}$$

B. Performance Analysis of ENSRA

For ease of exposition, we analyze the performance of ENSRA by assuming that the system randomness is independent and identically distributed (i.i.d.). Notice that with the technique developed in [16], we can obtain similar results under Markovian randomness.

We define the capacity region Λ as the closure of the set of arrival vectors that can be stably supported, considering all network selection and resource allocation algorithms. We assume that the mean traffic arrival is strictly interior to Λ, *i.e.*, there exists an $\eta > 0$ such that

$$\mathbb{E}\left\{\boldsymbol{A}\left(t\right)\right\} + \eta \cdot \boldsymbol{1} \in \Lambda. \tag{20}$$

We use P_{av}^{*} to denote the optimal expected time average power consumption of problem (18). The performance of ENSRA is described in the following theorem.

Theorem 1: ENSRA achieves:

$$P_{av}^{\mathsf{ENSRA}} \triangleq \limsup_{K\to\infty} \frac{1}{KT} \sum_{t=0}^{KT-1} \mathbb{E}\{P(\boldsymbol{\alpha}(t_T), \boldsymbol{p}(t))\} \le P_{av}^{*} + \frac{\Omega}{V}, \tag{21}$$

$$Q_{av,T}^{\mathsf{ENSRA}} \triangleq \limsup_{K\to\infty} \frac{1}{K} \sum_{l=1}^{L} \sum_{k=0}^{K-1} \mathbb{E}\{Q_l\left(kT\right)\} \le \frac{\Omega + V P_{\max}}{\eta}. \tag{22}$$

where $\Omega = \frac{1}{2} T L \left(A_{\max}^2 + r_{\max}^2\right)$, $P_{\max}$ is defined in (15), and η is defined in (20).

Theorem 1 implies that, by increasing parameter V, the operator can push the power consumption arbitrarily close to the optimal value, *i.e.*, P_{av}^{*}, but at the expense of the increase in the average traffic delay.

V. Network Selection and Resource Allocation With Prediction

We study the situation where the operator can predict the system randomness for the future frames. With the predictive future information, the operator is able to achieve better performance than ENSRA.

A. Information Prediction Model

We consider the structure of the *prediction window*, where the window size W is the number of frames in a window.

Algorithm 2. Predictive Energy-Aware Network Selection and Resource Allocation (P-ENSRA)

```
Set $t = 0$ and $\boldsymbol{Q}(0) = \boldsymbol{0}$;
while $t < t_{end}$ do
    if mod $(t, WT) = 0$
        Set $h = \frac{t}{WT}$ and solve problem (23) to determine
        $\boldsymbol{\alpha}(hWT + wT), w = 0, 1, \ldots, W-1, \boldsymbol{x}(\tau), \boldsymbol{p}(\tau), \tau \in \mathcal{W}_h$;
    end if
    Update $\boldsymbol{Q}(t+1)$, according to (17);
    $t \leftarrow t + 1$.
end while
```

Thus, we define the h-th ($h \in \{0, 1, \ldots\}$) window as the time interval that contains frames $\mathcal{T}_{hW}, \mathcal{T}_{hW+1}, \ldots, \mathcal{T}_{hW+W-1}$. We use $\mathcal{W}_h \triangleq \mathcal{T}_{hW} \cup \mathcal{T}_{hW+1} \cup \ldots \cup \mathcal{T}_{hW+W-1}$ to define the set of time slots within the h-th window. Equivalently, we have $\mathcal{W}_h = \{hWT, hWT+1, \ldots, hWT+WT-1\}$.

We assume that at time slot $t = hWT$, *i.e.*, the beginning of the h-th window, the operator accurately predicts the system randomness for the whole window: (a) $\boldsymbol{S}\left(hWT + wT\right), w = 0, 1, \ldots, W-1$, where $\boldsymbol{S}\left(hWT + wT\right)$ denotes users' locations during frame $\mathcal{T}_{hW+w}$; (b) $\boldsymbol{H}\left(\tau\right), \boldsymbol{A}\left(\tau\right), \tau \in \mathcal{W}_h$, where $\boldsymbol{H}\left(\tau\right)$ and $\boldsymbol{A}\left(\tau\right)$ denote users' channel conditions and traffic arrivals at time slot τ, respectively.

At time slot $t = hWT$, with the predictive information, the operator runs P-ENSRA or GP-ENSRA, and determines the operations for the whole window: (a) $\boldsymbol{\alpha}\left(hWT + wT\right), w = 0, 1, \ldots, W-1$, where $\boldsymbol{\alpha}\left(hWT + wT\right)$ denotes the network selection during frame $\mathcal{T}_{hW+w}$; (b) $\boldsymbol{x}\left(\tau\right), \boldsymbol{p}\left(\tau\right), \tau \in \mathcal{W}_h$, where $\boldsymbol{x}\left(\tau\right)$ and $\boldsymbol{p}\left(\tau\right)$ are the subchannel allocation and power allocation at time slot τ, respectively.

B. Predictive Energy-Aware Network Selection and Resource Allocation (P-ENSRA) Algorithm

We propose P-ENSRA in Algorithm 2. The basic idea of ENSRA in Section IV is to minimize the upper bound of the "drift-plus-penalty" term for a frame. Different from ENSRA, P-ENSRA guarantees a θ-controlled upper bound on the "drift-plus-penalty" term instead of minimizing the "drift-plus-penalty" term for a window. This is because P-ENSRA determines the network selection and resource allocation for several frames (*i.e.*, a window), and it needs to use a novel control parameter $\theta > 0$ to balance the queue lengths among different frames. With parameter θ, we can assign larger

$$\begin{aligned}
\min \quad & V \sum_{w=0}^{W-1} \sum_{\tau=(hW+w)T}^{(hW+w+1)T-1} P\left(\boldsymbol{\alpha}\left(hWT + wT\right), \boldsymbol{p}\left(\tau\right)\right) + \sum_{l=1}^{L} \sum_{w=0}^{W-1} Q_l\left(hWT + wT\right) \sum_{\tau=(hW+w)T}^{(hW+w+1)T-1} \left(A_l\left(\tau\right) + \theta\right) \\
& - \sum_{l=1}^{L} \sum_{w=0}^{W-1} Q_l\left(hWT + wT\right) \sum_{\tau=(hW+w)T}^{(hW+w+1)T-1} r_l\left(\boldsymbol{\alpha}\left(hWT + wT\right), \boldsymbol{x}^l\left(\tau\right), \boldsymbol{p}^l\left(\tau\right)\right) \\
\text{s.t.} \quad & \text{constraints (1), (2), (3), (4), (5),} \\
\text{var.} \quad & \boldsymbol{\alpha}\left(hWT + wT\right), w = 0, 1, \ldots, W-1, \boldsymbol{x}\left(\tau\right), \boldsymbol{p}\left(\tau\right), \tau \in \mathcal{W}_h.
\end{aligned} \tag{23}$$

weights to the transmission rates of the earlier frames than those of the latter frames within a prediction window, and thus reduce the time average queue length.

C. Performance Analysis of P-ENSRA

Similar as ENSRA, we characterize the performance of P-ENSRA under the i.i.d. system randomness and assume that the condition (20) is satisfied. We define $P_{av}^{\text{P-ENSRA}}$ as the expected time average power consumption of P-ENSRA, and define $Q_{av,T}^{\text{P-ENSRA}}$ as the expected time average value of user queue length at the beginning of each frame under P-ENSRA. The performance of P-ENSRA is described as follows.

Theorem 2: P-ENSRA achieves

$$P_{av}^{\text{P-ENSRA}} \triangleq \limsup_{H\to\infty} \frac{1}{HWT} \sum_{t=0}^{HWT-1} \mathbb{E}\left\{P\left(\boldsymbol{\alpha}\left(t_T\right), \boldsymbol{p}\left(t\right)\right)\right\} \leq P\left(\theta\right) + \frac{\Omega}{V}, \quad (24)$$

$$Q_{av,T}^{\text{P-ENSRA}} \triangleq \limsup_{H\to\infty} \frac{1}{HW} \sum_{l=1}^{L} \sum_{h=0}^{HW-1} \mathbb{E}\{Q_l\left(hT\right)\} \leq \frac{\Omega + VP\left(\theta\right)}{\theta}, \quad (25)$$

for any $V > 0$ and $\theta \in (0, \eta]$, where $P(\theta)$ is defined as the minimum power consumption required to stabilize the traffic arrival vector $\mathbb{E}\{\boldsymbol{A}(t)\} + \theta \cdot \mathbf{1}$, considering all network selection and resource allocation algorithms.

D. Greedy Predictive Energy-Aware Network Selection and Resource Allocation (GP-ENSRA)

In problem (23), shown at the bottom of the previous page, the network selections and resource allocations in different frames are tightly coupled by the queue lengths. Such coupling significantly increases the difficulty of directly solving problem (23). Here, we propose a greedy algorithm, GP-ENSRA, which approximately solves problem (23) for each window and significantly reduces the complexity.

We present GP-ENSRA in Algorithm 3. In order to simplify the description, we use $\boldsymbol{\beta}(hWT+wT) = (\boldsymbol{\alpha}(hWT+wT), \boldsymbol{x}(\tau), \boldsymbol{p}(\tau), \tau \in \mathcal{T}_{hW+w})$ to represent the operator's operations (network selection and resource allocation) over frame $\mathcal{T}_{hW+w}$, $w = 0, 1, \ldots, W-1$. From line (5) to line (12), the operator iteratively updates the operations for all frames within the window. As shown in line (11), we use F^i to denote the value of the objective function in (23) under the i-th iteration. The condition for ending the iteration (line (5)) implies that the decrease from F^{i-1} to F^i is no larger than a positive parameter ϵ. Such a condition is guaranteed to be achievable, and we leave the detailed proof in [15].

Algorithm 3. Greedy Predictive Energy-Aware Network Selection and Resource Allocation (GP-ENSRA)

```
1: Set $t = 0$ and $\boldsymbol{Q}(0) = \mathbf{0}$;
2: while $t < t_{end}$ do
3:     if $\frac{t}{WT} \in \mathbb{N}$
4:         Set $h = \frac{t}{WT}$, $i = 0$, and $\boldsymbol{\beta}(hWT + wT) = \mathbf{0}$, $\forall w = 0, 1, \ldots, W-1$;
5:         while $i < 2$ or $F^{i-1} - F^i > \epsilon$ do
6:             $i \leftarrow i + 1$;
7:             for $w = 0$ to $W - 1$ do
8:                 Minimize the objective function in problem (23) over $\boldsymbol{\beta}(hWT+wT)$ (fix $\boldsymbol{\beta}(hWT+w'T)$ for all $w' \neq w$);
9:                 Update $\boldsymbol{\beta}(hWT + wT)$ with the optimal solution obtained in line (8);
10:            end for
11:            Denote the value of the objective function in (23) under $(\boldsymbol{\beta}(hWT+wT), w = 0, 1, \ldots, W-1)$ by $F^i$;
12:        end while
13:        Output vector $\boldsymbol{\beta}(hWT+wT)$, $w = 0, 1, \ldots, W-1$, as the operations for the window;
14:    end if
15:    Update $\boldsymbol{Q}(t+1)$, according to (17);
16:    $t \leftarrow t + 1$.
17: end while
```

VI. Simulation

A. Simulation Settings

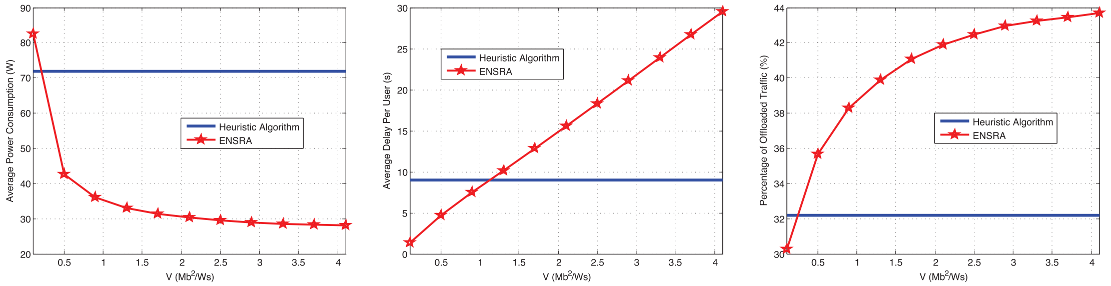
We simulate the problem with $L = 10$ users, 1 macrocell network, $N = 10$ Wi-Fi networks, and $S = 100$ locations. We set the time slot length to be 10 milliseconds, and the frame length to be 1 second, *i.e.*, $T = 100$. We run each experiment in MATLAB for 5,000 frames. We assume that the macrocell network covers all locations, and the channel gain follows the Rayleigh fading [17]. Furthermore, we assume that each Wi-Fi network is randomly distributed spatially, and each Wi-Fi network covers 1 ∼ 4 connected locations. We choose the Wi-Fi transmission rate function from [18], and the power consumption function from [19].

B. Simulation Results

1) Comparison Between ENSRA and Heuristic Algorithm: We compare ENSRA with the following *heuristic algorithm*.

Heuristic algorithm: At the beginning of each frame, the operator first assigns the users who are only covered by the macrocell network or are within 100m of the macrocell network. Then the operator sequentially checks the available Wi-Fi networks for each of the remaining users, and assigns each user to the Wi-Fi network with the lowest number of connected users; at every time slot, the operator determines the resource allocation based on a heuristic method [11].

In Figure 4, we compare ENSRA under different parameter V with the heuristic algorithm. In Figure 4(a), we plot the total power consumption of ENSRA against V. We observe that, as V increases, ENSRA's total power consumption decreases. According to (21), the upper bound of P_{av}^{ENSRA} decreases with the increasing of V, which is consistent with our observation here. Figure 4(a) also shows the total power consumption of the heuristic algorithm, which is independent of V. We notice that ENSRA consumes less power than the heuristic algorithm for any $V > 0.2\ \text{Mb}^2/\text{W} \cdot \text{s}$.

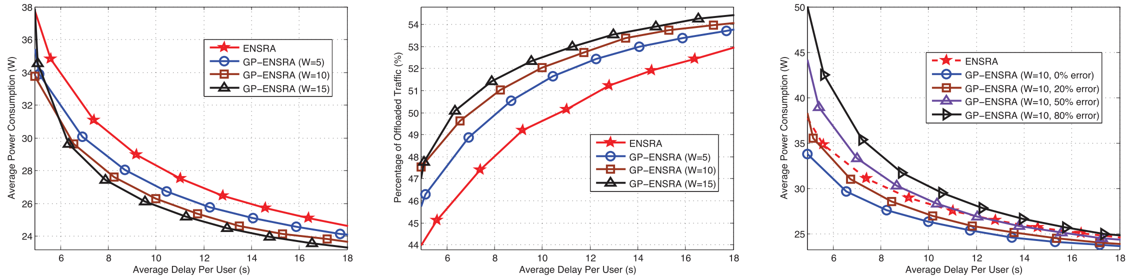


(a) Power vs Parameter V.

(b) Traffic Delay vs Parameter V.

(c) Data Offloading vs Parameter V.

Fig. 4. Comparison of ENSRA and Heuristic Algorithm.



(a) Power-Delay Tradeoff.

(b) Data Offloading vs Average Delay.

(c) Power-Delay Tradeoff under Prediction Errors.

Fig. 5. Comparison of ENSRA and GP-ENSRA.

In Figure 4(b), we plot the average traffic delay per user under ENSRA against V. As V increases, the average delay of ENSRA increases, which is consistent with the result in (22). Compared with the heuristic algorithm, ENSRA generates less delay for any $V < 1.1\ \mathrm{Mb}^2/\mathrm{W}\cdot\mathrm{s}$. Figure 4(a) and Figure 4(b) imply that, if the operator chooses $0.2\mathrm{Mb}^2/\mathrm{W}\cdot\mathrm{s} \leq V \leq 1.1\mathrm{Mb}^2/\mathrm{W}\cdot\mathrm{s}$, ENSRA outperforms the heuristic algorithm in both the power and delay. For example, ENSRA with $V = 0.5\ \mathrm{Mb}^2/\mathrm{W}\cdot\mathrm{s}$ saves 40.8% power and 47.8% delay over the heuristic algorithm.

In Figure 4(c), we plot the percentage of the traffic served in Wi-Fi against V. According to (19), a larger V implies that the operator focuses more on the power consumption than the traffic delay, and ENSRA will delay users' traffic to Wi-Fi networks to reduce the power cost. Hence, in Figure 4(c), the percentage of the traffic served in Wi-Fi increases with V.

2) Comparison Between ENSRA and GP-ENSRA: In Figure 5(a), we plot the average total power consumption against the average traffic delay per user for ENSRA and GP-ENSRA. We obtain these power-delay tradeoff curves by varying V. Comparing ENSRA with GP-ENSRA, we observe that when the traffic delay is above 6 s, GP-ENSRA always generates a smaller power consumption than ENSRA under the same traffic delay. For example, when the generated traffic delay is 8 s, the power consumptions of ENSRA and GP-ENSRA with window size $W = 15$ are 30.4 W and 27.4 W, respectively. Hence, the power saving of GP-ENSRA with $W = 15$ over ENSRA is 9.9%. The performance improvement of GP-ENSRA is more obvious in terms of the delay saving. For example, when the operator pursues a power consumption of 26 W, the average traffic delays under ENSRA and GP-ENSRA with window size $W = 15$ are 13.9 s and 9.7 s, respectively. This shows that GP-ENSRA with window size $W = 15$ saves 30.2% delay over ENSRA. In Figure 5(a), we also observe that the performance improvement increases with the size of the prediction window.

In Figure 5(b), we compare the percentages of the traffic offloaded to Wi-Fi under ENSRA and GP-ENSRA. We plot the percentage of the traffic served in Wi-Fi against the average traffic delay. When generating the same traffic delay, GP-ENSRA offloads a larger percentage of traffic than ENSRA. The reason is that the predictive information helps the operator design a network selection and resource allocation strategy that utilizes Wi-Fi networks more efficiently to reduce the total power consumption.

In Figure 5(c), we investigate the power-delay performance of GP-ENSRA under the prediction errors. For example, GP-ENSRA with 20% prediction error means that for each information (*i.e.*, users' locations, channel conditions, and traffic arrivals) of the future frames, with 0.8 probability the operator accurately predicts its value, while with 0.2 probability the operator obtains an incorrect value of the information. In Figure 5(c), we plot the average power consumption against the average traffic delay per user for ENSRA and GP-ENSRA with window size $W = 10$ under different percentages of the prediction errors. We observe that the power-delay performance of GP-ENSRA declines as the percentage of the prediction errors increases. However, GP-ENSRA with 20% prediction error still achieves a better power-delay tradeoff than the non-predictive algorithm ENSRA, which shows the robustness of GP-ENSRA against the prediction errors.

VII. Conclusion

In this paper, we studied the online network selection and resource allocation problem in the stochastic integrated cellular and Wi-Fi networks. We first proposed the ENSRA algorithm, which can generate a close-to-optimal power consumption at the expense of an increase in the average traffic delay. We then proposed the P-ENSRA algorithm and the GP-ENSRA algorithm by incorporating the prediction of the system randomness into the network selection and resource allocation. In our future work, we plan to analytically characterize the impact of the prediction errors on the predictive algorithms.

References

[1] H. Yu, M. H. Cheung, L. Huang, and J. Huang, "Predictive delay-aware network selection in data offloading," in *Proc. IEEE GLOBECOM*, Austin, TX, USA, Dec. 2014, pp. 1376–1381.

[2] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczók, "The global footprint of mobile communications: The ecological and economic perspective," *IEEE Commun. Mag.*, vol. 49, no. 8, pp. 55–62, Aug. 2011.

[3] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, "Toward dynamic energy-efficient operation of cellular network infrastructure," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 56–61, Jun. 2011.

[4] M. Ismail, W. Zhuang, E. Serpedin, and K. Qaraqe, "A survey on green mobile networking: From the perspectives of network operators and mobile users," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 3, pp. 1535–1556, Aug. 2015.

[5] J. B. Rao and A. O. Fapojuwo, "A survey of energy efficient resource management techniques for multicell cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 154–180, Feb. 2014.

[6] M. Ismail and W. Zhuang, "Network cooperation for energy saving in green radio communications," *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 76–81, Oct. 2011.

[7] Y. Yao, L. Huang, A. Sharma, L. Golubchik, and M. J. Neely, "Data centers power reduction: A two time scale approach for delay tolerant workloads," in *Proc. IEEE INFOCOM*, Orlando, FL, USA, Mar. 2012, pp. 1431–1439.

[8] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Rafael, CA, USA: Morgan & Claypool, 2010.

[9] L. Venturino, A. Zappone, C. Risi, and S. Buzzi, "Energy-efficient scheduling and power allocation in downlink OFDMA networks with base station coordination," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 1–14, Jan. 2015.

[10] C. Xiong, G. Y. Li, S. Zhang, Y. Chen, and S. Xu, "Energy-efficient resource allocation in OFDMA networks," *IEEE Trans. Commun.*, vol. 60, no. 12, pp. 3767–3778, Dec. 2012.

[11] J. Huang, V. G. Subramanian, R. Agrawal, and R. A. Berry, "Downlink scheduling and resource allocation for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 1, pp. 288–296, Jan. 2009.

[12] Z. Shen, J. G. Andrews, and B. L. Evans, "Adaptive resource allocation in multiuser OFDM systems with proportional rate constraints," *IEEE Trans. Wireless Commun.*, vol. 4, no. 6, pp. 2726–2737, Nov. 2005.

[13] G. Auer *et al.*, "How much energy is needed to run a wireless network?" *IEEE Wireless Commun.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[14] M. H. Cheung, R. Southwell, and J. Huang, "Congestion-aware network selection and data offloading," in *Proc. IEEE 48th Annu. Conf. Inf. Sci. Syst. (CISS)*, Princeton, NJ, USA, Mar. 2014, pp. 1–6.

[15] H. Yu, M. H. Cheung, L. Huang, and J. Huang, "Power-delay tradeoff with predictive scheduling in integrated cellular and Wi-Fi networks," *IEEE J. Sel. Areas Commun.*, 2015 [Online]. Available: http://arxiv.org/abs/1512.06428.

[16] L. Huang and M. J. Neely, "Max-weight achieves the exact [O(1/V),O(V)] utility-delay tradeoff under Markov dynamics," arXiv preprint arXiv:1008.0200, 2010

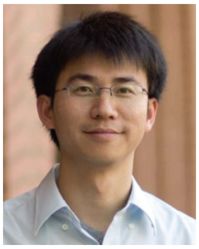
[17] V. Gajić, J. Huang, and B. Rimoldi, "Competition of wireless providers for atomic users," *IEEE/ACM Trans. Netw.*, vol. 22, no. 2, pp. 512–525, Apr. 2014.

[18] G. Bianchi, "Performance analysis of the IEEE 802.11 distributed coordination function," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 3, pp. 535–547, Mar. 2000.

[19] B. H. Jung, H. Jin, and D. K. Sung, "Adaptive transmission power control and rate selection scheme for maximizing energy efficiency of IEEE 802.11 stations," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2012, pp. 266–271.

Haoran Yu (S'14) is currently pursuing the Ph.D. degree at the Department of Information Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong. He is also a Visiting Student at the Yale Institute for Network Science (YINS) and the Department of Electrical Engineering, Yale University, New Haven, CT, USA. His research interests include wireless communications and network economics, with current emphasis on mobile data offloading, cellular/Wi-Fi integration, LTE in unlicensed spectrum, and economics of public Wi-Fi networks.

Man Hon Cheung received the B.Eng. and M.Phil. degrees in information engineering from the Chinese University of Hong Kong (CUHK), Hong Kong, in 2005 and 2007, respectively, and the Ph.D. degree in electrical and computer engineering from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2012. Currently, he is a Postdoctoral Fellow with the Department of Information Engineering, CUHK. His research interests include the design and analysis of wireless network protocols using optimization theory, game theory, and dynamic programming, with current focus on mobile data offloading, mobile crowd sensing, and network economics. He serves as a Technical Program Committee member in IEEE ICC, Globecom, and WCNC. He was the recipient of the IEEE Student Travel Grant for attending IEEE ICC 2009, the Graduate Student International Research Mobility Award by UBC, and the Global Scholarship Programme for Research Excellence by CUHK.



Longbo Huang (S'10–M'11) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, CA, USA, in 2011. He then worked as a Postdoctoral Researcher with the Electrical Engineering and Computer Sciences Department (EECS), University of California at Berkeley (UC Berkeley), Berkeley, CA, USA, from 2011 to 2012. Since 2012, he has been an Assistant Professor with the Institute for Interdisciplinary Information Sciences (IIIS), Tsinghua University, Beijing, China. He was a Visiting Scholar at the LIDS Laboratory, MIT, Cambridge, MA, USA, and at the EECS Department, UC Berkeley. He was also a Visiting Professor at the Chinese University of Hong Kong (CUHK), Hong Kong, and at Bell-labs France. He was selected into China's Youth 1000-talent program in 2013, and into the MSRA StarTrack Program in 2015. His research interests include learning and optimization for networked systems, mobile networks, data center networking, and smart grid. He has served/serves as the TPC Vice Chair for Submissions for WiOpt 2016, and as a TPC member for top-tier IEEE and ACM conferences including ACM Sigmetrics/Performance, MobiHoc, INFOCOM, WiOpt, and E-Energy. His paper in ACM MobiHoc 2014 was selected as a Best Paper Finalist. He was the recipient of the Google Research Award and the Microsoft Research Asia (MSRA) Collaborative Research Award in 2014.

Jianwei Huang (S'01–M'06–SM'11–F'16) received the Ph.D. degree from Northwestern University, in 2005. He is an Associate Professor with the Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong. He is a Distinguished Lecturer of IEEE Communications Society. He was the co-recipient of eight international Best Paper Awards, including IEEE Marconi Prize Paper Award in Wireless Communications in 2011.