Offloading in HCNs: Congestion-Aware Network Selection and User Incentive Design

Yuqing Li, Bingyu Shen, Jinbei Zhang, Xiaoying Gan, *Member, IEEE,* Jingchao Wang, and Xinbing Wang, *Senior Member, IEEE*

***Abstract*—To accommodate exponentially increasing traffic demands, operators are seeking to offload cellular traffic to small Base Stations (BSs) in Heterogeneous Cellular Networks (HCNs), which is promising in alleviating traffic congestion. In HCNs, operators are eager to balance the traffic globally, where users may be pushed to less preferred small BSs, resulting in possible conflict with user local preference. Thus it is a big challenge to achieve dynamic load balancing for operators and provide participation incentive for users simultaneously. Due to the dynamics of network state and user traffic demand, we are inspired to utilize Lyapunov optimization to develop a congestion-aware cellular offloading scheme. Specifically, an operator profit maximization problem involving network selection and rate control is formulated. To achieve long-term network stability, we propose a congestion-aware network selection algorithm, obtaining the BS alternative set that maintains traffic congestion constraint. By exploring the heterogeneity of user quality sensitivity, we devise the optimal quality-price contract which maximizes operator profit. With effective pricing and resource allocation, users are motivated to make proper association strategy chosen from the BS alternative set. Simulation results demonstrate the effectiveness of our scheme in improving operator profit. User incentive and network stability are also validated.**

***Index Terms*—Congestion-aware offloading, Lyapunov optimization, dynamic load balancing, contract-based incentive scheme.**

I. Introduction

In recent years, cellular networks have been facing an explosive growth in mobile data traffic, mainly driven by the proliferation of mobile devices such as smartphones or tablet computers [1]. As predicted by Ericsson's forecast, smartphone traffic is expected to increase by 11-fold between 2015 and 2021 [2]. To meet surging traffic demands, operators have realized that offloading part of cellular traffic from macrocell to other small BSs including picocells, femtocells and WiFi APs, is an effective paradigm to address such traffic overload problem [3], [4].

With the rapid development of cellular offloading service, small BSs are being increasingly widely deployed in HCNs [5], [6]. It is common that users are within the coverage of several small BSs at the same time. Moreover, these BSs may differ much in various aspects, such as transmit power, physical size and operation cost. Such BS heterogeneity makes the offloading quality (e.g., transmission rate) vary from one BS to another. Therefore, how to select a good BS association for each user is an extremely critical issue for cellular offloading.

A typical cellular offloading system usually consists of one single operator and multiple mobile users. In general, who determines user association strategy corresponds to two types of schemes, i.e., operator-initiated offloading [7], [8] and user-initiated offloading [9], [10]. In this paper, we focus on the operator-initiated offloading scheme, where operator is responsible for selecting a specific set of BSs to perform offloading service for users. Comparing with the user-initiated offloading (where users decide which BSs to access), the operator-initiated offloading globally furnishes operator with a better control on how to offload users to which BSs at what cost. Specifically, such scheme involves a combination of network selection, resource allocation and pricing. Ye *et al.* [11] presented a load-aware cell association method for HCNs by considering cell association and resource allocation jointly. Dong *et al.* [12] established reverse auction-based iDEAL scheme to minimize operation cost by leveraging resources from third-party owners.

Even in operator-initiated offloading scheme, due attention should be given to user participation incentive. Generally, users prefer accessing to near small BSs with high offloading quality. In operator's global load balancing, however, users especially for those in congested areas, are always aggressively pushed to less preferred small BSs so as to fully realize the utilization of lightly-loaded small BSs. Inevitably, this practice may result in possible conflict with user local preference, making user satisfaction reduced and reluctant to offload [13]. Thus it is necessary to develop an incentive scheme to promote user incentive and enhance operator profit. Several recent works have been devoted to incentive issues of cellular offloading [14], [15]. With user QoS requirement, Oo *et al.* [16] formulated traffic offloading as a joint optimization problem of interference mitigation, user association and resource allocation. Wang *et al.* [13] established an auction-based framework to transform the global proportional fairness problem into a matching problem by constructing association graph. The system utility is maximized while guaranteeing the optimality of user association. Unfortunately, statistical information about user traffic demand, which is beneficial for making pricing and network selection, is not considered in the above works. Contract theory is effective in designing incentive mechanisms by coordinating the provided service and differential pricing,

Y. Li is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China. Email: liyuqing@sjtu.edu.cn.

B. Shen, J. Zhang, X. Gan, and X. Wang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. Email: {sby2013, abelchina, ganxiaoying, xwang8}@sjtu.edu.cn. J. Wang is with the China Electronic Equipment System Engineering Company, Beijing 100141, China. Email: wangjc_61@163.com.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2724027, IEEE Transactions on Wireless Communications

especially under incomplete information scenario [17], [18]. Zhou *et al*. [19] developed a volume-price contract to study the issue of data offloading. Gao *et al*. [20] designed an optimal contract to cope with spectrum trading process.

In this paper, we propose a contract-based incentive scheme for cellular offloading. With the heterogeneity of user quality sensitivity, we classify users into different types, indicating whether they are quality-prone or price-prone. Operator offers users the optimal contract consisting of a set of offloading quality-price combinations, each specified for one association strategy. By leveraging pricing, users with low quality sensitivity are encouraged to associate with lightly loaded small BSs (which usually provide low-price but low-quality services), rather than those heavily loaded small BSs (which usually provide high-quality but high-price services). Accordingly, such incentive scheme is enabled for users with different quality sensitivity.

Despite the necessity of user incentive design, it's more practical to provide conditional incentive for users. General incentive mechanisms for cellular offloading service are enabled based on user local preference. Aryafar *et al*. [21] modeled BS selection as a non-cooperative game, where each user selects the best BS to maximize its throughput. Inevitably, such local preference may result in severe load imbalance and violate network performance. To achieve global load balancing, traffic congestion constraint should also be taken into account. **First**, random user traffic demand and time-varying channel condition make network dynamics nonnegligible. To avoid heavy congestion induced by surging or plunging traffic demands, there is an urgent need to adjust association strategy dynamically and guarantee network stability. **Second**, due to the differences in BS available resource and load capability, there always exists significant heterogeneity in BS tolerable congestion level (i.e., the maximum number of served users) [22]. Even for serving the same number of users, BSs may reach different congestion levels, e.g., some BSs are heavily loaded while others remain lightly loaded. Failure to make coordinated resource allocation will further aggravate the seriousness of traffic congestion and performance degradation. Li *et al*. [23] proposed the dynamic pricing strategy and resource scheduling policy to address the profit maximization problem. All of these bring the traffic congestion issue to the center of cellular offloading scheme design.

To this end, we are inspired to utilize Lyapunov optimization to exploit the optimality and stability of congestion-aware cellular offloading. Here congestion-aware mainly refers to guaranteeing global network stability (or in temporal domain) and BS tolerable congestion level (or in spatial domain). Accordingly, Lyapunov-based traffic scheduling scheme is proposed, both achieving dynamic load balancing for operators and providing participation incentive for users. Specifically, we first formulate the operator profit maximization (OPM) problem and further convert it into network selection and rate control subproblems. To guarantee long-term network stability and global load balancing, we develop a congestion-aware network selection algorithm, obtaining BS alternative set that maintains traffic congestion constraint. Since effective pricing can activate user willingness to participate, we devise the optimal quality-price contract which maximizes operator profit. Note that our contract is designed based on user conditional incentive. Different from traditional contract design, each contract item is chosen from BS alternative set. In addition, spectrum resource needs to be allocated effectively to enable the optimal quality available to users as much as possible.

Our main contributions are highlighted as follows.

• We propose a Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading. Such scheme involving network selection and rate control is modeled as the OPM problem. To the best of our knowledge, we are the first to consider user conditional incentive issue, i.e., providing user participation incentive under traffic congestion constraint.

• To achieve dynamic load balancing globally, we develop a congestion-aware network selection algorithm, obtaining the BS alternative set that maintains traffic congestion constraint.

• In view of user conditional incentive, we devise the contract-based incentive scheme to maximize operator profit, while guaranteeing traffic congestion constraints. The optimal quality-price contract is obtained under incomplete information scenario, where users are motivated to make proper association strategies chosen from BS alternative sets.

• Simulation results demonstrate the effectiveness of our scheme in improving operator profit. It is indicated that user incentive and network stability can be guaranteed by leveraging effective pricing and spectrum resource allocation.

The remaining part of this paper is organized as follows. We describe system model and problem formulation in Section II and III, respectively. In Section IV, we propose a Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading. Simulation results are given in Section V. Finally, we conclude the paper in Section VI.

II. System model

We give an overview of Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading, including a single operator and $|\mathcal{U}|$ randomly located mobile users in the user set $\mathcal{U}$. In general, operator's global load balancing may result in possible conflict with user local preference. We devise the optimal quality-price contract to promote user participation incentive and enhance operator profit. Here, offloading quality refers to the transmission rate obtained from accessing to BSs, and price represents the corresponding payment for this service. In practice, however, there always exists significant heterogeneity in BS tolerable congestion level, e.g., some BSs are heavily loaded while others remain lightly loaded. Such local preference will further aggravate the seriousness of traffic congestion and performance degradation. Thus it is necessary to provide conditional incentive for users, where users are motivated to make proper association strategies while guaranteeing traffic congestion constraint.

A. Network and Service Model

We consider a HCN consisting of $|\mathcal{B}|$ fixed BSs in the BS set $\mathcal{B} = \{1, 2, \cdots |\mathcal{B}|\}$, where one macrocell denoted by $\{1\}$ and multiple small BSs denoted by $\{2, \cdots |\mathcal{B}|\}$ are involved.

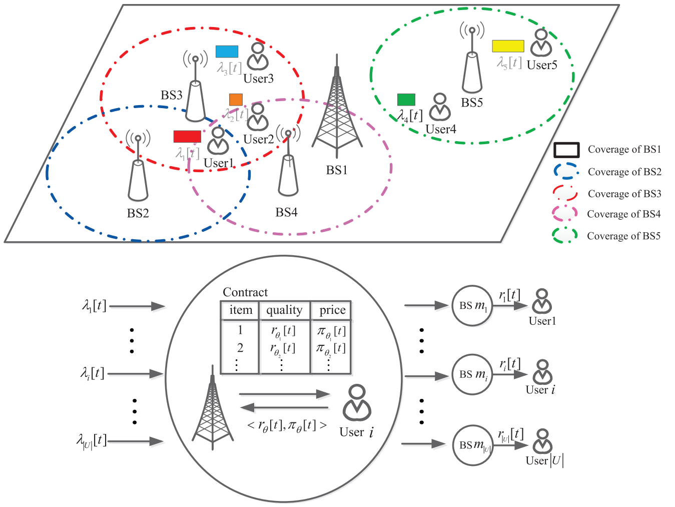


Fig. 1. (a) An example of system with five users (i.e., $\mathcal{U} = \{1, \cdots, 5\}$) and five networks including one macrocell and four WiFi APs (i.e., $\mathcal{B} = \{1, \cdots, 5\}$). The traffic of user $i \in \mathcal{U}$ can be offloaded to BS $m \in \mathcal{B}$, only if user i is within the coverage area of BS m. (b) An illustration of Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading.

With the increasingly widely deployed small BSs, BS coverage area[1] is always overlapping. In this way, users are more likely to be within the coverage of several small BSs at the same time. This offloading system is assumed to operate in a slotted manner with slots indexed by $t \in \mathcal{T} = \{0, 1, \cdots\}$. Let $\mathcal{B}_i[t] \subseteq \mathcal{B}$ denote the set of BSs that are available to user $i \in \mathcal{U}$ at time slot t, where all users are assumed to be in the coverage area of macrocell[2]. Note that in practice, it would be likely that $\mathcal{B}_i[t] = \{1\}$, indicating no small BS is available to user i. Then user i will access to the macrocell only if its participation incentive is guaranteed, otherwise it will give up the offloading service. In such cellular offloading system, users arrive at any time slot and their traffic demands can be satisfied at one particular slot. Denote the traffic demand of user i at slot t as $\lambda_i[t]$. The traffic of user $i \in \mathcal{U}$ can be offloaded to BS $m \in \mathcal{B}$, only if user i is within the coverage area of BS m.

Figure 1 illustrates an example of offloading system with one macrocell, four WiFi APs, and five users. In general, macrocell, i.e., BS 1, is more expensive than other small BSs. Intuitively, the traffic of User 1 can be offloaded to BS 1, BS 2, BS 3, and BS 4. To achieve high offloading quality under relatively low price, User 1 would prefer associating with BS 3 due to short transmission distance. In reality, however, BS 3 is rather congested compared with BS 2 and BS 4. Thus the desired choice of User 1 may further exacerbate the severity of load imbalance, which is not expected for operator to see.

We propose a Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading, achieving a good balance between user preference and load balancing. In view of user conditional incentive, we devise an optimal quality-price contract, where each user are encouraged to make association strategy that maintains traffic congestion constraint. According to user quality sensitivity, we classify users into different types denoted as θ. The optimal contract contains a set of quality-price contract items, each intended for a specific user type. In Fig. 1, operator offers the optimal contract to users. By comparing BS alternatives, users will choose a proper BS to maximize its utility. By leveraging pricing and resource allocation, users with low quality sensitivity (e.g., user 1) are encouraged to associate with lightly loaded small BSs which usually provide low-price but low-quality services (e.g., BS 2), rather than those heavily loaded small BSs which usually provide high-quality but high-price services (e.g., BS 3). Accordingly, such incentive scheme is enabled for users with different quality sensitivity.

B. BS Modeling

We consider a comparatively crowded scenario in this paper. To balance the traffic globally, effective resource allocation and pricing need to be performed carefully. On the one hand, **spectrum resource** has a great impact on offloading quality. If more spectrum resource is allocated, the offloading quality that BS can provide will increase. Thus it is essential to allocate resource reasonably to improve spectrum utilization and offloading efficiency. On the other hand, **pricing** is one of common ways to effectively foster user willingness to participate. Generally, high price always pushes users to other BSs, while low price may attract more users to associate.

1) *Offloading Quality:*

Suppose each BS transmits with a constant transmission power no matter which users are associated. For any BS $m \in \mathcal{B}_i[t]$, the transmission power is denoted as $P^t_{im}[t]$. In HCNs, one typical large scale fading, i.e., a log-distance path loss model, is considered[3]. In previous studies, the log-distance path loss model has been extensively studied since it can nicely characterize the result of signal attenuation caused by signal propagation over large distances [13], [16]. Accordingly, the signal power received at user i can be characterized as

$$P^r_{im}[t] = P^t_{im}[t](1 + d_{im}[t])^{-\gamma}, \tag{1}$$

where γ is pathloss exponent, and $d_{im}[t]$ is the distance between user i and BS m. The element $1 + d_{im}[t]$ is to ensure $P^r_{im}[t] \leq P^t_{im}[t]$. Let $P^{noise}_i[t]$ denote the received noise power, and $P^r_{in}[t]$ denote the received interference power from other small BSs. The Signal-to-Interference-plus-Noise-Ratio (SINR) received at user i from BS m is

$$\Gamma_{im}[t] = \frac{P^r_{im}[t]}{\sum\limits_{n \in \mathcal{B}_i[t] \setminus \{1,m\}} P^r_{in}[t] + P^{noise}_i[t]}. \tag{2}$$

Note that the term $n \in \mathcal{B}_i[t] \setminus \{1, m\}$ demonstrates BS m is only interfered by other small BSs, where macrocell is not included. There always exists heterogeneity in BS transmission power. The typical transmission power of a macrocell is

[1] BS coverage area is always determined by the received SINR, which will be discussed in detail later. Inspired by [24], we model the BS coverage area as a circle area centered at the BS.

[2] In traditional two-tier offloading structure, the macrocell coverage area is often partitioned according to Voronoi Cell, which is not overlapping with each other. Then each user can only access to one macrocell at any time [25]. Actually, the focus of our study is a typical two-tier offloading structure with only one macrocell, and that is why the macrocell is available to all users.

[3] Note that the small scale fading is not involved in our work. Under the influence of small scale fading, BS coverage area may exhibit in the shape of irregular circle, which will be left for future study.

around 43 dBm in practice, and that of small BSs is usually $20 \sim 30$ dBm [26]. In this case, high-power macrocell is very likely to interfere with low-power small BSs. To eliminate the induced mutual interference between macrocell and small BSs, we consider the orthogonal-channel deployment, where macrocell operates at different spectrum from small BSs. Since there is no great difference in transmission power of small BSs, it is acceptable to regard these small BSs use the same frequency band and interfere with each other. Actually, it is quite common to adopt such deployment in HCNs [16], [27].

With current channel condition, operator needs to make coordinated spectrum resource allocation. Obviously, inefficient allocation will increase the risk of network congestion and reduce the offloading efficiency. Let $s_m[t]$ denote the resource allocated to BS m at slot t. The spectrum resource allocation for this system can be given by vector $\boldsymbol{s_m}[t] = (s_1[t], \cdots, s_{|\mathcal{B}|}[t])$.

In this work, offloading quality is characterized by the achievable transmission rate. According to Shannon Theory, the offloading quality that user i obtains from BS m is

$$r_{im}[t] = s_m[t] \cdot log_2(1 + \Gamma_{im}[t]). \tag{3}$$

It is obvious the more spectrum resource allocated in the better channel condition, the higher offloading quality will obtain.

Remark. Given BS transmission power and spectrum resource allocation, offloading quality greatly depends on the SINR received at users, which actually involves channel fading. The further the distance is, the larger the attenuation will be, indicating smaller SINR will be received. The successful transmission can be realized only if the SINR is not less than the predetermined threshold. Thus we consider BS coverage area as a circle centered at the BS. In general, high-quality services are desired by rational users, and that is why users always prefer accessing to near BSs. Since user local preference may result in severe load imbalance, this paper is targeted at achieving a good balance between operator's load balancing and user local preference.

2) Association and Congestion Constraints:

Different from cellular network, small BSs such as WiFi APs, may not be available to users at all time due to user mobility. They can only provide intermittent network connectivity [28]. User i can potentially access to any BS m only if it wanders into the coverage of BS m. Under the definition of $\mathcal{B}_i[t]$, we can easily obtain that for any BS $m \in \mathcal{B}_i[t]$, it is available to user i or user i is in its coverage area. At each time slot, users need to choose one BS to access from the BS available set. On the basis of this, we introduce a binary decision variable $x_{im}[t]$ to characterize BS association strategy. Specifically, $x_{im}[t] = 1$ represents that BS m is selected to serve user i, and $x_{im}[t] = 0$ otherwise. The BS association strategy of user i can be further given by the vector $\boldsymbol{x_i}[t] = (x_{i1}[t], \cdots, x_{i|\mathcal{B}|}[t])$. In practice, user i can only associate with at most one BS at any time slot. Then we can obtain the BS association constraint, i.e.,

$$\sum_{m \in \mathcal{B}} x_{im}[t] = \sum_{m \in \mathcal{B}_i[t]} x_{im}[t] \leq 1, \quad \forall i \in \mathcal{U}. \tag{4}$$

With BS association strategy, the offloading quality received at user i can be given by

$$r_i[t] = \sum_{m \in \mathcal{B}} x_{im}[t] \cdot r_{im}[t]. \tag{5}$$

Accordingly, we further obtain the total offloading quality of this system, i.e.,

$$r[t] = \sum_{i \in \mathcal{U}} r_i[t]. \tag{6}$$

In congestion-aware cellular offloading, each BS is assumed to have limited offloading capability, especially for resource-constraint scenarios. Suppose the average number of users served by BS m does not exceed a tolerable congestion level α_m. Here $\alpha_m \geq \max \sum_{i \in \mathcal{U}} x_{im}[t]$ is assumed to be BS specific, since different BSs may have various available resources and load efficiency. Define the time-average number of users served by BS m as $\overline{x}_m = \underset{t \to \infty}{\text{limsup}} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left\{\sum_{i \in \mathcal{U}} x_{im}[\tau]\right\}$. The BS congestion constraint can be represented as

$$\overline{x}_m \leq \alpha_m, \forall m \in \mathcal{B}. \tag{7}$$

C. Mobile User Modeling

1) User Attribute:

Intuitively, high offloading quality and low price services are preferred by users. In practice, however, high quality is always accompanied by high price. Hence there exists a tradeoff between users' desired quality and price. To quantify the potential trend of users' desired association strategy, we introduce user quality sensitivity denoted as θ_i, to characterize whether they are quality-prone or price-prone. In this paper, we regard θ_i as user-dependent and it is specified for each user. In addition, due attention should be given to user traffic demand since surging traffic demand always implies high risk of traffic congestion. Moreover, the BS available set is closely related to user location. Based on the above, we can describe user attributes as follows.

Definition 1: (User Attributes) Each user $i \in \mathcal{U}$ is associated with:

- A quality sensitivity θ_i captures the heterogeneity of user preference for offloading quality.
- A user requesting rate $\lambda_i[t]$ characterizes user traffic demand, denoting the amount of traffic (measured in units of packets arrived at user i at time slot t.
- A BS available set $B_i[t]$ describes the set of BSs available to user i at time slot t.

2) User Association Strategy:

Once BS available set $B_i[t]$ is given, user association sytategy is determined by offloading quality and price. Denote the corresponding price paid to operator for offloading quality $r_i[t]$ as $\pi_i[t]$. Accordingly, each association strategy can be specified by a quality-price tuple $\langle r_i[t], \pi_i[t] \rangle$. In user conditional incentive issue, users are motivated to make proper association strategy, which involves user utility maximization under traffic congestion constraint. On the basis of the heterogeneity of user quality sensitivity, it is promising to enable the conditional incentive scheme by making differentiated pricing. Specifically, users can be guided to associate with lightly loaded BSs usually providing low-price but low-quality services, rather

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2724027, IEEE Transactions on Wireless Communications

than those heavily loaded BSs usually providing high-quality but high-price services.

3) User Satisfaction Function:

We introduce user satisfaction function to characterize how users are happy with offloading service. In this paper, the offloading quality mainly refers to the received transmission rate. Due to the difference in SINR and allocated spectrum resource, the obtained quality varies as well. Based on user attributes, we define quality sensitivity θ_i as the increment in user satisfaction for a unit increase of offloading quality. The larger θ_i is, the higher user i' requirement on offloading quality would be. Intuitively, high-quality services are preferred by users, which can make user satisfaction increase. We assume user satisfaction is proportional to offloading quality. As shown in [29], logarithm utility functions lead to proportional fairness among users. Thus we define user satisfaction function as the monotone increasing function of offloading quality, i.e.,

$$V_i[t] = \ln\left(1 + \theta_i r_i[t]\right). \tag{8}$$

D. Data Queue

It is necessary for operator to develop traffic scheduling scheme so as to keep the whole network stable. In particular, we model the dynamic traffic demand and being served of all users as a single data queue. At any slot t, user requesting rate in the whole system can be given by vector $\boldsymbol{\lambda}[t] = \left(\lambda_1[t], \cdots, \lambda_{|\mathcal{U}|}[t]\right)$. We regard the requesting rate of each user subscribing to offloading service as an individual traffic arrival. Then the traffic arrival of data queue can be characterized as $A[t] = \sum_{i\in\mathcal{U}} \lambda_i[t]$, which is i. i. d. over slots.

1) Queueing Dynamic:

The queueing dynamics play a key role in characterizing time-varying channel condition and control action. As for such data queue, we denote queue backlog as $Q[t]$, describing the unserved traffic waiting for being offloaded at the beginning of slot t. The data queue is assumed to be initially empty, i.e., $Q[0] = 0$. Accordingly, the queueing dynamic can be illustrated as

$$Q[t+1] = \max\left\{Q[t] - r[t], 0\right\} + A[t], \forall t \geq 0, \tag{9}$$

where $r[t]$ and $A[t]$ are the corresponding transmission rate and traffic arrival, respectively. Moreover, we suppose any traffic arrival occurs at the end of each slot, indicating that the packet cannot be served during that slot. The term $\max\{\cdot\}$ guarantees that the amount of served packets is no more than current queue backlog size.

2) Network Stability:

In view of time-varying characteristics of traffic arrival and transmission rate, we define strong stability to handle these two arbitrary stochastic processes [30].

Definition 2: A queue is called strongly stable if it has a bounded time average backlog, i.e.,

$$\overline{Q} = \limsup_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{Q[\tau]\} < \infty. \tag{10}$$

III. Problem Formulation

A. Conditional Incentive Problem

As for this congestion-aware cellular offloading, it is crucial to maintain traffic congestion constraint while providing user participation incentive, i.e., user incentive is conditionally guaranteed. Here the traffic congestion constraint mainly involves network stability (or in temporal domain) and load-balancing globally (or in spatial domain). As two rational entities in offloading service, both users and operator are attempting to pursue their own benefits. Generally, users prefer accessing to those BSs with high offloading quality. But in operator's load balancing, users are always pushed to less preferred small BSs. Hence, the key to solving this conditional incentive issue is to resolve the conflict between users' local preference and operator's global traffic scheduling. By exploring the heterogeneity of user price sensitivity, it is promising to encourage users to associate with lightly-loaded small BSs by conducting effective pricing and spectrum resource allocation. In the following, we will formally characterize this problem.

When accessing to BS m, user i should pay price $\pi_i[t]$ to operator for the obtained offloading quality $r_i[t]$. We define the service valuation perceived by user i as user satisfaction function. Thus the utility function of user i can be modeled as

$$U_i[t] = wV_i[t] - \pi_i[t], \tag{11}$$

where $w > 0$ is the scaling weight between service valuation and payment. Without loss of generality, we suppose $w = 1$. Substituting equation (8) into (11) yields

$$U_i[t] = \ln\left(1 + \theta_i r_i[t]\right) - \pi_i[t]. \tag{12}$$

From users' perspective, participation incentive is guaranteed only if user association strategy can satisfy the following two constraints, which are commonly used in incentive mechanism designs [17], [18], [19], [20].

Definition 4: (IR: Individual Rationality) User association strategy satisfies the IR constraint if at any time slot t, each user receives a non-negative utility, i.e.,

$$U_i[t] \geq 0. \tag{13}$$

Definition 5: (IC: Incentive Compatibility) The IC constraint is satisfied if at any time slot t, the selected association strategy is optimal to user i to other strategies, i.e.,

$$\ln\left(1 + \theta_i r_i[t]\right) - \pi_i[t] \geq \ln\left(1 + \theta_i {r'}_i[t]\right) - {\pi'}_i[t] \tag{14}$$

where $\langle {r'}_i[t], {\pi'}_i[t]\rangle$ is the offloading service obtained from user i' accessing to other small BSs.

Remark. Actually, IR constraint demonstrates any rational user would avoid a network that results in negative utility. IC constraint guarantees the optimality of user association strategy, indicating that each user can only associate with at most one BS at each slot. Thus the BS association constraint in (4) is actually embedded in IC constraint.

Inevitably, providing offloading service will incur operation cost. The operation cost is assumed as quality-specific cost, mainly including transmission cost through BS networks. Specifically, we regard such operation cost as a monotone increasing function of offloading quality, which is explicit and commonly used [31]. Thus the operation cost for serving user i can be modeled as

$$g_i[t] = r_i[t]c_i[t], \tag{15}$$

where $c_i[t]$ is the unit operation cost incurred by user i when accessing to BSs. We define operator profit, denoted by $R[t]$, as the difference between revenue and operation cost, i.e.,

$$R[t] = \sum_{i\in\mathcal{U}} \pi_i[t] - r_i[t]c_i[t]. \tag{16}$$

We obtain the expected time average operator profit, i.e.,

$$\overline{R} = \liminf_{t\to\infty} \frac{1}{t}\sum_{\tau=0}^{t-1} \mathbb{E}\{R[\tau]\}. \tag{17}$$

Thus the operator profit maximization problem (OPM) can be formulated as

OPM1:

$$\begin{aligned} \max \quad & \overline{R} && (18)\\ \text{s.t.} \quad & \overline{Q} < \infty, && (19)\\ & \overline{x}_m \le \alpha_m, \forall m \in \mathcal{B}, && (20)\\ & \text{IR constraint in (13) \& IC constraint in (14).} && (21) \end{aligned}$$

The OPM problem involves network selection and user incentive design subproblems. Specifically, conditions (19) and (20) represent network stability and BS congestion level constraints, respectively. User participation incentive is guaranteed in condition (21).

B. Our Solution

The OPM problem is challenging to address due to the following reasons. **First**, it is an infinite horizon stochastic optimization problem, making it hard to ensure that all the above conditions are satisfied for any user at any time slot. **Second**, conditions (13) and (14) guarantee for each user i, the selected scheduling $\langle r_i[t], \pi_i[t]\rangle$ is superior to other scheduling $\langle r'_i[t], \pi'_i[t]\rangle$ when accessing to other BSs. Such conflicting optimization further increases the difficulty of OPM.

To this end, we propose Lyapunov-based dynamic traffic scheduling scheme to exploit the optimality and stability of congestion-aware cellular offloading. Under this scheme, the OPM problem is further converted into network selection and rate control subproblems. In user conditional incentive issue, the offloading vector in the second one is actually determined by association vector in the first one. Thus one feasible solution is to address the incentive issue with condition (21) on the basis of traffic congestion constraint obtained from network selection problem with conditions (19) and (20). Specifically, we first develop congestion-aware network selection algorithm, and obtain the BS alternative set that maintains traffic congestion constraint. To effectively activate user willingness to participate, the optimal quality-price contract is carefully designed, where each user can choose the best one from the BS alternative set.

IV. Lyapunov-based dynamic traffic scheduling scheme for Congestion-Aware Cellular Offloading

A. Lyapunov Optimization

Lyapunov optimization is efficient in designing stable control algorithms. It has been extended to treat network stability and performance optimization simultaneously [30], [32], [33]. Considering the dynamics of network state and user traffic demand, we are inspired to implement Lyapunov optimization to study this congestion-aware cellular offloading.

1) Virtual Queues:

We introduce virtual queue $\boldsymbol{X}_m[t]$ for BS network $m \in \mathcal{B}$, with dynamic update equation as

$$X_m[t+1] = \max\{X_m[t] - \alpha_m, 0\} + x_m[t]. \tag{22}$$

In fact, $X_m[t]$ can be viewed as the queue backlog in the virtual queue of BS m with arrival rate $x_m[t]$ and serving rate α_m. Without loss of generality, we assume that all virtual queues are initially empty, i.e., $X_m[0] = 0, \forall m \in \mathcal{B}$.

According to the definition 2, if our scheme could stabilize all virtual queues, BS congestion level constraints will be satisfied. Therefore, the constraints in (19) and (20) are equivalent to stabilizing all actual and virtual queues in the network.

2) Drift-plus-Penalty Function:

Let $\boldsymbol{\Theta}[t] = [\boldsymbol{Q}[t], \boldsymbol{X}[t]]$ as the aggregate queue vector. We define the Lyapunov function as

$$L(\boldsymbol{\Theta}[t]) = \frac{1}{2}\left[Q^2[t] + \sum_{m\in\mathcal{B}} X_m^2[t]\right], \tag{23}$$

where the factor $1/2$ is used for notational convenience. The Lyapunov drift can be defined as

$$\Delta(\boldsymbol{\Theta}[t]) = \mathbb{E}\{L(\boldsymbol{\Theta}[t+1]) - L(\boldsymbol{\Theta}[t]) | \boldsymbol{\Theta}[t]\}. \tag{24}$$

It characterizes the expected change in the quadratic function of queue backlog over each slot.

We incorporate operator profit into Lyapunov drift, providing network stability and profit maximization (or penalty minimization) jointly. At every time slot, we try to minimize the drift-plus-penalty problem greedily, i.e.,

$$\min \;\; \Delta(\boldsymbol{\Theta}[t]) - V\mathbb{E}\{R[t]|\boldsymbol{\Theta}[t]\}, \tag{25}$$

where $-\mathbb{E}\{R[t]|\boldsymbol{\Theta}[t]\}$ is penalty, and V is a control parameter to deal with the tradeoff between operator profit and delay.

Remark. In (25), the first term $\Delta(\boldsymbol{\Theta}[t])$ shows that the constraints in (19) and (20), i.e., network stability and BS congestion level constraints, will be satisfied if there exists an upper bound of the drift-plus-penalty function. The second term $-\mathbb{E}\{R[t]|\boldsymbol{\Theta}[t]\}$ indicates that operator profit maximization is also involved in the minimization problem in (25).

Based on the above analysis, the OPM problem can be further rewritten as

$$\begin{aligned} &\text{OPM2:}\\ &\min \quad \Delta(\boldsymbol{\Theta}[t]) - V\mathbb{E}\{R[t]|\boldsymbol{\Theta}[t]\}\\ &\;\text{s.t.} \quad \text{IR constraint in (13) \& IC constraint in (14)}\;. \end{aligned} \tag{26}$$

3) Performance Analysis:

The key to solve the drift-plus-penalty minimization problem in (25) lies in finding an upper bound of drift-plus-penalty function, which is defined in the following lemma.

Lemma 1: For all slots t, we have

$$\begin{aligned}\Delta(\mathbf{\Theta}[t]) - V\mathbb{E}\{R[t]|\mathbf{\Theta}[t]\} \le B - V\mathbb{E}\{R[t]|\mathbf{\Theta}[t]\}\\ +\mathbb{E}\left\{Q[t]\left(A[t]-r[t]\right)|\mathbf{\Theta}[t]\right\}\\ +\mathbb{E}\left\{\sum_{m\in\mathcal{B}} X_m[t]\left(x_m[t]-\alpha_m\right)|\mathbf{\Theta}[t]\right\},\end{aligned} \tag{27}$$

where B is a finite and positive constant satisfying the following condition for all t, i.e., $B \ge \frac{1}{2}E\left\{r^2[t]+A^2[t]+\sum\limits_{m\in B} x_m^2[t]+\alpha_m^2|\Theta[t]\right\}$.

Proof. See Appendix A. □

Remark. According to Lemma 1, the drift-plus-penalty minimization problem in (25) is equivalent to minimizing the Right-Hand-Side (RHS) of (27).

Theorem 1: (Lyapunov Optimization) Let $\mathbb{E}\{L(\mathbf{\Theta}(0))\} < \infty$ and define R^* as our desired "target" profit. Suppose there exist finite and positive constants V, ε, B such that for every time slot t and aggregate queue backlog vector $\mathbf{\Theta}[t] = [\boldsymbol{Q}[t], \boldsymbol{X}[t]]$, the Lyapunov drift satisfies

$$\Delta(\mathbf{\Theta}[t]) - V\mathbb{E}\{R[t]|\mathbf{\Theta}[t]\} \le B - \varepsilon\left(Q[t] + \sum_{m\in\mathcal{B}} X_m[t]\right) - VR^* \tag{28}$$

then we have

$$\begin{aligned}\overline{Q}+\overline{X} = \limsup_{t\to\infty}\tfrac{1}{t}\textstyle\sum_{\tau=0}^{t-1}\left(\mathbb{E}\{Q[\tau]\} + \sum\limits_{m\in\mathcal{B}}\mathbb{E}\{X_m[\tau]\}\right)\\ \le \tfrac{B+V(\overline{R}-R^*)}{\varepsilon},\end{aligned} \tag{29}$$

$$\overline{R} = \liminf_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\{R(\tau)\} \ge R^* - \frac{B}{V}, \tag{30}$$

where $\overline{Q}$ and $\overline{X}$ are the corresponding average queue backlogs.

Proof. See Appendix B. □

Remark. Theorem 1 provides an upper bound of average queue backlog and a lower bound of average profit, respectively. According to Little's law, queue backlog is proportional to delay. Here the delay refers not only to actual delay in data queue, but also to BS tolerable congestion level reflected in virtual queues. Theorem 1 specifies a drift condition which ensures operator profit can be pushed arbitrarily close to target profit R^* by increasing control parameter V. From equation (29), we obtain the increase of V leads to longer delay since queue backlog is linear in V. Thus there exists a tradeoff between operator profit and delay performance.

4) Optimizing the Drift Bound:

We attempt to minimize the RHS of (27) at every time slot. Taking expectations of (27) with respect to the distribution of $\mathbf{\Theta}[t]$ and using the law of iterated expectations yield

$$\begin{aligned}&\Delta(\mathbf{\Theta}[t]) - V\mathbb{E}\{R[t]\} \le B\\ &\quad -V\mathbb{E}\left\{\sum_{i\in\mathcal{U}}\left(\pi_i[t]-c_i[t]r_i[t]\right)\right\}\\ &\quad +\mathbb{E}\left\{\sum_{i\in\mathcal{U}}\sum_{m\in\mathcal{B}} x_{im}[t]Q[t]\left(\lambda_i[t]-\tfrac{r_{im}[t]}{V}\right)\right\}\\ &\quad -\mathbb{E}\left\{\sum_{i\in\mathcal{U}}\sum_{m\in\mathcal{B}} x_{im}[t]X_m[t]\left(\frac{\alpha_m}{\sum\limits_{l\in\mathcal{U}\setminus\{i\}} x_{lm}[t]}-1\right)\right\}.\end{aligned} \tag{31}$$

Minimizing the drift-plus-penalty is equivalent to minimizing the RHS of (31). The minimization problem in RHS of (31) can be converted into network selection (i.e., the third term of the RHS of (31)) and rate control (i.e., the second term of the RHS of (31)) subproblems. Specifically, the former determines the BS alternative set that maintains traffic congestion constraint, and the latter involves how to select the best one for each user from its BS alternative set. In user conditional incentive issue, offloading quality is dependent on traffic congestion constraint, and each quality can correspond to one particular price by making effective pricing. The decision vector in the second one is determined by association vector in the first one. Therefore, one feasible solution is first to cope with network selection problem, and then on that basis, to address rate control problem. Such practice can maintain the equivalence to the OPM problem. In the following, we present the details of these two subproblems.

- ***Network Selection:***

We introduce association vector $\overline{x}_{im}[t]$ to characterize the BS alternative set that maintains traffic congestion constraint. Specifically, $\overline{x}_{im}[t] = 1$ indicates BS m is in the BS alternative set of user i, and $\overline{x}_{im}[t] = 0$ otherwise. Note that $\overline{x}_{im}[t]$ is not the final BS association strategy $x_{im}[t]$, which is determined by rate control problem. For each user, there is always more than one potential BS satisfying $\overline{x}_{im}[t] = 1$.

We determine association vector $\overline{x}_{im}[t]$ by observing queue backlogs $Q[t]$ and $X_m[t]$.

$$\begin{aligned}&\mathrm{OPM2-1}:\\ &\min\sum_{i\in U}\sum_{m\in B}\bar{x}_{im}[t]\left\{Q[t]\left(\lambda_i[t]-\tfrac{r_{im}[t]}{V}\right)-X_m[t]\left(\tfrac{\alpha_m}{\bar{x}'_{im}[t]}-1\right)\right\}\end{aligned} \tag{32}$$

where $\overline{x}'_{im}[t] = \sum_{l\in\mathcal{U}\setminus\{i\}}\overline{x}_{im}[t]$. This network selection subproblem can be solved in a distributed manner. If the objective function in (32) is negative, set $\overline{x}_{im}[t] = 1$ and regard network m as one BS alternative offered to user i. Otherwise, set $\overline{x}_{im}[t] = 0$. The solution to this problem can be viewed as BS alternative set that maintains traffic congestion constraint. After that, we can determine the offloading quality specified for each of BS alternative set, i.e., $r_i[t] \in \Psi = \{r_{im}[t]|\overline{x}_{im}[t] = 1\}$. We present congestion-aware network selection algorithm in Algorithm 1.

- ***Rate Control:***

To guarantee user conditional incentive, user association strategy needs to be chosen from the BS alternative set obtained from Algorithm 1. That is, the offloading vector $\langle r_i[t], \pi_i[t]\rangle$ is determined by association vector $\overline{x}_{im}[t]$. Thus

Algorithm 1 Congestion-Aware Network Selection Algorithm

```
Initialization: $Q[0] = 0$, $X_m[0] = 0, \forall m \in \mathcal{B}$;
                $\overline{x}_{im}[t] = 0, \forall i \in \mathcal{U}, m \in \mathcal{B}$;
for each time slot $t \in \{0, 1, 2, \cdots\}$ do
    Queue-Backlog Effect Minimization:
    Set $Z_{im}[t] = Q[t]\lambda_i[t] - \frac{Q[t]r_{im}[t]}{V} - X_m[t]\left(\frac{\alpha_m}{\overline{x}'_{im}[t]} - 1\right)$
    while $Z_{im}[t] < 0$ do
        $\overline{x}_{im}[t] = 1$
    end while
    Updating Rule:
        $Q[t+1] = \max\{Q[t] - r[t], 0\} + A[t]$
        $X_m[t+1] = \max\{X_m[t] - \alpha_m, 0\} + x_m[t]$,
end for
```

the rate control problem can be characterized as

$$\begin{aligned} &\text{OPM2-1:} \\ &\max \quad \sum_{i \in \mathcal{U}} \pi_i[t] - r_i[t]c_i[t] \\ &\text{s.t.} \quad \text{IR constraint in (13) \& IC constraint in (14)} \\ &\qquad\qquad r_i[t] \in \Psi. \end{aligned} \tag{33}$$

As both sides of this transaction, operator and users have different objective functions, which are in conflict. We are inspired to utilize contract theory to pursue a mutually beneficial resolution of disputes. By leveraging pricing and resource allocation, the designed contract can realize operator profit maximization, where user incentive is guaranteed under traffic congestion constraints. Different from traditional contract design, all contract items in our work are chosen from BS alternative set obtained from Algorithm 1. Note that many complex interactions among operator and users are involved in this issue, making it not feasible to implement such incentive scheme at each time slot, especially for rapidly changing network states. Thus we conduct the contract-based incentive scheme on the basis of long-time network stability. In particular, we will further study the optimal quality-price contract design in Section IV-B.

B. Quality-Price Contract Design

1) Contract Formulation Under Incomplete Information Scenario:

In general, users may have different preferences for offloading service. Quality sensitivity is introduced to capture whether the user is quality-prone or price-prone. For users with high quality sensitivity such as staff in stock market, offloading quality is the major concern in making association strategy. While for users with low quality sensitivity such as students in college, what they care more about is the price paid to operator. We classify users into different types according to quality sensitivity, and refer to user i as a type-θ user if its quality sensitivity $\theta_i = \theta$.

Suppose there are enough users in HCNs, and then it is acceptable to consider user type θ continuous. In this work, we investigate contract-based incentive scheme under incomplete information scenario. That is, user type is private information only known to itself. Operator, however, cannot know the exact type of particular user, and is only aware of the distribution of user type, which is determined by the probability mass function $f(\theta)$ on an interval $[\theta_l, \theta_u]$.

The quality-price contract is composed of a set of quality-price combinations, where each combination corresponds to one BS association strategy. Once given user traffic demand and location, the selected association strategy mainly depends on user type. In such contract design, only one particular strategy will be chosen by each type of users. Then quality $r_i[t]$ and price $\pi_i[t]$ can be written as $r_\theta[t]$ and $\pi_\theta[t]$, respectively. We denote Ω as the set of all possible qualities and Π as the set of all possible prices, where each quality $r_i[t] \in \Omega$ corresponds to only one price $\pi_i[t] \in \Pi$. The association strategy $\langle r_i[t], \pi_i[t] \rangle$ for type-θ user i can be characterized as quality-price contract item $\langle r_\theta[t], \pi_\theta[t] \rangle$, indicating that the specified association strategy is identical for all type-θ users. Thus user utility can be further represented as

$$U_\theta[t] = \ln(1 + \theta r_\theta[t]) - \pi_\theta[t]. \tag{34}$$

With the statistical information about user quality sensitivity, operator profit can be given by

$$R[t] = \int_{\theta_l}^{\theta_u} (\pi_\theta[t] - r_\theta[t]c_\theta[t]) \cdot f(\theta) d\theta. \tag{35}$$

Accordingly, the contract optimization (CO) problem can be formulated as

CO-1:

$$\max_{\{\langle r_\theta[t], \pi_\theta[t] \rangle, \forall \theta \in [\theta_l, \theta_u]\}} \int_{\theta_l}^{\theta_u} (\pi_\theta[t] - r_\theta[t]c_\theta[t]) \cdot f(\theta) d\theta \tag{36}$$

$$\text{s.t.} \qquad \ln(1 + \theta r_\theta[t]) - \pi_\theta[t] \geq 0, \tag{37}$$

$$\ln(1 + \theta r_\theta[t]) - \pi_\theta[t] \geq \ln\left(1 + \theta r_{\hat{\theta}}[t]\right) - \pi_{\hat{\theta}}[t] \tag{38}$$

$$r_\theta[t] \in \Psi. \tag{39}$$

where constraints (37) and (38) are actually user IR and IC constraints in the context of contract theory. BS alternative condition (39) indicates offloading quality needs to satisfy traffic congestion constraint. Before solving CO problem, we first simplify the IR and IC constraints [10], [19].

Lemma 2: As for the optimal contract under strongly incomplete information scenario in CO problem, the IR constraint can be replaced by

$$\ln(1 + \theta_l r_{\theta_l}[t]) - \pi_{\theta_l}[t] \geq 0, \tag{40}$$

given that the IC constraint holds.

Proof. See Appendix C. □

Lemma 3: If user utility function satisfies Spence-Mirrlees Condition (SMC), i.e., $\frac{\partial}{\partial\theta}\left[-\frac{\partial U/\partial r}{\partial U/\partial \pi}\right] > 0$, the IC constraint in CO problem is equivalent to the following two conditions:

Monotonicity:

$$\frac{dr_\theta[t]}{d\theta} \geq 0, \tag{41}$$

Local incentive compatibility:

$$\frac{\theta r'_\theta[t]}{1 + \theta r_\theta[t]} = \pi'_\theta[t]. \tag{42}$$

Proof. See Appendix D. □

2) Optimality of Contract:

According to Lemma 2 and Lemma 3, the CO problem can be rewritten as

$$\begin{array}{ll} \text{CO-2:} & \\ \max\limits_{\{\langle r_\theta[t],\pi_\theta[t]\rangle,\forall\theta\in[\theta_l,\theta_u]\}} & \int_{\theta_l}^{\theta_u}\left(\pi_\theta[t]-r_\theta[t]c_\theta[t]\right)\cdot f(\theta)d\theta \\ \text{s.t.} & \ln\left(1+\theta_l r_{\theta_l}[t]\right)-\pi_{\theta_l}[t]\geq 0, \\ & \frac{dr_\theta[t]}{d\theta}\geq 0, \\ & \frac{\theta r'_\theta[t]}{1+\theta r_\theta[t]}=\pi'_\theta[t], \\ & r_\theta[t]\in\Psi. \end{array} \tag{43}$$

To address the problem in (43), we first solve the relaxed CO problem without monotonicity condition and BS alternative condition. After that, check whether the obtained solutions satisfy these two conditions.

Define

$$\begin{aligned} W_\theta[t] &= \ln\left(1+\theta r_\theta[t]\right)-\pi_\theta[t] \\ &= \max_{\hat\theta}\left(\ln\left(1+\theta r_{\hat\theta}[t]\right)-\pi_{\hat\theta}[t]\right). \end{aligned} \tag{44}$$

According to the envelope theorem [17], we have $\frac{dW_\theta[t]}{d\theta}=\frac{\partial W_\theta[t]}{\partial\theta}|_{\hat\theta=\theta}=\frac{r_\theta[t]}{1+\theta r_\theta[t]}$. At the optimal contract, the IR constraint of the lowest type is binding, i.e., $W_{\theta_l}[t]=0$. Integrating both sides of this equation yields $W_\theta[t]=\int_{\theta_l}^{\theta}\frac{r_x[t]}{1+xr_x[t]}dx$. By equation (44), we have $\pi_\theta[t]=\ln(1+\theta r_\theta[t])-W_\theta[t]$. The operator profit can be represented as

$$\begin{aligned} R[t] &= \int_{\theta_l}^{\theta_u}\left(\ln\left(1+\theta r_\theta[t]\right)-r_\theta[t]c_\theta[t]\right)f(\theta)d\theta \\ &\quad -\int_{\theta_l}^{\theta_u}\int_{\theta_l}^{\theta}\frac{r_x[t]}{1+xr_x[t]}f(\theta)d\theta \\ &= \int_{\theta_l}^{\theta_u}\left(\ln\left(1+\theta r_\theta[t]\right)-r_\theta[t]c_\theta[t]\right)f(\theta) \\ &\quad -\frac{r_\theta[t]}{1+\theta r_\theta[t]}\left(1-F(\theta)\right)d\theta. \end{aligned} \tag{45}$$

Algorithm 2 Quality-Price Contract Design

```
Require: user type set $\Theta=[\theta_l,\theta_u]$; BS alternative set $\Psi$;
Ensure: optimal contract $\langle r^*_\theta[t],\pi^*_\theta[t]\rangle$; operator profit $R$;
 1: for $\theta\in\Theta$ do
 2:     Set $R(r_\theta[t])=(\ln(1+\theta r_\theta[t])-r_\theta[t]\cdot c_\theta[t]-\frac{r_\theta[t]}{1+\theta r_\theta[t]}\frac{1-F(\theta)}{f(\theta)})f(\theta)$
 3:     Set $r^*_\theta[t]=\arg\max_r R(r_\theta[t])$
 4: end for
 5: while $r^*_\theta[t]$ is not feasible do
 6:     Make some adjustment to $\boldsymbol{s_m}[t]$
 7:     Find an infeasible region $[a,b]\subseteq\Theta$
 8:     Set $r^*_\theta[t]=\arg\max_r\int_a^b R(r_\theta[t])d\theta,\ \forall\theta\in[a,b]$
 9: end while
10: for $\theta\in\Theta$ do
11:     Set $\pi^*_\theta[t]=\left(\ln(1+\theta r^*_\theta[t])-\frac{r^*_\theta[t]}{1+\theta r^*_\theta[t]}\frac{1-F(\theta)}{f(\theta)}\right)f(\theta)$
12:     Set $R=\int_{\theta\in\Theta}\left(\pi^*_\theta[t]-r^*_\theta[t]c_\theta[t]\right)f(\theta)d\theta$
13: end for
```

The maximization of $R[t]$ with respect to $r_{(\cdot)}[t]$ requires that the term under this integral be maximized with respect to $r_{(\cdot)}[t]$. Therefore, the relaxed CO problem can be written as

$$\max_{r_\theta[t]}\left(\ln(1+\theta r_\theta[t])-r_\theta[t]c_\theta[t]-\frac{r_\theta[t]}{1+\theta r_\theta[t]}\frac{1-F(\theta)}{f(\theta)}\right)f(\theta). \tag{46}$$

By solving this problem, we obtain optimal quality $\hat{r}^*_\theta[t]$ for the relaxed CO problem.

Moreover, we also need to check the **feasibility** of the solution $\hat{r}^*_\theta[t]$. The solution is feasible only if both monotonicity condition in (41), i.e., $dr_\theta[t]/d\theta\geq 0$ and BS alternative condition in (39), i.e., $r_\theta[t]\in\Psi$, are satisfied. We regard the solution satisfying these two conditions as our desired optimal quality $r^*_\theta[t]$. Otherwise, we need to make modifications to the infeasible solution $\hat{r}^*_\theta[t]$. Actually, the offloading quality can be adjusted by making effective resource allocation, which enables the optimal offloading quality available to users as much as possible. Due attention should be given spectrum resource allocation $\boldsymbol{s_m}[t]$ in modifying infeasible solutions. In this work, congestion-based resource allocation is adopted, where the allocated resource depends on BS congestion level and operator prefers allocating more resource to lightly-loaded BSs. On the basis of this, we implement "Bunching and Ironing" algorithm to make the infeasible region to be feasible. More details of this algorithm are shown in [20].

We obtain the corresponding optimal price for each quality $r^*_\theta[t]$, i.e.,

$$\pi^*_\theta[t]=\left(\ln(1+\theta r^*_\theta[t])-\frac{r^*_\theta[t]}{1+\theta r^*_\theta[t]}\frac{1-F(\theta)}{f(\theta)}\right)f(\theta). \tag{47}$$

The contract design is presented in Algorithm 2. In particular, the monotonicity and BS alternative conditions are validated, and those infeasible $\hat{r}^*_\theta[t]$ are modified in Lines 5-9.

V. Simulation Results

We validate the performance of our scheme in improving operator profit, and demonstrate how to guarantee user conditional incentive by leveraging pricing and resource allocation.

Our simulation is conducted in a general HCN, where the marcocell is located at the center of region and 16 WiFi APs are uniformly installed around it. We assume that users are distributed according to a homogeneous Poisson Point Process (PPP), which is commonly used in previous studies [16]. In addition, the packet arrival (i.e., requesting rate) at each user follows Bernoulli distribution [23]. The optimal contract is designed in continuous-user-type scenario, where user type θ follows a uniform distribution on an interval [0.1, 10].

We compare the performance of our proposed scheme against the associate-to-nearest offloading scheme. In associate-to-nearest scheme, users try to access to the nearest small BSs. Generally, compared with the macrocell, these BSs can provide high transmission rate at relative low price. Users will offload traffic immediately only if their nearest small BSs are available; otherwise they will make a service request for the cellular network [13].

A. Network Stability

The offloading quality represented by transmission rate is closely related to the allocated spectrum resources. In operator's load balancing, effective resource allocation can facilitate alleviating traffic congestion by encouraging users to associate with those lightly-loaded BSs. To characterize the impact of resource allocation on network stability, we

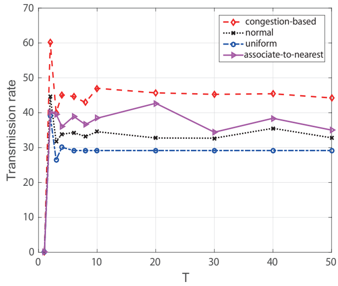


Fig. 2. Transmission rate vs. Time T.

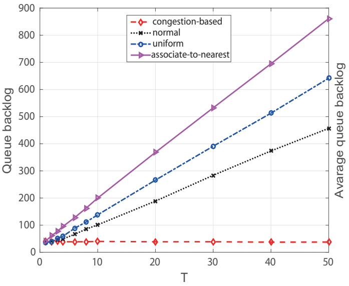


Fig. 3. Queue backlog vs. Time T.

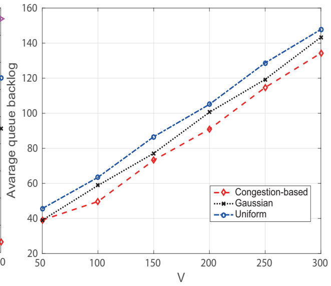


Fig. 4. Average queue backlog vs. Control parameter V.

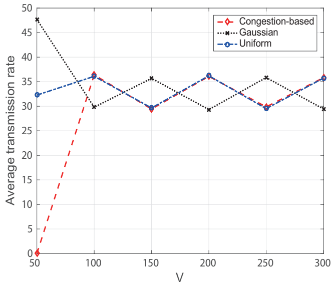


Fig. 5. Average transmission rate vs. Control parameter V.

implement Lyapunov-based traffic scheduling algorithm under the following three resource allocation schemes, respectively.

• Congestion-based allocation: The allocated resource depends on BS congestion level, and operator prefers to allocate more resource to lightly-loaded BSs to improve offloading quality.

• Gaussian allocation: The resource allocated to each BS follows a Gaussian distribution.

• Uniform allocation: Under this allocation, each BS obtains the same spectrum resource.

As for load balancing, we illustrate the variation of offloading quality and queue backlog in terms of time T as shown in Fig. 2 and Fig. 3. As expected, the proposed scheme is superior to the associate-to-nearest scheme in maintaining network stability. From Fig. 2, we observe that compared with the associate-to-nearest scheme, the proposed scheme shows high speed of convergence[4]. The intuitive is that in the proposed scheme, operator conducts traffic scheduling from the global perspective and encourage users to associate with lightly-loaded small BSs, which is effective in alleviating traffic congestion. In contrast, the associate-to-nearest scheme is implemented based on user local preference, indicating that high transmission rate will be provided. In practice, such local preference may result in heavy traffic congestion issues, making it hard for those BSs with limited service capacity to provide any satisfactory offloading service. That is why the transmission rate shows high fluctuation. As shown in Fig. 3, the large queue backlog can be observed due to the potential traffic congestion issues.

Our conditional incentive scheme is enabled under congestion-based allocation, where Gaussian and Uniform allocation schemes only act as benchmarks. Figure 2 presents operator obtains higher offloading quality under congestion-based allocation. In traffic scheduling, users are more likely to be pushed to remote small BSs. The increase of transmission distance makes offloading quality decreased under uniform allocation, and operator can only provide limited contract items for users. Under Gaussian allocation, the randomness in allocated resource may facilitate traffic alleviation to some extent. While under congestion-based case, operator prefers allocating more resource to lightly-loaded small BSs, making it possible for these BSs to provide relatively high-quality services. In this way, more users are encouraged to associate with lightly-loaded small BSs. Thus small queue backlog can be observed as shown in Fig. 3.

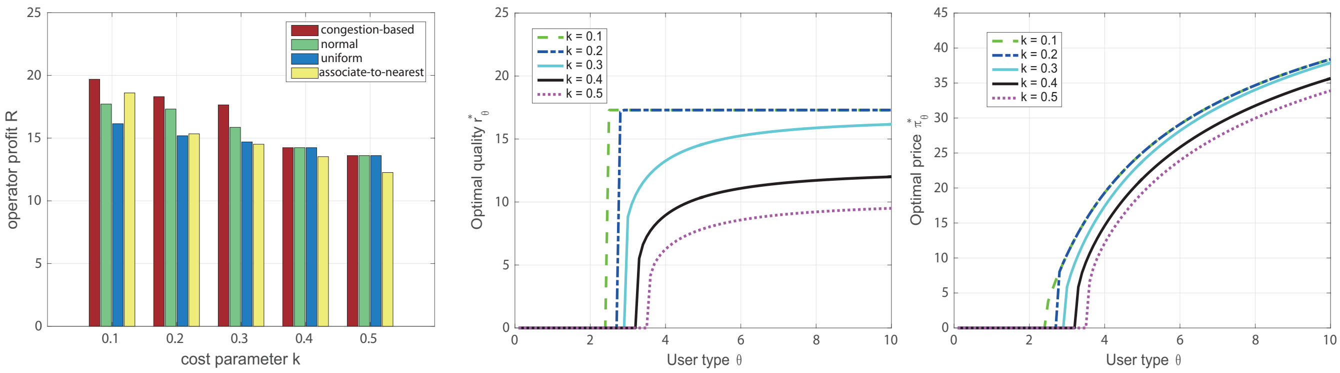
According to Theorem 1, there exists a tradeoff between operator profit and traffic congestion. In the context of Lyapunov optimization, control parameter V is introduced to characterize the heterogeneity in how much emphasis put on operator profit. In particular, we illustrate the impact of V on network stability by varying it from 50 to 300. From Fig. 4, we observe congestion-baed allocation is superior to other cases in decreasing queue backlog. Furthermore, with the increase of V, queue backlog under these three allocations increase. Figure 5 presents all average transmission rates are approaching to a steady state, and converge quickly as V increases.

B. Optimal Contract Design

Our contract-based incentive scheme is developed on the basis of long-term network stability. Specifically, we devise the optimal contract to maximize operator profit when $V = 300$ and $T = 50$. Due attention should be given to quality-specific operation cost, which always shows significant heterogeneity. We introduce cost parameter k, which is proportional to the unit operation cost $c_\theta[t]$, to illustrate the impact of offloading quality on operation cost. Actually, cost parameter is determined by network characteristics, e.g., channel condition and device energy consumption. To capture such heterogeneity, we set k to be 0.1, 0.2, 0.3, 0.4 and 0.5, respectively.

We demonstrate the comparison of operator profit in terms of cost parameter, shown in Fig. 6(a). We observe that operator profit in the proposed scheme is basically larger than that in the associate-to-nearest scheme. In the associate-to-nearest scheme, users always obtain relatively high service quality, especially in good channel condition. When k is small, quality has little impact on operation cost and thus operator can still gain high profit from high-quality users, even in serious network congestion. As k increases, the impact of quality becomes larger, and operator profit decreases accordingly. When k is large enough, the payment from high-quality users is not enough to compensate for the cost induced by huge quality impact, making operator profit decrease greatly. In the proposed scheme, however, under IC constraint, operator will

[4] Note that the unit of convergence time can match the time slot in network selection algorithm. The duration of each time slot depends on the specific dynamics of network state and operator's ability of traffic scheduling.



(a) Operator profit

(b) Optimal transmission rate

(c) Optimal price

Fig. 6. Operator profit R and optimal contract $\langle r^*_\theta[t], \pi^*_\theta[t] \rangle$ with respect to cost parameter k.

charge more money from high-quality users. Thus operator profit may not decrease too much.

As for Lyapunov-based scheduling scheme, we observe operator can achieve much more profit under congestion-based allocation. The effectiveness of resource allocation determines whether sufficient contract items can be provided. Under congestion-based allocation, users are guided to select the contract item specified for their types. High user satisfaction will attract more users to participate in offloading service, making operator profit increase accordingly. In addition, when k is small (e.g., $k = 0.1$), quality has a small impact on operation cost. Under congestion-based allocation, operator can provide more high-quality services for those users in lightly-loaded small BSs and users should pay much more money for such service. While under Gaussian and uniform allocation, operator obtains relatively low profit since only low-quality services can be provided. Thus there exists a large profit gap between these schemes. When k turns large (e.g., $k = 0.5$), quality has a huge impact on operation cost and user payment for high-quality service is not sufficient to compensate for it. Then a reduction in profit gap can be observed.

Figures 6(b) and 6(c) demonstrate the optimal quality-price contract with respect to user type and cost parameter under congestion-based allocation, which is similar to that under Gaussian and uniform allocations. As shown in Fig. 6(b), the optimal quality $r^*_\theta[t]$ basically increases with θ, which is consistent with the monotonicity condition in (41). With the increase of k, quality has a growing impact on operation cost. Thus given user payment, the obtained quality decreases. When θ is small, $r^*_\theta[t]$ remains unchanged at 0 under IR constraint. We can also observe it is upper bounded by a constant quality. It is an interesting observation and we can understand it in this way. To maintain network stability, the optimal quality is chosen from BS alternative set, i.e., finite contract items can only provided. Figure 6(c) shows the optimal price $\pi^*_\theta[t]$ increases with θ as well. The intuitive is that under IC constraint, users should pay much more money to operator for the increased quality. In addition, the increase of k always accompanies with high operation cost and low offloading quality. Then the optimal price decreases.

VI. Conclusion

We propose Lyapunov-based traffic scheduling scheme for congestion-aware cellular offloading, where user participation incentive is guaranteed conditionally. Specifically, we develop a congestion-aware network selection algorithm, achieving network stability and load-balancing. On the basis of this, we devise the optimal quality-price contract to motivate users to make proper association strategy. Simulation results demonstrate the effectiveness of our scheme in improving operator profit. User incentive and network stability are also validated.

There are several possible extensions which may deserve further study. The first is to consider how to incorporate predictive scheduling into dynamic offloading to facilitate balancing the traffic globally. The second is to study the effect of user mobility on BS available set. For example, one user can walk into the coverage area of one certain BS which was not available before. We believe that user mobility can be further guided with effective pricing and resource allocation, which is promising to enhance system performance such as network stability and operator profit.

Appendix A
Proof of Lemma 1

Proof. Since $Q[t+1] = \max\{Q[t] - r[t], 0\} + A[t]$, we have

$$Q^2[t+1] \leq Q^2[t] + r^2[t] + A^2[t] + 2Q[t]\left(A[t] - r[t]\right).$$

Similarly, according to the dynamic of virtual queues in (22), we obtain

$$X^2_m[t+1] \leq X^2_m[t] + {\alpha_m}^2 + x^2_m[t] + 2X_m[t]\left(x_m[t] - \alpha_m\right).$$

Therefore, summing the above yields

$$\begin{aligned}
&\tfrac{1}{2}\left[Q^2[t+1] - Q^2[t] + \sum_{m\in\mathcal{B}}\left(X^2_m[t+1] - X^2_m[t]\right)\right] \\
&\leq \tfrac{1}{2}\left[r^2[t] + A^2[t] + \sum_{m\in\mathcal{B}} x^2_m[t] + \alpha^2_m\right] \\
&\quad + Q[t]\left(A[t] - r[t]\right) + \sum_{m\in\mathcal{B}} X_m[t]\left(x_m[t] - \alpha_m\right) \\
&\leq B + Q[t]\left(A[t] - r[t]\right) + \sum_{m\in\mathcal{B}} X_m[t]\left(x_m[t] - \alpha_m\right).
\end{aligned} \tag{48}$$

After adding $-VR[t]$ to both sides, we take the conditional expectation on both sides. Thus Lemma 1 can be proven. □

Appendix B
Proof of Theorem 1

Proof. Define

$$x[t] = \varepsilon \left(Q[t] + \sum_{m \in \mathcal{B}} X_m[t] \right) + VR^*, \tag{49}$$

$$y[t] = B + VR[t]. \tag{50}$$

Thus the Lyapunov drift condition in (27) can be written as

$$\Delta(\mathbf{\Theta}[t]) \le \mathbb{E}\{y[t] | \mathbf{\Theta}[t]\} - \mathbb{E}\{x[t] | \mathbf{\Theta}[t]\}.$$

Taking expectations of the above inequality with respect to the distribution of $\mathbf{\Theta}[t]$ and using the law of iterated expectations yields

$$\mathbb{E}\left\{L(\mathbf{\Theta}[t+1])\right\} - \mathbb{E}\left\{L(\mathbf{\Theta}[t])\right\} \le \mathbb{E}\{y[t]\} - \mathbb{E}\{x[t]\}.$$

The above inequality holds for all t. Summing both sides over $\tau \in \{0, 1, \cdots, M-1\}$ yields

$$\mathbb{E}\left[L(\mathbf{\Theta}(M))\right] - \mathbb{E}\left\{L(\mathbf{\Theta}(0))\right\} \le \sum_{\tau=0}^{M-1} \mathbb{E}\{y(\tau)\} - \sum_{\tau=0}^{M-1} \mathbb{E}\{x(\tau)\}. \tag{51}$$

Rearranging terms, dividing by M, and using non-negativity of $L(\mathbf{\Theta}[t])$, we have

$$\frac{1}{M} \sum_{\tau=0}^{M-1} \mathbb{E}\{x(\tau)\} \le \frac{1}{M} \sum_{\tau=0}^{M-1} \mathbb{E}\{y(\tau)\} + \frac{\mathbb{E}\left\{L(\mathbf{\Theta}(0))\right\}}{M}. \tag{52}$$

(Part 1: Proof of queue backlog) Taking the superior limits both sides of (52) as $M \to \infty$ and substituting t for M yield

$$\limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{x(\tau)\} \le \limsup_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{y(\tau)\}. \tag{53}$$

According to the definitions of $x[t]$ and $y[t]$, we have

$$\begin{aligned} &\limsup_{t \to \infty} \tfrac{1}{t} \textstyle\sum_{\tau=0}^{t-1} \varepsilon \cdot \mathbb{E}\left\{ Q[\tau] + \sum\limits_{m \in \mathcal{B}} X_m[\tau] \right\} + VR^* \\ &\le \limsup_{t \to \infty} \tfrac{1}{t} \textstyle\sum_{\tau=0}^{t-1} \mathbb{E}\left\{B + VR[\tau]\right\} \le B + V\overline{R}. \end{aligned} \tag{54}$$

Shifting terms and dividing by ε yield (29).

(Part 2: Proof of operator profit) Taking the inferior limits both sides of (52) as $M \to \infty$ and substituting t for M yield

$$\begin{aligned} &\liminf_{t \to \infty} \tfrac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{x(\tau)\} \le \liminf_{t \to \infty} \tfrac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{y(\tau)\} \\ &\le \liminf_{t \to \infty} \tfrac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{B + VR(\tau)\} \\ &= B + V \liminf_{t \to \infty} \tfrac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R(\tau)\}. \end{aligned} \tag{55}$$

Note that $x[t] \ge VR^*$ and we can get

$$VR^* \le B + V \liminf_{t \to \infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\{R(\tau)\}.$$

Dividing both sides by V and shifting terms yield the average operator profit in (30).

Combining the above two cases, we have completed the proof of Theorem 1. □

Appendix C
Proof of Lemma 2

Proof. We assume that $\theta_l \le \theta_1 \le \theta_2 \le \theta_u$. According to the assumption, IC constraint is satisfied for all types. Then we can conclude that

$$\begin{aligned} \ln\left(1 + \theta_2 r_{\theta_2}[t]\right) - \pi_{\theta_2}[t] &\ge \ln\left(1 + \theta_2 r_{\theta_1}[t]\right) - \pi_{\theta_1}[t] \\ &\ge \ln\left(1 + \theta_1 r_{\theta_1}[t]\right) - \pi_{\theta_1}[t]. \end{aligned} \tag{56}$$

By iterating, we can obtain that $\ln\left(1 + \theta_2 r_{\theta_2}[t]\right) - \pi_{\theta_2}[t] \ge \ln\left(1 + \theta_l r_{\theta_l}[t]\right) - \pi_{\theta_l}[t]$. Due to the random selection of θ_2, we have

$$\ln\left(1 + \theta r_\theta[t]\right) - \pi_\theta[t] \ge \ln\left(1 + \theta_l r_{\theta_l}[t]\right) - \pi_{\theta_l}[t], \forall \theta \in [\theta_l, \theta_u].$$

In order to satisfy IR constraint for all contract items, we only need to guarantee $\ln\left(1 + \theta_l r_{\theta_l}[t]\right) - \pi_{\theta_l}[t] \ge 0$. Thus we complete the proof of Lemma 2. □

Appendix D
Proof of Lemma 3

Proof. It is easy to verify that user utility function in (34) satisfies the SMC constraint. Next, we will prove the monotonicity condition and local incentive compatibility constraint hold if the IC constraint holds. Suppose that $r_\theta[t]$ and $\pi_\theta[t]$ are differentiable about θ. Given that the optimality of $\hat{\theta}$, the following first- and second-order conditions are satisfied at $\hat{\theta} = \theta$:

$$\frac{\hat{\theta} r'_\theta[t]}{1 + \hat{\theta} r_\theta[t]} - \pi'_\theta[t] = 0, \tag{57}$$

$$\frac{\hat{\theta} r''_\theta[t] \left(1 + \hat{\theta} r_\theta[t]\right) - \left(\hat{\theta} r'_\theta[t]\right)^2}{\left(1 + \hat{\theta} r_\theta[t]\right)^2} - \pi''_\theta[t] \le 0. \tag{58}$$

Obviously, the first condition is the same as the local incentive compatibility condition. If we further differentiate the expression in (57) with respect to θ, we have

$$\frac{r'_\theta[t] + \theta r''_\theta[t]}{1 + \theta r_\theta[t]} - \frac{\theta r'_\theta[t] \cdot \left(r_\theta[t] + \theta r'_\theta[t]\right)}{\left(1 + \theta r_\theta[t]\right)^2} - \pi''_\theta[t] = 0. \tag{59}$$

Based on the fact that $\frac{r'_\theta[t]}{1 + \theta r_\theta[t]} \ge 0$, subtracting (58) from (59) yields $r'_\theta[t] \ge 0$.

On the other hand, we will prove that if the monotonicity condition and local incentive compatibility condition hold, the IC constraint holds. By contradiction, we suppose that the IC constraint is violated for at least one type θ, i.e.,

$$\ln\left(1 + \theta r_\theta[t]\right) - \pi_\theta[t] < \ln\left(1 + \theta r_{\hat{\theta}}[t]\right) - \pi_{\hat{\theta}}[t], \tag{60}$$

for at least one $\hat{\theta} \ne \theta$. After integrating, we obtain

$$\int_\theta^{\hat{\theta}} \frac{\theta r'_x[t]}{1 + \theta r_x[t]} - \pi'_x[t] \, dx > 0. \tag{61}$$

According to the monotonicity condition, we have $\frac{dr_x[t]}{dx} \ge 0$. If $\hat{\theta} > \theta$, we can get

$$\frac{\theta r'_x[t]}{1 + \theta r_x[t]} < \frac{x r'_x[t]}{1 + x r_x[t]}. \tag{62}$$

Since the local incentive compatibility condition holds, we obtain

$$\int_{\theta}^{\hat{\theta}} \frac{\theta r_x'[t]}{1+\theta r_x[t]} - \pi_x'[t]\, dx < 0, \tag{63}$$

which contradicts with (61). Similarly, we can derive contradiction when $\hat{\theta} < \theta$. Thus we obtain if the monotonicity and local incentive compatibility conditions hold, IC constraint holds.

Combining the above two cases, we have completed the proof of this lemma.

□

Acknowledgement

This work was supported in part by NSF China under Grant 61672342, Grant 91438115, Grant 61671478, Grant 61532012, Grant 61325012, Grant 61602303, Grant 61521062, and Grant 61601287.

References

[1] Cisco White Paper, "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2015-2020," Feb. 2016.

[2] Ericsson, Stockholm, Sweden, "Ericsson Mobility Report," Ericsso Press Release, Jun. 2016.

[3] K. Lee, I. Rhee, J. Lee, Y. Yi, and S. Chong, "Mobile Data Offloading: How Much Can WiFi Deliver?" in *Proc. of ACM CoNEXT*, Nov. 2010.

[4] A. Balasubramanian, R. Mahajan, and A. Venkataramani, "Augmenting Mobile 3G Using WiFi," in *Proc. of ACM MobiSys*, Jun. 2010.

[5] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in Heterogeneous Networks: Modeling, Analysis, and Design Insights," in *IEEE Transactions on Wireless Communications*, vol. 12, no. 5, pp. 2484-2497, May. 2013.

[6] S. Singh and J. G. Andrews, "Joint Resource Partitioning and Offloading in Heterogeneous Cellular Networks," in *IEEE Transactions on Wireless Communications*, vol. 13, no. 2, pp. 888-901, Feb. 2014.

[7] H. Yu, M. H. Cheung, L. Huang and J. Huang, "Predictive Delay-Aware Network Selection in Data Offloading," in *Proc. of IEEE GLOBECOM*, Dec. 2014.

[8] X. Kang, Y-K. Chia, S. Sun, and H. F. Chong, "Mobile Data Offloading Through A Third-Party WiFi Access Point: An Operator's Perspective," in *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5340-5351, Oct. 2014.

[9] M. H. Cheung and J. Huang, "DAWN: Delay-Aware Wi-Fi Offloading and Network Selection," in *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 6, pp. 1214-1223, Jun. 2015.

[10] Y. Li, J. Zhang, X. Gan, L. Fu, H. Yu, and X. Wang, "A Contract-Based Incentive Mechanism for Delayed Traffic Offloading in Cellular Networks," in *IEEE Transactions on Wireless Communications*, vol. 15, no. 8, pp. 5314-5327, Aug. 2016.

[11] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, et.al, "User Association for Load Balancing in Heterogeneous Cellular Networks," in *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706-2716, Jun. 2013.

[12] W. Dong, S. Rallapalli, R. Jana, L. Qiu, K. Ramakrishnan, L. Razoumov, Y. Zhang, and T. Cho, "iDEAL: Incentivized Dynamic Cellular Offloading via Auctions," in *Proc. of IEEE INFOCOM*, Apr. 2013.

[13] W. Wang, X. Wu, L. Xie, and S. Lu, "Femto-Matching: Efficient Traffic Offloading in Heterogeneous Cellular Networks," in *Proc. of IEEE INFOCOM*, Apr. 2015.

[14] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of Mobile Data Offloading," in *Proc. of IEEE INFOCOM*, Apr. 2013.

[15] L. Gao, G. Iosifidis, J. Huang, L. Tassiulas, and D. Li, "Bargaining-Based Mobile Data Offloading," in *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1114-1125, Jun. 2014.

[16] T. Z. Oo, N. H. Tran, W. Saad, D. Niyato, Z. Han, and C. S. Hong, "Offloading in HetNet: A Coordination of Interference Mitigation, User Association and Resource Allocation," in *IEEE Transactions on Mobile Computing*, accepted.

[17] P. Bolton and M. Dewatripont, "Contract Theory," The MIT Press, 2005.

[18] L. Duan, L. Gao, and J. Huang, "Cooperative Spectrum Sharing: A Contract-Based Approach," in *IEEE Transactions on Mobile Computing*, vol. 13, no. 1, pp. 174-187, Jan. 2014.

[19] Z. Zhou, X. Feng, X. Gan, F. Yang, X. Tian, and X. Wang, "Data Offloading in Two-tire Networks: A Contract Design Approach," in *Proc. of IEEE GLOBECOM*, Dec. 2014.

[20] L. Gao, X. Wang, Y. Xu, and Q. Zhang, "Spectrum Trading in Cognitive Radio Networks: A Contract-Theoretic Modeling Approach," in *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 4, pp. 843-855, Apr. 2011.

[21] E. Aryafar, A. Keshavarz-Haddad, M. Wang, and M. Chiang, "RAT Selection Games in Hetnets," in *Proc. of IEEE INFOCOM*, Apr. 2013.

[22] M. H. Cheung, R. Southwell, and J. Huang, "Congestion-Aware Network Selection and Data Offloading," in *Proc. of IEEE CISS*, Mar. 2014.

[23] S. Li, J. Huang, and S. R. Li, "Dynamic Profit Maximization of Cognitive Mobile Virtual Network Operator," in *IEEE Transaction on Mobile Computing*, vol. 13, no. 3, pp. 526-540, Mar. 2014.

[24] H. Beyranvand, M. Lvesque, M. Maier, J. A. Salehi, C. Verikoukis, and D. Tipper, "Toward 5G: FiWi Enhanced LTE-A HetNets With Reliable Low-Latency Fiber Backhaul Sharing and WiFi Offloading," in *IEEE/ACM Transaction on Networking*, vol. 25, no. 2, pp. 690-707, Apr. 2017.

[25] Q. Chen, G. Yu, H. Shan, A. Maaref, G. Y. Li, and A. Huang, "Cellular Meets WiFi: Traffic Offloading or Resource Sharing?" in *IEEE Transactions on Wireless Communications*, vol. 15, no. 5, pp. 3354-3367, May 2016.

[26] K. Son, S. Lee, Y. Yi, and S. Chong, "REFIM: A Practical Interference Management in Heterogeneous Wireless Access Networks," in *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 6, pp. 1260-1272, Jun. 2011.

[27] C. K. Ho, D. Yuan, and S. Sun, "Data Offloading in Load Coupled Networks: A Utility Maximization Framework," in *IEEE Transactions on Wireless Communications*, vol. 13, no. 4, pp. 1921-1931, Apr. 2014.

[28] F. Mehmeti and T. Spyropoulos, "Performance Analysis of Mobile Data Offloading in Heterogeneous Networks," in *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 482-497, Feb. 2017.

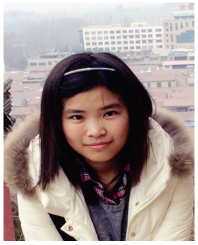
[29] R. Urgaonkar and M. J. Neely, "Opportunistic Scheduling with Reliability Guarantees in Cognitive Radio Networks," in *Proc. of IEEE INFOCOM*, Apr. 2008.

[30] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Resource Allocation and Cross-Layer Control in Wireless Networks," in *Foundations and Trends in Networking*, vol. 1, no. 1, pp. 1-149, 2006.

[31] S. Paris, F. Martignon, I. Filippini and L. Chen, "A Bandwidth Trading Marketplace for Mobile Data Offloading," in *Proc. of IEEE INFOCOM*, Apr. 2013.

[32] M. J. Neely, "Stochastic Network Optimization with Application to Communication and Queueing Systems," Morgan & Claypool, 2010.

[33] Y. Qin, J. Zheng, X. Wang, H. Luo, H. Yu, et.al, "Opportunistic Scheduling and Channel Allocation in MC-MR Cognitive Radio Networks," in *IEEE Transaction on Vehicular Technology*, vol. 63, no. 7, pp. 3351-3368, Sep. 2014.

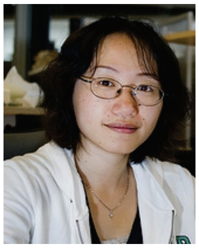


Yuqing Li received the B.S. degree in Communication Engineering from Xidian University, Xi'an, China, in 2014, and is currently pursuing the Ph.D. degree in Electronic Engineering at Shanghai Jiao Tong University, Shanghai, China. Her current research interests include network economics, social aware networks, heterogeneous cellular networks, and network security.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TWC.2017.2724027, IEEE Transactions on Wireless Communications

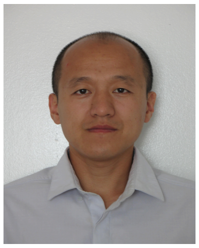
Bingyu Shen is an undergraduate student in Department of Computer Science at Shanghai Jiao Tong University, China. He is currently working as an research intern supervised by Prof. Xiaoying Gan. His research interests include network performance optimization, network security and system dependability.

Jinbei Zhang received the B.S. degree in Electronic Engineering from Xidian University, Xi?an, China, in 2010, and the Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2016. His current research interests include network security, asymptotic analysis, and coded caching.



Xiaoying Gan received the Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006. From 2009 to 2010, she was a Visiting Researcher with the California Institute for Telecommunications and Information, University of California at San Diego, San Diego, CA, USA. She is currently an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University. Her current research interests include network economics, social aware networks, heterogeneous cellular networks, multiuser multi-channel access, and dynamic resource management.

Jingchao Wang received a Ph.D. degree in Electronics Engineering from Tsinghua University, Beijing, China. His research interests include space information networks and satellite communications.



Xinbing Wang received the B.S. degree (with honors) from the Department of Automation, Shanghai Jiao Tong University, Shanghai, China, in 1998, and the M.S. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2001, and the Ph.D. degree (major from the Department of Electrical and Computer Engineering and minor from the Department of Mathematics) from the North Carolina State University, Raleigh, NC, USA, in 2006. He is currently a Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University. He has been an Associate Editor of the IEEE TRANSACTIONS ON NETWORKING/ACM Transactions on Networking and the IEEE TRANSACTIONS ON MOBILE COMPUTING, and a member of the technical program committees of several conferences, including the ACM MobiCom 2012, the ACM MobiHoc 2012 and 2013, and the IEEE INFOCOM from 2009 to 2014.