Data Demand Dynamics in Wireless Communications Markets

Shaolei Ren, *Student Member, IEEE*, and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—In this paper, we focus on the users' aggregate data demand dynamics in a wireless communications market served by a monopolistic wireless service provider (WSP). Based on the equilibrium data demand, we optimize the WSP's data plans and long-term network capacity decisions to maximize its profit. First, by considering a market where only one data plan is offered, we show that there exists a unique equilibrium in the data demand dynamics regardless of the data plans, and that the convergence of data demand dynamics is subject to the network congestion cost, which is closely related to the WSP's long-term capacity decision. A sufficient condition on the network congestion cost indicates that the WSP needs to provide a sufficiently large network capacity to guarantee the convergence of data demand dynamics. We also propose a heuristic algorithm that progressively optimizes the WSP's data plan to maximize its equilibrium revenue. Next, we turn to a market where two different data plans are offered. It is shown that the existence of a unique equilibrium data demand depends on the data plans, and the convergence of data demand dynamics is still subject to the network congestion cost (and hence, the WSP's network capacity, too). We formalize the problem of optimizing the WSP's data plans and network capacities to maximize its profit. Finally, we discuss the scenario in which the data plans are offered by two competing WSPs and conduct extensive simulations to validate our analysis.**

***Index Terms*— Congestion, data demand dynamics, pricing, wireless markets.**

I. Introduction

WE have witnessed over the last decade a successful proliferation of wireless networks, which support a variety of services and applications, and increasingly heated competition among the wireless service providers (WSPs). To sustain their competitive positions in the market and increase revenues, WSPs themselves will need to appropriately price their scarce network resources and expand their network capacities to support the unprecedented amount of wireless traffic. Hence, it becomes of paramount importance for these WSPs to understand how the aggregate data demand of all the subscribers evolves and how the demand is affected by various pricing plans.

Manuscript received January 17, 2011; revised August 13, 2011; accepted November 04, 2011. Date of publication December 01, 2011; date of current version March 06, 2012. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Martin Schubert. This work was supported in part by the National Science Foundation by Grant 0830556.

The authors are with Electrical Engineering Department, University of California, Los Angeles, CA 90095 USA (e-mail: rsl@ee.ucla.edu; mihaela@ee.ucla.edu).

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/TSP.2011.2177826

In this paper, we are interested studying the users' aggregate data demand dynamics, and optimizing the WSP's data plans and network capacities in a wireless communications market. In general, the WSP's network capacity is difficult to change once it is deployed and hence, it is a long-term strategy for the WSP to decide its network capacity [8]. In contrast, the WSP can adjust its data plans over the lifespan of its network infrastructure, although the data plans cannot be updated as frequently as the users change their data plan subscription. Overall, we will assume that the users may change their data plan subscription frequently based on their short-term (e.g., a few days or weeks per period) decisions, the WSP's data plans are changed less frequently based on the WSP's medium-term (e.g., several months or years per period) decisions, while the WSP's network capacity decision is a long-term (e.g., several years per period) decision. In order to evaluate and compare the long-term profitability of networks with different capacities, the WSP needs to predict its maximum profit for each network capacity configuration. To maximize revenue given the network capacity and the associated cost, the WSP needs to know the users' aggregate data demand and their willingness to pay for the service, and then choose its optimal data plans. Hence, by using backward induction, we study first the users' dynamic decisions as to whether or not they subscribe to the WSP's data plans (i.e., short-term problem), then the WSP's revenue-maximizing data plans (i.e., medium-term problem), and finally the WSP's network capacity decision (i.e., long-term problem). Note that we assume in our study that the medium-term period is sufficiently short compared to the long-term period, while it is sufficiently long compared to the short-term period.

We consider a wireless market with a monopolistic WSP serving a sufficiently large number of users. For the sake of analysis, we consider that the WSP can offer one or two data plans, while each user can subscribe to one of the available data plans. Due to the resource constraint (e.g., network capacity), congestion effects are observed when multiple users share the same network, degrading the network performance (e.g., increasing delays). Essentially, congestion effects are a type of negative network externalities and have similar impacts to prices on the users' experiences (i.e., utilities). Thus, congestion effects are also referred to as *congestion costs* in the literature [14], [23]. Taking into consideration the charged price and congestion cost, each user can dynamically decide whether to subscribe to the WSP's service and which data plan to subscribe to. First, by considering a market where only one data plan is offered, we show that there exists a unique equilibrium in the data demand dynamics regardless of the data plan or congestion costs. Nevertheless, the convergence of data demand dynamics is subject to the network congestion cost, which is closely related to the WSP's long-term capacity decision. We derive a sufficient condition for the convergence of data demand dynamics, indicating that the WSP needs to provide a sufficiently large network capacity. A heuristic algorithm is also proposed to progressively optimize the WSP's data plan such that its equilibrium revenue is maximized. Next, we turn to a market where two different data plans are offered. We show that the existence of a unique equilibrium data demand depends on the data plans. Moreover, the convergence of data demand dynamics is still subject to the network congestion cost (and hence, the WSP's network capacity, too). The problem of optimizing the WSP's data plans and network capacities is formalized and solved by numerical methods to maximize its profit. Next, we discuss the scenario in which the data plans are offered by two competing WSPs (i.e., a duopoly market) and find that the two WSPs only need to adjust their data plans a few times before reaching an equilibrium. Finally, extensive simulations are conducted to validate our analysis. Numerical results shows that, to maximize the profit, the WSP needs to increase the network capacity for its capped data plan while reducing the network capacity for its unlimited data plan. This coincides with the current trend that some WSPs have discontinued the offering of unlimited data plans [20].

The rest of this paper is organized as follows. We review the related literature in Section II. Section III describes the model. In Section IV and Section V, we study the data demand dynamics, data plan decision and capacity desicion for a wireless market where one and two data plans are offered, respectively. In Section VI, we provide numerical results to validate our analysis. Finally, we conclude this paper in Section VII.

II. Related Works

The engineering community has recently started to analyze as well as consider the design of existing and emerging wireless markets from various perspectives. Because of the space limitation, we only provide an incomplete list of related literature. In our previous work [1], we study the user subscription dynamics and revenue maximization in both monopoly and duopoly communications markets, based on a general distribution of users' valuation of quality-of-service (QoS) and a general QoS function that captures negative network externalities. Focusing on two specific access technologies (i.e., wide and local area network), the authors in [2] apply a stochastic geometric model and study the convergence of user subscription dynamics. In [3], the authors showed that noncooperative communications markets suffer from unfair revenue distribution among the service providers and proposed a revenue-sharing mechanism that requires cooperation among the service providers. The behavior of users and its impact on the revenue distribution, however, were not explicitly considered in [3]. Reference [4] studies technology adoption and competition between incumbent and emerging network technologies. The model characterizing the users' valuation of QoS is restricted to uniform distributions, and only constant QoS functions and positive network effects are considered in [4]. The user evolution in wireless social community networks is investigated in [5], where a key assumption is that the social community network provides a higher QoS to each user as the number of subscribers increases. While this assumption is valid if the network coverage is the only factor that determines the QoS, it does not model the QoS degradation due to, for instance, user traffic congestions at the WSP. By taking into account the congestion cost (i.e., negative network externalities), [6] studies the feasibility of Paris Metro pricing (PMP) and shows sufficient conditions on the congestion cost functions, under which PMP leads to a higher revenue or social welfare than flat-rate pricing. Pricing decisions (restricted to unlimited data plans) and network capacity decisions in the presence of network congestion effects are studied in [8], where a missing part is the analysis of users' subscription decisions. Reference [10] investigates market dynamics emerging when next-generation networks and conventional networks coexist, by applying a market model that consists of content providers, service providers, and users. Nevertheless, the level of QoS that a certain technology can provide is not considered in the model. The authors in [11] formulate a rate allocation problem by incorporating the participation of content providers into the model, and derive equilibrium prices and data rates. In [13], time-dependent pricing is studied from the perspective of its efficiency in terms of revenues. In [14], an upper bound on the efficiency loss as a result of price competition is derived in the context of congested markets, where an infinite number of users can selfishly route their traffic through the network.

In the aforementioned works, however, several key points are neglected. First, user heterogeneity in terms of data demand is not considered in these works (except for [11]). Specifically, it has been an implicit yet common assumption in these works that every user has the same data demand when it subscribes to the service provider. In other words, the QoS provided by the service provider only depends on the number of subscribers, regardless of their actual demand. Hence, user heterogeneity in terms of data demand cannot be captured and the QoS characterization may not be accurate under this assumption. Second, in most of the works, only a single data plan (e.g., flat-rate or "unlimited," usage-based price) is considered. Nevertheless, with the exploding popularity of smart phones, multiple pricing schemes are emerging in the market. For instance, capped data pricing plans and unlimited data pricing plans are both available in current wireless markets. Last but not least, it remains unknown how the congestion costs affect the aggregate data demand dynamics, in terms of both the equilibrium point and convergence, and the resulting revenue of the WSP. To address all these concerns, we propose a unified model that captures the user heterogeneity in terms of data demand and various practical data plans. Then, we study the users' data demand dynamics, and the WSP's data plan decision and network capacity decision.

III. Model

Consider a wireless communications market where one monopolistic WSP, denoted by $\mathcal{W}$, offers to N users data communications service, which takes up an overwhelming majority of the wireless traffic [21]. By assuming that N is sufficiently large such that each user is negligible,[1] we use a continuum user population model and normalize the number of users to 1 [1]–[8]. In general, WSP $\mathcal{W}$ may offer multiple data plans, and users can choose any of the plans depending on their own preferences (the user choice shall be detailed later). As in [8], to keep the analysis tractable, we assume that WSP $\mathcal{W}$ offers up to two data plans, represented by $\mathcal{P}_1$ and $\mathcal{P}_2$, respectively. For notational convenience, we also refer to users that subscribe to the plan $\mathcal{P}_i$ as $\mathcal{P}_i$-users (or $\mathcal{P}_i$-subscribers), for $i = 1, 2$. Next, we shall provide the modeling details of the WSP and users.

A. WSP Model

Before entering a market, the WSP needs to first make investment in infrastructure. In this paper, we concentrate on the WSP' capacity deployment which, once determined, is difficult to adjust and hence is an irreversible long-term decision [7]. Denote by $C_i \geq 0$ the network capacity (normalized by the number of users N) that the WSP allocates to its data pricing plan $\mathcal{P}_i$, for $i = 1, 2$. Assuming that the WSP incurs an average cost of τ per unit capacity,[2] we can express the WSP's equilibrium profit per short-term period (i.e., users' subscription period) as

$$\Pi_{\mathcal{W}} = \sum_{i=1,2} \{R_i - \tau C_i\} \tag{1}$$

where R_i is the equilibrium revenue per short-term period derived from $\mathcal{P}_i$-users. Note that in (1), we neglect the recurring cost of serving the users, which can also be absorbed into the revenue R_i [7]. To maximize its profit given the users' rational decisions, the WSP shall strategically determine its capacities $\mathbf{C} = \{C_1, C_2\}$. After building the network, the WSP decides its data plans and may alter them throughout the network's lifespan.

In today's wireless market, the most popular data plans are "unlimited," "capped," and "usage-based," all of which can be represented by a unified pricing model specified by (p, d^*, γ): each subscriber pays a fixed subscription fee p that allows it to transmit and receive up to d^* units of data; for each unit of additional data usage exceeding the capped data limit d^*, the subscriber pays γ. In special cases, a capped data plan characterized by (p, d^*, γ) becomes a usage-based one if $p = 0$ and $d^* = 0$, and an unlimited data plan if $d^* = \infty$ or $\gamma = 0$. For analytical tractability and to gain insights on how the congestion costs affect the data demand dynamics, we assume that the WSP's data plan $\mathcal{P}_1 = (p_1, +\infty, 0)$ is "unlimited" whereas its data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ is "capped." [3] This assumption, which may seem strong, can be justified by noting that some WSPs have (partially) resorted to capped data plans in view of the soaring wireless data service demand that frequently clogs their network infrastructure.[4] Moreover, even if the WSP offers two capped data plans, it is likely that one of the data plans has a very high data limit, which only a negligible fraction of subscribers can exceed in practice, and thus this data plan is almost "unlimited" (see, e.g., [20]).

B. User Model

Due to the capacity constraint, the network becomes more congested (i.e., negative network externalities or effect) as more data flow is transmitted [2], [13]. Such an effect is quantified by the congestion cost, which has similar impacts to prices on the users' experiences (i.e., utilities) [14]. We denote the congestion cost associated with the data pricing plan $\mathcal{P}_i$ by $g_i(D_i, C_i)$, where $i = 1, 2$ and $D_i \geq 0$ is the aggregate data demand (i.e., the total data demand of all the $\mathcal{P}_i$-users over a certain period) and C_i is the capacity allocated to $\mathcal{P}_i$-users. Without causing ambiguity, we simplify $g_i(D_i, C_i)$ as $g_i(D_i)$ by removing C_i wherever applicable. An implicit assumption in the model is that congestion costs for different data plans are independent of each other, which may be achieved by splitting network capacity among the plans [7].

Users are heterogeneous in the sense that they may have different data service demand and different benefits of utilizing the WSP's communications service. To model the user heterogeneity, each user k is characterized by a two-element tuple (θ_k, d_k), where θ_k indicates user k's benefit from data service and d_k denotes its data demand over a certain period (e.g., a month or a day). The values of θ_k and d_k can be determined by various approaches. For instance, (θ_k, d_k) may be user k's intrinsic characteristic and not influenced by the WSP's pricing schemes. In such scenarios, each individual user has *inelastic* demand [13], [14], although the aggregate demand of all the users is still elastic and influenced by the prices. Mathematically speaking, when user k subscribes to the WSP's data plan $\mathcal{P}_i$, its utility is given by

$$u_{k,i} = \theta_k - g_i(D_i) - p_i - \gamma_i \left[d_k - d_i^*\right]^+ \tag{2}$$

where $[x]^+ = \max\{0, x\}$, and if its data demand exceeds the granted data limit d_i^*, the term $\gamma_i[d_k - d_i^*]^+$ is positive and represents the additional cost user k incurs. Similar utility functions have been used in [4], [6], [11]–[13], and references therein. The utility function in (2) can be interpreted as follows: θ_k represents the benefit that user k receives from d_k units of data service, $g_i(D_i)$ indicates the congestion cost (i.e., negative network externalities), and $p_i + \gamma_i[d_k - d_i^*]^+$ is the payment made to WSP W_i [12]. Users that do not subscribe to any data plans obtain zero utility. Now, we impose some standard assumptions on the users' data demand and their benefits, users' subscription decisions, and the congestion function $g_i(D_i)$.

Assumption 1: The users' benefits and their data demand follow a two-dimensional distribution whose joint density function $f(\theta, d)$ is defined on $\mathcal{U} = \{(\theta, d) | 0 \leq \theta \leq \theta_{\max}, 0 \leq d \leq d_{\max}\}$. For completeness of definition, we have $f(\theta, d) = 0$ for all $(\theta, d) \notin \mathcal{U}$. The cumulative density function is given by $F(\theta, d) = \int_{-\infty}^{d} \int_{-\infty}^{\theta} f(x, y) dx dy$ for $(\theta, d) \in \mathbb{R}^2$.

[1]Another interpretation of the continuum model is that there is a representative user which has the same characteristics (e.g., data demand) as each user i in the market with a certain probability.

[2]The cost is averaged over the lifespan of the network infrastructure. For instance, if a network with a lifespan of T short-term periods (i.e., users' subscription period) is built at a cost of $\tilde{\tau}$ per unit capacity, then the average cost per unit capacity is $\tau = \frac{\tilde{\tau}}{T}$.

[3]In the most general case where both data plans are "capped,"the approach of analysis in this paper is still applicable, although the analysis becomes more complicated.

[4]Starting from June 7, 2010, AT&T discontinued unlimited data plans to its new iPhone users and adopts a capped data plan as considered in this paper [20].

Assumption 2: Each user k subscribes to the data plan $\mathcal{P}_i$ if $u_{k,i} > u_{k,j}$ and $u_{k,i} \geq 0$ for $i, j \in \{1, 2\}$ and $i \neq j$. If $u_{k,1} = u_{k,2} \geq 0$, user k subscribes to the unlimited data plan $\mathcal{P}_1$.[5]

Assumption 3: $g_i(D_i)$ is a nonnegative, nondecreasing, and differentiable[6] function in $D_i \in [0, D_{\max}]$, where $D_{\max}$ is the maximum possible aggregated data demand, normalized with respect to the total population, and given by

$$D_{\max} \triangleq \int_{y=0}^{d_{\max}} \int_{x=0}^{\theta_{\max}} y f(x, y) dx dy. \tag{3}$$

We briefly explain the above three assumptions. Assumption 1 can be considered as an expression of user diversity in terms of the benefits and their data demand. The lower bound on the interval is set as zero to simplify the analysis, and this will be the case when there is enough diversity in the users so that there are nonsubscribers for any positive price [7], [8]. Assumption 2 captures the user rationality. A rational user will subscribe to the data plan that provides a higher utility if at least one data plan provides a nonnegative utility, and to neither data plan otherwise. Assumption 3 indicates an intuitive fact that the congestion cost that each user experiences when subscribing to the data plan $\mathcal{P}_i$ becomes larger when the aggregate data demand increases.

Before concluding this section, it is worthwhile to provide the following remarks regarding our model.

Remark 1: As in [13], we assume for the convenience of analysis that each individual user k has an inelastic and fixed demand d_k (and benefit θ_k, too). Alternatively, d_k can be determined by solving a utility maximization problem and θ_k is the maximum benefit that user k receives [11]. Nevertheless, given the WSP's data plans, (θ, d) still follows a certain distribution over all the users and thus, our approach can be viewed as a proxy to determine the users' demand and benefit, provided that the distribution does not change significantly with the data plans.

Remark 2: Compared to the congestion cost function used in the existing literature that disregards the user heterogeneity in terms of data demand and is defined solely in terms of the number of subscribers [1], [6], [8], $g_i(D_i)$ is more accurate in modeling the congestion effect. Whilst the actual congestion cost also depends on when the users utilize the network, we consider the congestion cost *averaged* over time and ignore the time dependency to keep the analysis tractable [14].

Remark 3: The shape of the congestion cost function $g_i(D_i)$ may be determined by various factors, including the network capacity, resource allocation schemes and/or scheduling algorithms used for the data plan $\mathcal{P}_i$. While our analysis applies to a general function $g_i(D_i)$ satisfying Assumption 3, we shall explicitly focus on the impacts of network capacities on $g_i(D_i)$ when we derive specific results or study the WSP's long-term capacity decision. For instance, a concrete example is given by $g_i(D_i) = \frac{D_i}{C_i}$, which has been widely used (with minor modification, e.g., assuming all the users have the same data demand) in the prior works [6], [8], [18], [19].[7]

Remark 4: In addition to negative network externalities (i.e., congestion costs in this paper), positive network externalities may also be observed in a communications network. For instance, when more users subscribe to the WSP's data plan, the value of communications service may become higher as more users can communicate with each other [4]. As in prior research (e.g., [2], [6]–[8], [14], [15], [23]), we neglect the positive network externalities and concentrate on the impacts of congestion effects on the users' subscription decisions.

IV. Wireless Communications Market: Single Data Plan

In this section, we study the wireless communications market where the WSP offers a single data plan. Without loss of generality, we assume that the offered data plan is $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ and, as aforementioned, the unlimited data plan $\mathcal{P}_1 = (p_1, \infty, 0)$ is a special case when the data limit is infinity. The timing (i.e., order of moves) can be described as follows.

Stage 1 (long-term): The WSP decides its network capacity C_2 to deploy to maximize its profit.

Stage 2 (medium-term): Given C_2, the WSP chooses its optimal data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ by specifying p_2, d_2^* and γ_2 to maximize its revenue.

Stage 3 (short-term): By jointly considering the congestion cost and offered data plan, users decide whether or not to subscribe to the WSP's service.

From the described timing, we see that the WSP can be regarded as the leader whereas the users are followers. Thus, in order to identify the optimal data plan and network capacity, the WSP needs to first know how the users make their subscription decisions. Therefore, we proceed with our analysis using backward induction.

A. Users' Subscription Decisions

Due to rationality, users will not choose to subscribe to the WSP's data plan $\mathcal{P}_2$ if they cannot obtain nonnegative utilities. Essentially, the subscription decision stage can be formalized as a noncooperative game with an infinite number of players, the solution to which is (Nash) equilibrium. At an equilibrium, if any, no users can gain more benefits by deviating from their decisions. In other words, the aggregate data demand of those users subscribing to the WSP's data plan does not change at the equilibrium. Given the WSP's long-term capacity decision and an aggregate data demand D_2 of the subscribers, the congestion cost is uniquely given by $g_2(D_2)$. Moreover, the users' subscription decisions are also determined based on the sign of (2), i.e., user k subscribes to the data plan $\mathcal{P}_2$ if and only if $u_{k,2} = \theta_k - g_2(D_2^*) - \gamma_2[d_k - d_2^*]^+ \geq 0$. Hence, we study

[5]Online surveys show that users generally prefer an unlimited data plan to a capped one [22]. Moreover, specifying an alternative tie-breaking rule (e.g., random selection between the two data plans) in case of $u_{k,1} = u_{k,2} \geq 0$ will not significantly affect the analysis of this paper.

[6]Since $g_i(\cdot)$ is defined on $[0, D_{\max}]$, we use a one-sided limit to define the derivative of $g(\cdot)$ at 0 and $D_{\max}$, e.g., $g_i'(0) = \frac{\lim_{D_i \to 0^+}[g_i(D) - g_i(0)]}{(D_i - 0)}$.

[7]Another congestion cost function widely adopted in the literature is $g_i(D_i) = \frac{1}{(C_i - D_i)}$, which satisfies Assumption 3. Thus, our analysis is also applicable if $g_i(D_i) = \frac{1}{(C_i - D_i)}$ is considered.

the users' subscription decisions at the equilibrium by specifying the equilibrium (aggregate) data demand D_2^*. First, we can mathematically express the equilibrium data demand as

$$D_2^* = h_2(D_2^*) = \int_{y=0}^{d_{\max}} \int_{x=g_2(D_2^*)+p_2+\gamma_2[y-d_2^*]^+}^{\theta_{\max}} y f(x,y)\,dx\,dy. \quad (4)$$

Thus, an equilibrium data demand exists if and only if the mapping $h_2(D_2^*)$ in (4) has at least one fixed point. Next, we formally define the equilibrium data demand as follows.

Definition 1: When only data plan $\mathcal{P}_2$ is offered, $D_2^* \in [0, D_{\max}]$ is an *equilibrium* data demand if it satisfies

$$h_2(D_2^*) = D_2^*. \quad (5)$$

We establish in the following proposition the existence and uniqueness of an equilibrium data demand D_2^*.

Proposition 1: For any data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, there exists a unique equilibrium data demand satisfying (5).

Proof: See Appendix A. □

It can be seen from Proposition 1 that the data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ uniquely determines the equilibrium data demand. Although it is in general rather difficult to express D_2^* as an explicit function of $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, we summarize in Proposition 2 the relation between the data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ and the equilibrium data demand D_2^*.

Proposition 2: For any congestion cost function $g_2(D_2)$ satisfying Assumption 3, the equilibrium data demand D_2^* has the following properties:

1) $D_2^* > 0$ if and only if $0 \le p_2 < [\theta_{\max} - g_2(0)]^+$;
2) D_2^* is nonincreasing in $p_2 \ge 0$;
3) D_2^* is nonincreasing in $\gamma_2 \ge 0$;
4) D_2^* is nondecreasing in $d_2^* \ge 0$.

■

Property 1 shows that no users will subscribe to the WSP's data plan $\mathcal{P}_2$ if the fixed subscription fee p_2 exceeds the maximum benefit among all the users minus the minimum congestion cost. Properties 2 and 3 are consistent with the standard demand-price relation: increasing the price will not increase the demand. Property 4 indicates that the data demand will increase or at least remain the same if the data limit d_2^*, which each subscriber can enjoy without incurring additional costs, increases. This stems from the fact that increasing the data limit results in the decrease of payment for users with high data demand exceeding the limit d_2^*.

In practice, the users do not have complete information regarding each other and hence, they may not make directly the subscription decisions that lead to an equilibrium. Instead, an adjustment process where the users update their subscription decisions based on limited information is required. To formally describe the adjustment process, we consider a discrete-time model denoted by $\{D_2^t\}_{t=0}^{\infty}$, where $D_2^t \ge 0$ is the (aggregate) data demand in the tth time period and $D_2^0 \in [0, D_{\max}]$ is the initial data demand. A natural and well-studied approach to modeling the adjustment process is the best-response dynamics, in which each decision maker chooses the best action in response to the decisions made by the others. As in [1], [5], [9], [15], we consider the best-response dynamics based on naive (or static) expectation, and assume that the users can only change their subscription decisions (e.g., opt out of the plan $\mathcal{P}_2$) at discrete time periods indexed by $t = 1, 2, \cdots$. Specifically, at the beginning of the time period t, user k holds a (static) belief on the congestion cost, denoted by $\tilde{g}_{2,k}(D_2^t) = g_2(D_2^{t-1})$, and makes its subscription decision in a myopic way [4], [5].[8] When only one data plan $\mathcal{P}_2$ is offered by the WSP, each user has a choice of whether to subscribe to the plan at the beginning of each time period. In particular, user k subscribes to the data plan $\mathcal{P}_2$ in the time period t if and only if it believes that its utility $\theta_k - \tilde{g}_{2,k}(D_2^t) - p_2 - \gamma_2[d_k - d_2^*]^+ = \theta_k - g_2(D_2^{t-1}) - p_2 - \gamma_2[d_k - d_2^*]^+ \ge 0$. Note that, in order to make subscription decisions at time t, the users need to know the data plan $\mathcal{P}_2$ and receive a signal indicating the congestion cost $g_2(D_2^{t-1})$ at $t-1$. The best-response decision model implies that, for $t = 1, 2, \ldots$, the data demand dynamics evolves following a sequence $\{D_2^t\}_{t=0}^{\infty}$ specified by

$$D_2^t = h_2\left(D_2^{t-1}\right) = \int_{y=0}^{d_{\max}} \int_{x=g_2(D_2^{t-1})+p_2+\gamma_2[y-d_2^*]^+}^{\theta_{\max}} y f(x,y)\,dx\,dy \quad (6)$$

starting from an initial point $D_2^0 \in [0, D_{\max}]$. Essentially, the dynamics in (6) is a fixed point iteration for $h_2(\cdot)$ and it converges regardless of the initial point if $|h_2'(D_2)| < 1$ for $D_2 \in [0, D_{\max}]$ [25]. Nevertheless, $|h_2'(D_2)| < 1$ may not hold for all congestion cost functions, resulting in oscillation in the data demand dynamics. In accordance, the WSP's revenue becomes instable and may cause higher risks for the WSP's operation in the market. Let us consider a hypothetical example to explain this point. Suppose that the network is highly underutilized in the time period t and each subscriber incurs a low congestion cost. Users expect that the congestion cost will remain low in the period $t+1$, and thus more users subscribe to the data plan $\mathcal{P}_2$, leading to a high congestion cost in the time period $t+1$. The increase of congestion cost in turn will induce a small amount of data demand in the time period $t+2$. When the congestion cost function is very sensitive to the aggregate data demand, the data demand dynamics may oscillate around or diverge away from the equilibrium point. In the following proposition, we provide a sufficient condition under which the data demand dynamics is guaranteed to converge regardless of the initial points.

Proposition 3: For any data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, the data demand dynamics specified by (6) converges to the unique equilibrium point starting from any initial point $D_2^0 \in [0, \bar{d}]$ if

$$\max_{D_2 \in [0, D_{\max}]} g_2'(D_2) < \frac{2}{K \cdot d_{\max}^2} \quad (7)$$

where $D_{\max}$ is given by (3), d_{max} is the maximum individual demand and $K = \max_{(\theta,d)\in\mathcal{U}} f(\theta, d)$.

Proof: See Appendix B. □

[8]This model of belief formation is called naive or static expectations [24]. A similar dynamic model of belief formation and decision making has been extensively adopted in the existing literature such as [1], [4], [5], [9].

Proposition 3 states the relation between the congestion cost function and the distribution of (θ, d) such that the data demand dynamics converges, and holds for a general yet practical data plan. Although the convergence condition (7) is sufficient but not necessary, it provides us with the insight that, for a given distribution function $f(\theta, d)$, if the congestion cost increases too fast (i.e., $g_2'(D_2)$ is larger than $\frac{2}{(K \cdot d_{\max}^2)}$ for some $D_2 \in [0, D_{\max}]$), the data demand dynamics may oscillate or diverge. A similar insight was reported in [17] in the context of the decentralized spectrum access in cognitive networks. Next, by considering $g_2(D_2) = \frac{D_2}{C_2}$ for $D_2 \in [0, D_{\max}]$, we investigate the impacts of the WSP's long-term capacity decision on the convergence of the users' data demand dynamics. The result is summarized as follows.

Corollary 1: Suppose that $g_2(D_2) = \frac{D_2}{C_2}$ for $D_2 \in [0, D_{\max}]$. For any data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, the data demand dynamics specified by (6) converges to the unique equilibrium point starting from any initial point $D_2^0 \in [0, D_{\max}]$ if

$$C_2 > \frac{K \cdot d_{\max}^2}{2} \tag{8}$$

where $K = \max_{(\theta,d)\in\mathcal{U}} f(\theta, d)$. If $f(\theta, d) = 1$ over $\mathcal{U} = \{(\theta, d) | 0 \le \theta \le 1, 0 \le d \le 1\}$, then (8) becomes $C_2 > \frac{1}{2}$. ■

Corollary 1 indicates that the network capacity allocated to $\mathcal{P}_2$-users needs to be greater than a certain threshold such that the data demand dynamics is guaranteed to converge for any data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$. In particular, if (θ, d) is uniformly distributed over $\mathcal{U} = \{(\theta, d) | 0 \le \theta \le 1, 0 \le d \le 1\}$, then the capacity threshold (normalized with respect to the total number of users) corresponds to data demand averaged over all the users in the market. This implies that the network for the plan $\mathcal{P}_2$ needs to be able to accommodate all the users' data demand. Moreover, we see from (8) that the capacity threshold $K \cdot \frac{d_{\max}^2}{2}$ does not explicitly depend on $D_{\max}$. Instead, it is closely related to $d_{\max}$. In particular, if $d_{\max}$ increases, then a more stringent requirement is imposed on the WSP's network capacity in order to guarantee the convergence of data demand dynamics regardless of the initial points or data plans. On the other hand, if the network capacity is not large enough, then the users may experience excessive delays (i.e., high congestion costs) and the data demand dynamics may oscillate without convergence.

Before studying the WSP's data plan decision, we make two remarks regarding the users' subscription decisions.

Remark 5: The dynamics specified by (6) requires that all the users update subscription decisions at the beginning of each time period. In practice, if only a fraction $\epsilon \in (0, 1]$ of the user update subscription decisions each time, then the sequence becomes

$$D_2^t = \epsilon h_2 \left(D_2^{t-1}\right) + (1 - \epsilon) D_2^{t-1} \tag{9}$$

where $h_2(D_2^{t-1})$ is given by (6). The equilibrium analysis is not affected, whereas the convergence condition in (7) is modified as $\max_{D_2\in[0,D_{\max}]} g_2'(D_2) < \frac{2}{\epsilon\cdot(K \cdot d_{\max}^2)}$, which is more easily satisfied for a smaller $\epsilon \in (0, 1]$. In other words, the parameter $\epsilon \in (0, 1]$ smooths the data demand update process and makes the dynamics easier to converge by slowing down the convergence rate. Nevertheless, if the network capacity is large enough to serve all the users' data demand in practice, then the convergence can always be observed even though all the users update their subscription decisions. It should also be noted that another approach to modeling the users' data demand dynamics is considering a continuous-time dynamics specified as

$$\frac{dD_2}{dt} = \rho \cdot [h_2(D_2) - D_2] \tag{10}$$

where ρ is referred to as the diffusion rate [4]. For (10), the equilibrium is still defined the same as that in Definition 1, while the convergence is guaranteed. The considered discrete-time data demand dynamics has been studied in prior works (see, e.g., [5] and [9]) and is more appropriate for scenarios in which the users' subscription decisions can only change in discrete time instants (e.g., at the beginning of a day or month). Moreover, ϵ in (9) is essentially the same as ρ in (10) and the discrete-time dynamics considered in this paper will become (10) if the duration of a time period is sufficiently small.

Remark 6: As in the existing literature [1]–[6], the cost in updating the subscription decisions (e.g., time spent in calling the customer service, activation fees and early termination fees) are not considered in the paper. Here, we briefly discuss the impacts of this cost on the data demand dynamics. For simplicity, we assume that the cost of activating the data plan and that of terminating the subscription are the same, and we refer to this cost as *switching* cost denoted by c_s. With a switching cost, the users' subscription decisions are affected. Specifically, if user k is a subscriber in the time period t, it will continue the subscription in the next time period $t + 1$ if

$$\theta_k - g_2\left(D_2^t\right) - p_2 - \gamma_2 \left[d_k - d_2^*\right]^+ \ge -c_s. \tag{11}$$

On the other hand, if user k is not a subscriber in the time period t, it will choose to subscribe to the data plan in the next time period $t + 1$ if

$$\theta_k - g_2\left(D_2^t\right) - p_2 - \gamma_2 \left[d_k - d_2^*\right]^+ - c_s \ge 0. \tag{12}$$

It should be noted that if the cost is taken into account when the users make their subscription decisions, there may exist multiple equilibrium data demand points, and the convergence is subject to the initial point. For instance, in the extreme case in which the cost is so high (e.g., greater than $\theta_{\max}$) that no users would like to update their subscription decisions, every possible value of (aggregate) data demand $D_2 \in [0, D_{\max}]$ is an equilibrium point. We shall show in the numerical results the impact of switching cost c_s on the users' subscription decisions, while rigorous analysis of c_s is left as our future work.

B. WSP's Data Plan Decision

Over the entire lifespan of the network infrastructure, the WSP can change its data plans to maximize its revenue, although the change of data plans is sufficiently slow compared to the users' subscription decisions. In other words, the duration of a medium-term period corresponds to that of a sufficiently large number of short-term periods. We note that given the WSP's data plan, the data demand dynamics converges rapidly (e.g., within a few iterations) to the equilibrium point if the convergence condition (7) is satisfied. Thus, the WSP's average revenue per short-term period (i.e., users' subscription period) is approximately equal to its equilibrium revenue when the data demand reaches the unique equilibrium. Next, we derive the expression of the WSP's equilibrium revenue as follows:

$$R_2 = p_2 \int_{y=0}^{d_{\max}} \int_{x=g_2\left(D_2^*\right)+p_2+\gamma_2\left[y-d_2^*\right]^+}^{\theta_{\max}} f(x,y)dxdy + \int_{y=d_2^*}^{d_{\max}} \int_{x=g_2\left(D_2^*\right)+p_2+\gamma_2\left(y-d_2^*\right)}^{\theta_{\max}} \gamma_2\left(y-d_2^*\right) f(x,y)dxdy \quad (13)$$

where the first term on the right hand side is the subscription fee that every subscriber pays and the second term is the additional fee that users with demand higher than d_2^* pay. Although the equilibrium data demand D_2^* is uniquely determined by the WSP's data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ and hence can be expressed as an implicit function of $\mathcal{P}_2$, it is rather challenging to maximize the equilibrium revenue in (13). The difficulties are mainly: (1) D_2^* cannot be expressed explicitly in a closed-form function in terms of $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$; (2) due to the integral, the equilibrium revenue is not an explicit function of $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$. Thus, we resort to numerical methods to find the optimal $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ maximizing the equilibrium revenue. Specifically, we search over all the possible values of (p_2, d_2^*, γ_2) and select the one that yields the maximum equilibrium revenue. In practice, the data plan is typically confined within a small finite set of options[9] and hence, the complexity associated with the exhaustive search is not prohibitive.

In the following, we propose a heuristic algorithm that progressively chooses the (locally) optimal data plan in a greedy manner. For the ease of presenting the algorithm and deriving more specific results, we consider uniformly distributed (θ, d), i.e., $f(\theta, d) = 1$ over $\mathcal{U} = \{(\theta, d) | 0 \leq \theta \leq 1, 0 \leq d \leq 1\}$, although other forms of $f(\theta, d)$ can also be applied. Under the assumption of uniformly distributed of (θ, d) over $\mathcal{U} = \{(\theta, d) | 0 \leq \theta \leq 1, 0 \leq d \leq 1\}$, we rewrite the equilibrium revenue in (13) as

$$R_2 = p_2\left[1 - g_2\left(D_2^*\right) - p_2\right] + \frac{\gamma_2\left(1-d_2^*\right)^2}{2}\left[1 - g_2\left(D_2^*\right) - 2p_2 - \frac{2\gamma_2\left(1-d_2^*\right)}{3}\right] \quad (14)$$

where $p_2 \leq 1 - g_2(D_2^*)$, $d_2^* \leq 1$ and $\gamma_2 \geq 0$.[10] Note that, even if we artificially assume that the congestion cost $g_2(D_2^*)$ is independent of the data plan, (14) is nonconcave in (p_2, d_2^*, γ_2). Thus, there exist no efficient algorithms to find the optimal (p_2, d_2^*, γ_2). In the proposed heuristic algorithm, instead of jointly optimizing (p_2, d_2^*, γ_2), we optimize p_2, d_2^*, and γ_2 separately. Specifically, by assuming that the equilibrium data demand D_2^* is independent of p_2 and treating $g_2(D_2^*)$, d_2^* and γ_2 as fixed values, we choose the optimal p_2 to maximize (14). Then, we apply the same technique to optimize d_2^* and γ_2, and the same process repeats until the stopping criterion is satisfied (e.g., convergence or the maximum number of iterations is reached). To summarize, the heuristic algorithm is described in Algorithm I.

Algorithm 1 Find (p_2, d_2^*, γ_2)

$R_2 = 0$, $d_2^* \leftarrow \infty$, $\gamma_2 \leftarrow 0$, and $k \leftarrow 1$
while $k \leq MaxIterate$ **do**
  $temp \leftarrow R_2$
  Optimize p_2: $p_2 \leftarrow \left[\frac{1-g_2(D_2^*)-\gamma_2(1-d_2^*)^2}{2}\right]^+$
  Optimize d_2^*: $d_2^* \leftarrow \left[\frac{1-g_2(D_2^*)-p_2}{\gamma_2}\right]_0^1$
  Optimize γ_2: $\gamma_2 \leftarrow \left[\frac{3[1-g_2(D_2^*)-p_2]}{4(1-d_2^*)}\right]^+$
  Recalculate R_2 based on (14)
  if $\mathrm{abs}(temp - R_2) \leq \mathrm{threhold}$ **then**
    break
  end if
  Update D_2^*, $g_2(D_2^*)$ and $k++$
end while
return (p_2, d_2^*, γ_2)

C. WSP's Capacity Decision

We assume that the WSP's network capacity is chosen to guarantee the convergence of data demand dynamics to the unique equilibrium point regardless of the initial points.[11] For instance, if (θ, d), i.e., $f(\theta, d) = 1$ is uniformly distributed over $\mathcal{U} = \{(\theta, d) | 0 \leq \theta \leq 1, 0 \leq d \leq 1\}$ and $g_2(D_2) = \frac{D_2}{C_2}$, then the network capacity allocated to $\mathcal{P}_2$-users should be greater than $\frac{1}{(2\epsilon)}$, where ϵ is the fraction of users that update their subscription decisions in each time period. As can be seen from (13), given the WSP's capacity, it is rather difficult to find explicitly the optimal value of (p_2, d_2^*, γ_2) maximizing the WSP's equilibrium revenue. As a result, we can only numerically find the optimal network capacity to maximize the WSP's equilibrium profit in (1).

Finally, we note that if only the unlimited data plan $\mathcal{P}_1 = \{p_1, +\infty, 0\}$ is offered, the above analysis still applies and the corresponding result can be easily obtained by letting $d_2^* = \infty$ and $\gamma_2 = 0$.

V. Wireless Communications Market: Two Data Plans

In this section, we turn to the analysis of a wireless communications market where the WSP offers two data plans $\mathcal{P}_1$ and $\mathcal{P}_2$. Although we mainly focus on the scenario that these two data plans are offered by the same monopolistic WSP, we shall also briefly discuss at the end of this section the case in which they are offered by two competing WSPs.

A. Users' Subscription Decisions

As in a market with only one data plan $\mathcal{P}_2$, we study the users' equilibrium subscription decisions by specifying the equilibrium data demand (D_1^*, D_2^*). By Assumption 2, we see that the

[9]In practice, the subscription fee p_2 is usually selected from 9.99, 19.99, 24.99, 29.99 or a similar set of options.

[10]Since the maximum demand is $d_{\max} = 1$, $d_2^* > 1$ and $d_2^* = 1$ are essentially the same.

[11]This requires that the average cost of τ per unit capacity be sufficiently small such that the WSP can receive a nonnegative profit.

equilibrium data demand (D_1^*, D_2^*) satisfies the following equations:

$$D_1^* = h_{d,1}(D_1^*, D_2^*) = \int_{y=\tilde{d}}^{d_{\max}} \int_{x=g_1(D_1^*)+p_1}^{\theta_{\max}} yf(x,y)dxdy \quad (15)$$

$$D_2^* = h_{d,2}(D_1^*, D_2^*) = \int_{y=0}^{\tilde{d}} \int_{x=g_2(D_2^*)+p_2+\gamma_2[y-d_2^*]^+}^{\theta_{\max}} yf(x,y)dxdy \quad (16)$$

if $p_1 + g_1(D_1^*) > p_2 + g_2(D_2^*)$, and

$$D_1^* = h_{d,1}(D_1^*, D_2^*) = \int_{y=0}^{d_{\max}} \int_{x=g_1(D_1^*)+p_1}^{\theta_{\max}} yf(x,y)dxdy \quad (17)$$

$$D_2^* = h_{d,2}(D_1^*, D_2^*) = 0 \quad (18)$$

if $p_1 + g_1(D_1^*) \le p_2 + g_2(D_2^*)$. In (15) and (16), $\tilde{d}$ is given by

$$\tilde{d} = d_2^* + \frac{1}{\gamma_2}[p_1 - p_2 + g_1(D_1^*) - g_2(D_2^*)] \quad (19)$$

which specifies the data demand of marginal users that are "indifferent" between subscribing to the plan $\mathcal{P}_1$ and the plan $\mathcal{P}_2$ (see [1] and [6] for a detailed explanation of "indifferent"). Note that there are two regimes of the equilibrium data demand in the market with two data plans, and which regime governs the equilibrium depends on the relative values of the *effective* full price (not including the additional cost if the data demand exceeds the granted data limit), i.e., $p_1 + g_1(D_1^*)$ and $p_2 + g_2(D_2^*)$. Next, we give the formal definition of the equilibrium point (D_1^*, D_2^*), which is similar to Definition 1.

Definition 2: When two data plans $\mathcal{P}_1$ and $\mathcal{P}_2$ are offered, (D_1^*, D_2^*) is an *equilibrium* data demand if it satisfies

$$h_{d,1}(D_1^*, D_2^*) = D_1^* \quad \text{and} \quad h_{d,2}(D_1^*, D_2^*) = D_2^* \quad (20)$$

where $h_{d,1}(D_1^*, D_2^*)$ and $h_{d,2}(D_1^*, D_2^*)$ are given in (15)–(18).

When the unlimited data plan $\mathcal{P}_1$ is available in the market, there may not exist an equilibrium data demand if the plan $\mathcal{P}_2$ is "capped" (i.e., $d_2^* < d_{\max}$ and $\gamma_2 > 0$). Suppose, as a counterexample, that $g_2(D_2) = 0$ is a constant for $D_2 \in [0, D_{\max}]$. Thus, $h_{d,1}(D_1, D_2)$ in (15) and (17) is independent of D_2 and can be rewritten compactly as $h_{d,1}(D_1)$. From (15) and (17), we see that the integration interval is not continuous, implying that $\tilde{h}_{d,1}(D_1) = h_{d,1}(D_1) - D_1$ may not be a continuous function in $D_1 \in [0, D_{\max}]$. Specifically, if $p_1 + g_1(D_1) \le p_2$, then the integration interval is $[0, d_{\max}]$, i.e., no users subscribes to the plan $\mathcal{P}_2$, whereas if $p_1 + g_1(D_1) > p_2$, the integration interval is $[\tilde{d}, d_{\max}]$. According to Definition 2, the equilibrium data demand should satisfy $\tilde{h}_{d,1}(D_1^*) = h_{d,1}(D_1^*) - D_1^* = 0$. Although it is easy to show that $\tilde{h}_{d,1}(D_1)$ is strictly decreasing in $D_1 \in [0, D_{\max}]$, $\tilde{h}_{d,1}(0) \ge 0$ and $\tilde{h}_{d,1}(D_{\max}) \le 0$, it is not guaranteed that $\tilde{h}_{d,1}(D_1)$ has a root, since $\tilde{h}_{d,1}(D_1)$ may not be a continuous function in $D_1 \in [0, D_{\max}]$. In other words, an equilibrium data demand may not exist. Next, we provide a sufficient condition that establishes the existence and uniqueness of an equilibrium point in Proposition 4, whose proof is deferred to Appendix C.

Proposition 4: For any data plans $\mathcal{P}_1 = (p_1, +\infty, 0)$ and $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, there exists a unique equilibrium data demand (D_1^*, D_2^*) satisfying (15)–(18) if

$$d_2^* = 0 \quad \text{and} \quad \gamma_2 > 0. \quad (21)$$

Moreover, the equilibrium data demand (D_1^*, D_2^*) satisfies $D_1^* = h_{d,1}(D_1, 0^*)$ and $D_2^* = 0$ if $p_2 + g_2(0) \ge p_1 + g_1(D_1^*)$.

Proposition 4 indicates that, if the two data plans $\mathcal{P}_1$ and $\mathcal{P}_2$ are unlimited and usage-based, respectively, then the data demand admits a unique equilibrium point. It also shows that, if the effective subscription cost of the data plan $\mathcal{P}_1$ evaluated at D_1^* is always smaller than or equal to that of the data plan $\mathcal{P}_2$, then no users subscribe to the data plan $\mathcal{P}_2$ at the equilibrium point.

Following Section IV-A, we consider a discrete-time best-response dynamics to model the users' subscription decision process. With two data plans $\mathcal{P}_1 = (p_1, +\infty, 0)$ and $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ offered in the market, each user has three possible choices at the beginning of each time period: subscribe to the plan $\mathcal{P}_1$, subscribe to the plan $\mathcal{P}_2$, and subscribe to neither. The users expect that the congestion cost incurred when subscribing to a data plan in the time period t is equal to that in the previous period $t-1$ and make their subscription decisions to myopically maximize their utility in the time period t [1], [4], [5]. We assume that, other than the subscription price, there is no cost involved (e.g., initiation fees, termination fees, device prices) when users switch between the data plans $\mathcal{P}_1$ and $\mathcal{P}_2$ [4]. By Assumption 2, at period $t = 1, 2 \cdots$, user k subscribes to the data plan $\mathcal{P}_1$ if and only if

$$g_1(D_1^{t-1}) + p_1 \le g_2(D_2^{t-1}) + p_2 + \gamma_2[d_k - d_2^*]^+ \quad (22)$$

$$\theta_k - g_1(D_1^{t-1}) - p_1 \ge 0, \quad (23)$$

to the data plan $\mathcal{P}_2$ if and only if

$$g_2(D_2^{t-1}) + p_2 + \gamma_2[d_k - d_2^*]^+ < g_1(D_1^{t-1}) + p_1 \quad (24)$$

$$\theta_k - g_2(D_2^{t-1}) - p_2 - \gamma_2[d_k - d_2^*]^+ \ge 0, \quad (25)$$

and to neither data plan if and only if

$$\theta_k - g_1(D_1^{t-1}) - p_1 < 0 \quad (26)$$

$$\theta_k - g_2(D_2^{t-1}) - p_2 - \gamma_2[d_k - d_2^*]^+ < 0. \quad (27)$$

Therefore, given the data plans $\mathcal{P}_1 = (p_1, +\infty, 0)$ and $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$, the data demand dynamics is described by a sequence $\{(D_1^t, D_2^t)\}_{t=0}^{\infty}$ in $\mathcal{D} = \{(D_1, D_2) \in \mathbb{R}_+^2 | D_1 + D_2 \le D_{\max}\}$ generated by $D_1^t = h_{d_1}(D_1^{t-1}, D_2^{t-1})$ and $D_2^t = h_{d_2}(D_1^{t-1}, D_2^{t-1})$, where $h_{d_1}(D_1^{t-1}, D_2^{t-1})$ and $h_{d_2}(D_1^{t-1}, D_2^{t-1})$ are obtained by substituting (D_1^{t-1}, D_2^{t-1}) into (15)–(18).

Since an equilibrium point may not exist if the data plan $\mathcal{P}_2$ is unlimited or capped, we restrict the analysis in the remainder of this paper to the case that the plan $\mathcal{P}_2$ is usage-based (although an initial subscription fee p_2 may be charged) such that a unique equilibrium point is guaranteed to exist. Next, we provide a sufficient condition for the data demand dynamics $\{(D_1^t, D_2^t)\}_{t=0}^{\infty}$ to converge.

Proposition 5: For data plans $\mathcal{P}_1 = (p_1, +\infty, 0)$ and $\mathcal{P}_2 = (p_2, 0, \gamma_2)$ where $\gamma_2 > 0$, the data demand dynamics converges to the unique equilibrium point starting from any initial point $(D_1^0, D_2^0) \in \mathcal{D} = \{(D_1, D_2) \in \mathbb{R}_+^2 | D_1 + D_2 \leq D_{\max}\}$ if the following condition is satisfied:

$$\max_{(D_1,D_2)\in[0,d_{\max}]^2} \{g_1'(D_1), g_2'(D_2)\} < \left[K\cdot\left(\frac{d_{\max}^2}{2}+\frac{d_{\max}}{\gamma_2}[\theta_{\max}-p_1]^+ + \frac{d_{\max}}{\gamma_2}[\theta_{\max}-p_2]^+\right)\right]^{-1} \tag{28}$$

where $d_{\max}$ is the maximum individual demand, $\theta_{\max}$ is the maximum benefit derived from subscribing to the WSP's service and $K = \max_{(\theta,d)\in\mathcal{U}} f(\theta, d)$. □

Proof: See Appendix D.

We can obtain more specific conditions regarding the network capacities for the convergence of data demand dynamics by plugging $g_1(D_1) = \frac{D_1}{C_1}$ and $g_2(D_2) = \frac{D_2}{C_2}$ into (28). The result is similar to Corollary 1 and omitted for brevity. Note that (28) imposes a more stringent requirement on the congestion costs (e.g., the WSP needs to allocate larger capacities to the subscribers) than (7) does. However, (28) provides us with a similar insight that, if congest costs increase too rapidly, the data demand dynamics may exhibit oscillation or divergence. Another important observation from (28) is that the two data plans also affect the convergence. Specifically, given higher prices, it is easier for the congestion costs to satisfy the convergence condition. Intuitively, higher prices result in lower aggregate data demand. Therefore, there is less fluctuation in the data demand dynamics and the requirement on the congestion costs becomes less stringent.

B. WSP's Data Plan Decision

Following Section IV-B, we first write the the WSP's equilibrium revenues for the data plans $\mathcal{P}_1$ and $\mathcal{P}_2$ as

$$R_1 = \int_{y=\tilde{d}}^{d_{\max}} \int_{x=g_1(D_1^*)+p_1}^{\theta_{\max}} p_1 f(x,y)dxdy \tag{29}$$

$$R_2 = \int_{y=0}^{\tilde{d}} \int_{x=g_2(D_2^*)+p_2+\gamma_2 y}^{\theta_{\max}} (p_2+\gamma_2 y) f(x,y)dxdy \tag{30}$$

if $p_1 + g_1(D_1^*) > p_2 + g_2(D_2^*)$, and as

$$R_1 = \int_{y=0}^{d_{\max}} \int_{x=g_1(D_1^*)+p_1}^{\theta_{\max}} p_1 f(x,y)dxdy \quad \text{and} \quad R_2 = 0 \tag{31}$$

if $p_1 + g_1(D_1^*) \leq p_2 + g_2(D_2^*)$, where $\tilde{d}$ is given by $\tilde{d} = \frac{1}{\gamma_2}[p_1 - p_2 + g_1(D_1^*) - g_2(D_2^*)]$. The expressions of equilibrium revenues in (29)–(31) are even more complicated than (13) and hence, lose analytical tractability. As a consequence, we resort to numerical search to identify the optimal $\mathcal{P}_1 = (p_1, +\infty, 0)$ and $\mathcal{P}_2 = (p_2, 0, \gamma_2)$ maximizing $R_1 + R_2$.

C. WSP's Capacity Decision

It is mathematically challenging to analytically find the optimal capacities $\mathbf{C} = (C_1, C_2)$ to maximize the WSP's profit, since the optimal data plans can only be numerically found. Thus, as in Section IV-C, we find the WSP's optimal capacities through exhaustive search.

In the above analysis, we have considered that the two data plans $\mathcal{P}_1$ and $\mathcal{P}_2$ are offered by the same WSP. Nevertheless, in a wireless communications market, it is possible that these two plans are offered by two different WSPs competing against each other (i.e., duopoly market). The order of moves is almost the same as that described at the beginning of Section IV, with the exception that in the long-term and medium-term periods, each of the two WSPs decide their own network capacities and data plans, respectively. Specifically, for the long-term capacity decision, the two WSPs simultaneously and independently invest in the network capacities. Then, given the capacity decisions, the two WSPs play a noncooperative subgame in which they strategically make data plan decisions. Best-response dynamics can be applied to model the two WSPs' data plan decision process. That is, given its competitor's data plan, each WSP chooses an optimal data plan to selfishly maximize its revenue. In the short-term period, the users' subscription dynamics is unaffected and the same as that studied in Section V-A. Unfortunately, it is mathematically intractable to analyze the competition between the two WSPs, as explicitly expressing the optimal decisions of the two WSPs in response to each other's decision is not possible. With a simpler model, some (partial) analytical results regarding the competition between the WSPs are available in [1] and [8], whereas in this paper, we shall illustrate the WSP competition through numerical results.

VI. Numerical Results

In the numerical results, we assume that the congestion costs are given by $g_1(D_1) = \frac{D_1}{C_1}$ and $g_2(D_2) = \frac{D_2}{C_2}$, which capture the congestion externalities effects in time-sharing communications networks [6], [8]. For the ease of presentation, we consider uniformly distributed (θ, d), i.e., $f(\theta, d) = 1$ in $\mathcal{U} = \{(\theta, d) | 0 \leq \theta \leq 1, 0 \leq d \leq 1\}$. Note that our analysis also applies to other settings, provided that Assumptions 1–3 specified in Section III are satisfied.

A. Single Data Plan

First, we illustrate in Fig. 1 the oscillation and convergence of the data demand dynamics. The lower plot in Fig. 1 shows that the equilibrium data demand D_2^* decreases when the fixed subscription fee p_2 increases. Fig. 1 verifies that, even for the same data plan, different congestion cost functions may result in different convergence behaviors of the data demand dynamics. We plot the continuous-time data demand dynamics specified by (10) in Fig. 2. It can be seen that the continuous-time and discrete-time data demand dynamics converges to the same equilibrium point. The impacts of switching costs on the data demand dynamics are shown in Fig. 3, in which the upper plot indicates that switching costs may make the data demand dynamics

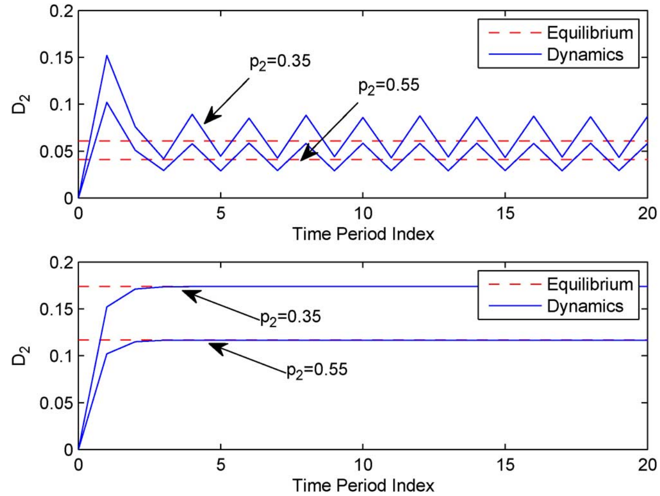


Fig. 1. Single data plan: oscillation and convergence of data demand dynamics. $\epsilon = 0.5$, $d_2^* = 0.5$, $\gamma_2 = 0.2$. $C_2 = \frac{1}{8}$ in upper plot and $C_2 = \frac{2}{3}$ in lower plot.

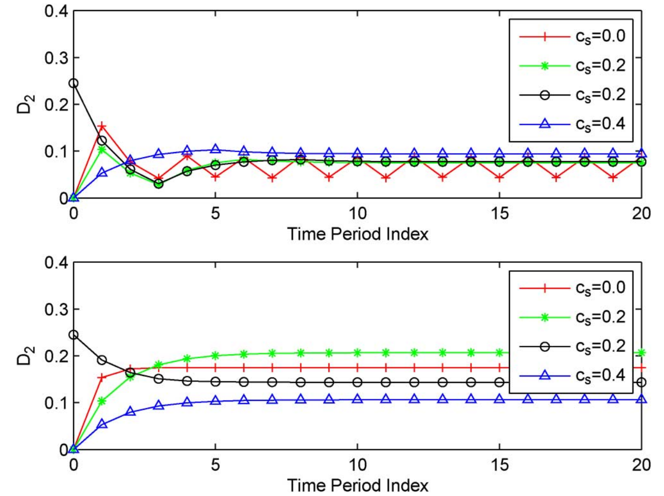


Fig. 3. Single data plan: oscillation and convergence of data demand dynamics with switching cost. $\epsilon = 0.5$, $p_2 = 0.35$, $d_2^* = 0.5$, $\gamma_2 = 0.2$. $C_2 = \frac{1}{8}$ in upper plot and $C_2 = \frac{2}{3}$ in lower plot.

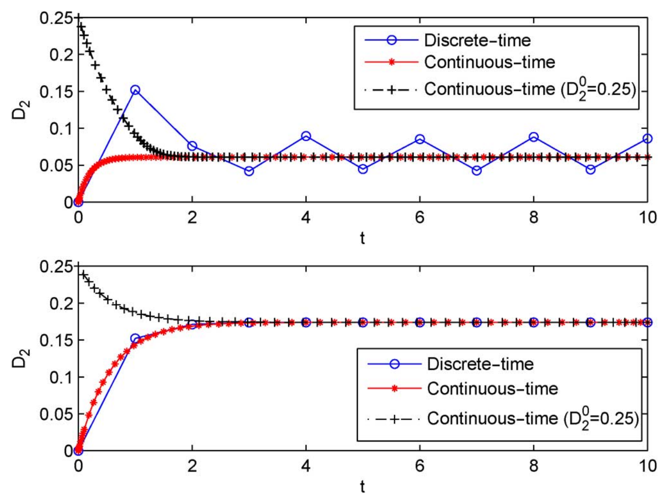


Fig. 2. Single data plan: comparison between discrete-time and continuous-time data demand dynamics. $\rho = 1$, $\epsilon = 0.5$, $d_2^* = 0.5$, $\gamma_2 = 0.2$. $C_2 = \frac{1}{8}$ in upper plot and $C_2 = \frac{2}{3}$ in lower plot.

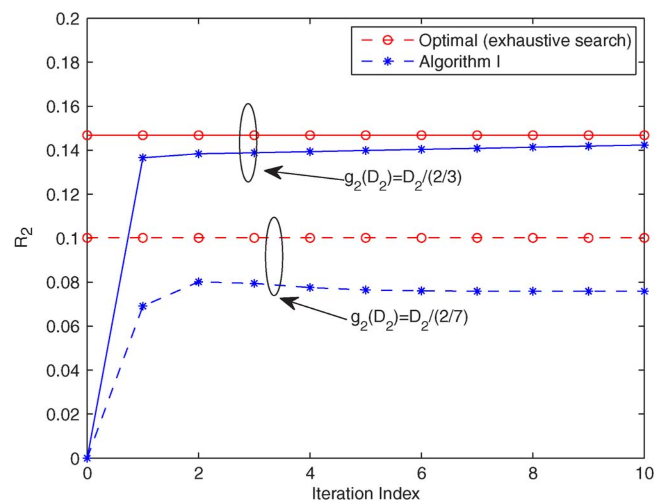


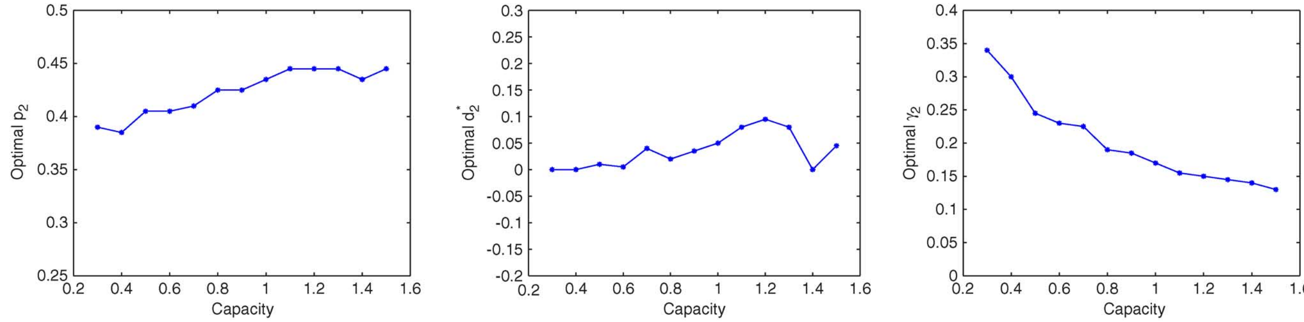
Fig. 4. Single data plan: comparison between the optimal revenue and that yielded by Algorithm I. $\epsilon = 0.5$.

converge even though the network capacity is not large enough. We explain this point by noting that, with switching costs, fewer users will not change their subscription decisions and hence the data demand dynamics converges under milder conditions. It can also be seen from the lower plot in Fig. 3 that there may exist multiple equilibrium data demand points and the equilibrium, to which the data demand dynamics converges, depends on the initial point. Next, we show in Fig. 4 that the proposed heuristic Algorithm I can yield a revenue close to the optimum, especially when the network capacity is large. Thus, Algorithm I may be used to find a suboptimal data plan if finding the optimal one is prohibitive. We also plot the optimal data plans[12] $\mathcal{P}_2 = \{p_2, d_2^*, \gamma_2\}$ in Fig. 5 under different network capacities. In Fig. 6, we show the WSP's profit versus its deployed network capacity under different capacity costs. It indicates that if the average capacity cost is smaller (e.g., the network's lifespan is long and/or the deployment cost is small), then the WSP needs to enlarge its investment in the network capacity. With a larger network capacity, the congestion effects will be reduced and the WSP can attract more users (hence, more revenue) to subscribe to its service.

[12]The optimal data plans are obtained by exhaustive search over all the possible data plans.

B. Two Data Plans

Convergence and oscillation of the data demand dynamics in a wireless market with two data plans are illustrated in Fig. 7. As intuitively expected and reflected in Proposition 5, a more stringent requirement on the congestion costs (i.e., the network capacities) is imposed to guarantee the convergence of the data demand dynamics with two data plans, compared to a market with only one data plan. Thus, even though a certain network capacity may guarantee the convergence of data demand dynamics with one data plan, it does not necessarily guarantee the convergence with two data plans. Next, we show in Fig. 8 the profits under various network capacities. To maximize the profit, the WSP needs to increase the network capacity for its capped data



(a) (b) (c)

Fig. 5. Single data plan: optimal data plan $\mathcal{P}_2 = (p_2, d_2^*, \gamma_2)$ versus capacity C_2. $\epsilon = 0.5$.

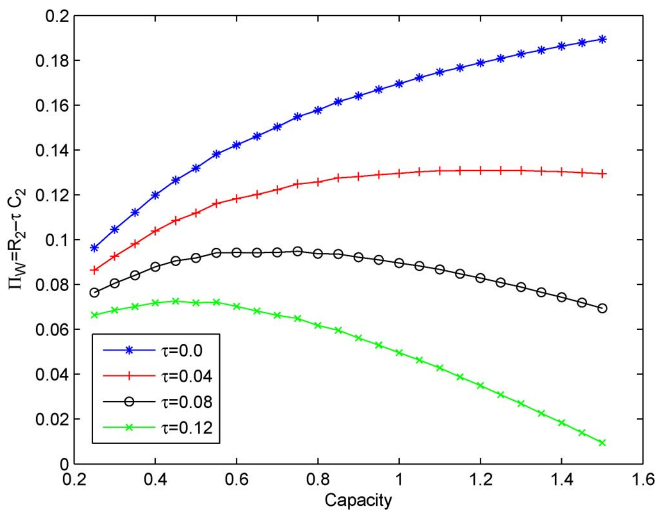


Fig. 6. Single data plan: optimal profit versus capacity. $\epsilon = 0.5$.

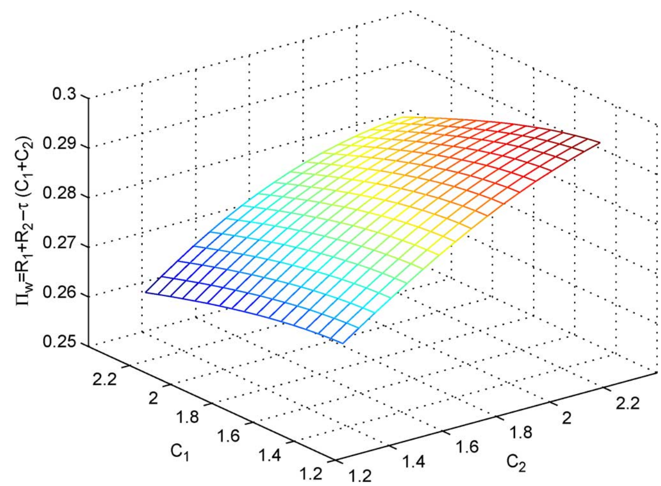


Fig. 8. Two data plans: optimal profit. $\epsilon = 0.5$, $\tau = 0.02$.

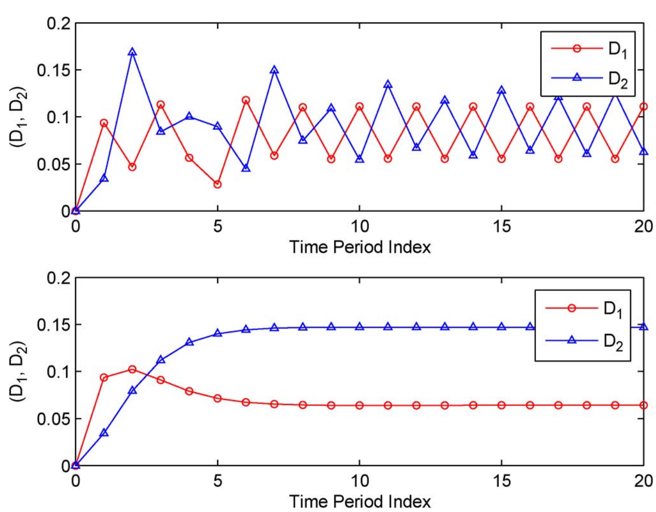


Fig. 7. Two data plans: oscillation and convergence of data demand dynamics. $\epsilon = 0.5, p_1 = 0.5, p_2 = 0.35, \gamma_2 = 0.3$. $C_1 = C_2 = \frac{1}{3}$ in upper plot and $C_1 = C_2 = 2.0$ in lower plot.

plan while reducing the network capacity for its unlimited data plan. This can be explained as follows: when an unlimited data plan is offered, subscribers with high data demand will cause excessive congestion costs for the other subscribers, reducing the profitability of the unlimited data plan. This also coincides with the current trend that some WSPs have discontinued unlimited data plans [20]. Finally, we show the competition between two WSPs in Fig. 9. It can be seen that if the two WSPs choose their optimal data plans independently in response to the competitor's data plan, then the competition will quickly lead to an equilibrium. This means that, given the long-term capacity investment, the WSPs only adjust their data plans a few times before reaching an equilibrium.[13] Fig. 9 also shows that if the capacity investment by the WSP offering the unlimited data plan decreases, the corresponding revenue will be reduced, whereas its competitor's revenue will significantly increase. This is because with a decreased capacity for the unlimited data plan, the resulting congestion cost will increase significantly (due to heavy users) and thus, many users will switch to the usage-based data plan. Note that the two WSPs also need to compete against each other by strategically choosing their long-term network capacities. The result is similar to Fig. 9 and hence, is omitted here for brevity.

VII. Conclusion

In this paper, we considered a wireless communications market where one monopolistic WSP serves a large number of users. The users' data demand dynamics, the WSP's data

[13]Similar results are also observed for other simulation settings, although in some (rare) cases the data plan competition between the two WSPs does not converge.

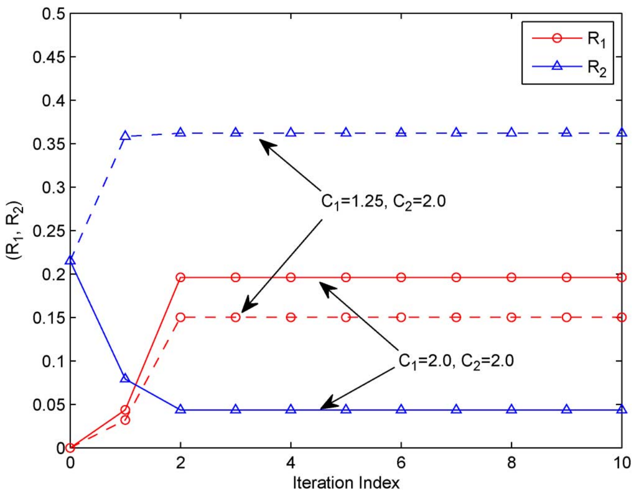


Fig. 9. Two data plans: iteration of revenues in a duopoly market. $\epsilon = 0.5$.

plan decision and network capacity decision were studied. In our analysis, the user heterogeneity in terms of their benefits and data demand, as well as the network congestion costs, were explicitly taken into consideration. For the data demand dynamics, we showed that: (1) the existence of an equilibrium data demand is independent of the congestion cost, although for certain data plans, there may not exist any equilibrium data demand if two data plans are both offered in the market; (2) in order to guarantee the convergence of data demand dynamics, the congestion costs should not increase too rapidly when the aggregate data demand increases, implying that the WSP needs to deploy a large network capacity to support the users' demand. We also proposed a heuristic algorithm that can achieve a close-to-optimal equilibrium revenue if only one data plan is offered by the WSP. For general cases, the WSP's data plan decision and network capacity decision were formalized and solved numerically to maximize the WSP's profit. Finally, we conducted extensive simulations to verify our analysis. Numerical results indicate that to maximize the profit, the WSP should increase the network capacity for its capped data plan while decreasing the network capacity for its unlimited data plan.

Appendix A
Proof of Proposition 1

To facilitate the proof, we first define an auxiliary function $\tilde{h}_2(D_2) = h_2(D_2) - D_2$ for $D \in [0, \infty)$, where $h_2(\cdot)$ is defined in (6). By Definition 1, D_2^* is an equilibrium point if and only if it is a root of $\tilde{h}_2(\cdot)$. Hence, it suffices to show that $\tilde{h}_2(\cdot)$ has a unique root on its domain.

Let $D_{\max} \geq D_{2,a} > D_{2,b} \geq 0$ be two arbitrarily-chosen real numbers. Then, it follows that

$$\begin{aligned}
&\tilde{h}_2(D_{2,a}) - \tilde{h}_2(D_{2,b}) \\
&\quad = h_2(D_{2,a}) - D_{2,a} - [h_2(D_{2,b}) - D_{2,b}] \\
&\quad = -\int_{y=0}^{d_{\max}} \int_{x=g_2(D_{2,b})+p_2+\gamma_2[y-d_2^*]^+}^{g_2(D_{2,a})+p_2+\gamma_2[y-d_2^*]^+} yf(x,y)dxdy \\
&\qquad - (D_{2,a} - D_{2,b}).
\end{aligned} \tag{32}$$

Since $g_2(\cdot)$ is nondecreasing in $[0, D_{\max}]$, we have $g_2(D_{2,a}) + p_2 + \gamma_2[y-d_2^*]^+$ is greater than or equal to $g_2(D_{2,b}) + p_2 + \gamma_2[y-d_2^*]^+$ and hence, $\int_{y=0}^{d_{\max}} \int_{x=g_2(D_{2,b})+p_2+\gamma_2[y-d_2^*]^+}^{g_2(D_{2,a})+p_2+\gamma_2[y-d_2^*]^+} yf(x,y)dxdy$ is nonnegative. Thus, it can be seen that $\tilde{h}_2(D_{2,a}) - \tilde{h}_2(D_{2,b}) \leq -(D_{2,a} - D_{2,b}) < 0$ for any $D_{\max} \geq D_{2,a} > D_{2,b} \geq 0$. That is, the auxiliary function $\tilde{h}_2(\cdot)$ is strictly decreasing in $[0, D_{\max}]$.

On the one hand, $\tilde{h}_2(0) = h_2(0) - 0 \geq 0$, and on the other hand, $\tilde{h}_2(D_{\max}) = h_2(D_{\max}) - D_{\max} \leq 0$ Since $\tilde{h}_2(\cdot)$ is continuous on $[0, D_{\max}]$, we see that $\tilde{h}_2(\cdot)$ has a unique root $D_2^* \in [0, D_{\max}]$, by applying the intermediate value theorem. This proves Proposition 1. ■

Appendix B
Proof of Proposition 3

We prove the convergence of the data demand dynamics based on contraction mapping theorem, which is formally stated as follows [26].

Definition 3 [26]: A mapping $\mathbf{T} : \mathcal{X} \to \mathcal{X}$, where $\mathcal{X}$ is a closed subset of $\mathbb{R}^n$, is called a contraction if there is a real number $\kappa \in [0, 1)$ such that

$$\|\mathbf{T}(x_1) - \mathbf{T}(x_2)\| \leq \kappa \cdot \|x_1 - x_2\|, \quad \forall x_1, x_2 \in \mathcal{X} \tag{33}$$

where $\|\cdot\|$ is some norm defined on $\mathcal{X}$.

In [26, Proposition 1.1, Ch. 3] shows an important property of a contraction mapping $\mathbf{T}$ that the update sequence generated by $x^{t+1} = \mathbf{T}(x^t)$, $t = 1, 2, \ldots$, converges to a (unique) fixed point x^* satisfying $\mathbf{T}(x^*) = x^*$ starting from any initial value $x^0 \in \mathcal{X}$. To prove Proposition 3, we shall show that the function $h_2(\cdot)$, defined in (6), is a contraction mapping on $[0, D_{\max}]$ with respect to the absolute value norm if (7) is satisfied.

Let $D_{2,a}$ and $D_{2,b}$ be two arbitrarily chosen real numbers such that $D_{\max} \geq D_{2,a} > D_{2,b} \geq 0$. Then, it can be shown that

$$\begin{aligned}
&|h_2(D_{2,a}) - h_2(D_{2,b})| \\
&\quad = \int_{y=0}^{d_{\max}} y \left\{ \int_{x=g_2(D_{2,b})+p_2+\gamma_2[y-d_2^*]^+}^{g_2(D_{2,a})+p_2+\gamma_2[y-d_2^*]^+} f(x,y)dx \right\} dy.
\end{aligned} \tag{34}$$

Denote $F(\theta|y) = \int_{x=-\infty}^{\theta} f(x,y)dx$. Thus, we can obtain the following inequalities:

$$\int_{x=g_2(D_{2,b})+p_2+\gamma_2[y-d_2^*]^+}^{g_2(D_{2,a})+p_2+\gamma_2[y-d_2^*]^+} f(x,y)dx \tag{35}$$

$$\begin{aligned}
&= F\left(g_2(D_{2,a}) + p_2 + \gamma_2 [y - d_2^*]^+ |y\right) \\
&\quad - F\left(g_2(D_{2,b}) + p_2 + \gamma_2 [y - d_2^*]^+ |y\right) && (36) \\
&= f(D_{2,\gamma}, y) \cdot g_2'(D_{2,\gamma}) \cdot (D_{2,a} - D_{2,b}) && (37) \\
&\leq K \cdot g_2'(D_{2,\gamma}) \cdot (D_{2,a} - D_{2,b}) && (38) \\
&\leq K \cdot \left[\max_{D_2 \in [0, D_{\max}]} g_2'(D_2)\right] \cdot (D_{2,a} - D_{2,b}) && (39)
\end{aligned}$$

where (37) follows from the intermediate value theorem and chain rule, $D_{2,\gamma}$ is a certain value in $[D_{2,b}, D_{2,a}]$, and $K = \max_{(\theta,d)\in\mathcal{U}} f(\theta, d)$. Then, by plugging the inequality (39) into (34), we have

$$\begin{aligned} &|h_2(D_{2,a}) - h_2(D_{2,b})| \\ &\leq \int_{y=0}^{d_{\max}} yK \left[\max_{D_2\in[0,D_{\max}]} g_2'(D_2)\right](D_{2,a} - D_{2,b})dy \\ &= \frac{d_{max}^2}{2} K \left[\max_{D_2\in[0,D_{\max}]} g'(D_2)\right] \cdot |D_{2,a} - D_{2,b}|. \end{aligned} \quad (40)$$

Therefore, if (7) is satisfied, then $\kappa = \frac{d_{max}^2}{2} K[\max_{D_2\in[0,D_{\max}]} g_2'(D_2)] \in [0,1)$ and $|h_2(D_{2,a}) - h_2(D_{2,b})| \leq \kappa|D_{2,a} - D_{2,b}|$, for any $D_{\max} \geq D_{2,a} > D_{2,b} \geq 0$. In other words, $h_2(\cdot)$ is a contraction mapping on $[0, D_{\max}]$ with respect to the absolute value norm. Thus, by applying [26, Proposition 1.1, Ch. 3], we see that the data demand dynamics converges if (7) is satisfied. Proposition 3 is therefore proved. ■

Appendix C
Proof of Proposition 4

To facilitate the proof, we first define two auxiliary functions $\tilde{h}_{d,1}(D_1, D_2) = h_{d,1}(D_1, D_2) - D_1$ and $\tilde{h}_{d,2}(D_1, D_2) = h_{d,2}(D_1, D_2) - D_2$ for $(D_1, D_2) \in \mathcal{D}$, where $h_{d,1}(D_1, D_2)$ and $h_{d,2}(D_1, D_2)$ are defined in (15)–(18). By Definition 2, (D_1^*, D_2^*) is an equilibrium point if and only if

$$\tilde{h}_{d,1}(D_1^*, D_2^*) = 0 \quad \text{and} \quad \tilde{h}_{d,2}(D_1^*, D_2^*) = 0. \quad (41)$$

Hence, it suffices to show that the equation set in (41) has a unique solution on its domain $\mathcal{D}$.

Let us first assume that $D_2 \in [0, D_{\max}]$ is a fixed number. Following Appendix A, we can show that, if $\gamma_2 > 0$ and $d_2^* = 0$, $\tilde{h}_{d,1}(D_1, D_2)$ is a strictly decreasing and continuous function of $D_1 \in [0, D_{\max} - D_2]$. Moreover, for any fixed value of $D_2 \in [0, D_{max}]$, we have $\tilde{h}_{d,1}(D_{\max} - D_2, D_2) \leq 0$ and $\tilde{h}_{d,1}(0, D_2) \geq 0$. Therefore, by applying the intermediate value theorem, it follows that $\tilde{h}_{d,1}(D_{\max} - D_2, D_2)$ has a unique root D_1^* given any fixed value of D_2. Thus, D_1^* can be expressed as a function in terms of D_2, and $\tilde{h}_{d,1}(D_1^*, D_2)$ and $\tilde{h}_{d,2}(D_1^*, D_2)$ can be rewritten in a compact form as $\tilde{h}_{d,1}(D_2)$ and $\tilde{h}_{d,2}(D_2)$, respectively. It can also be easily proved that D_1^* is a decreasing function of D_2 in $D_2 \in [0, D_{\max}]$. Next, we need to show that $\tilde{h}_{d,2}(D_2)$ has a unique root in $D_2 \in [0, D_{\max}]$ in order to prove Proposition 4.

Lemma 1: $h_t(D_2) = h_{d,2}(D_1^*, D_2) + D_1^*$ is decreasing in D_2.

Proof: Note that $h_{d,2}(D_1^*, D_2) + D_1^*$ is the sum data demand of WSP W_1 operating at its equilibrium data demand point and W_2 at the next period, when WSP W_2 currently has a data demand of D_2. Thus, $h_{d,2}(D_1^*, D_2) + D_1^*$ can be expressed as follows:

$$\begin{aligned} &h_{d,2}(D_1^*(D_2), D_2) + D_1^*(D_2) \\ &= D_{max} - \int_{y=0}^{\tilde{d}} \int_{x=0}^{g_2(D_2)+p_2+\gamma_2 y} yf(x,y)dxdy \\ &\quad - \int_{y=\tilde{d}}^{d_{\max}} \int_{x=0}^{g_1(D_1^*)+p_1} yf(x,y)dxdy \end{aligned} \quad (42)$$

where $\tilde{d} = \frac{1}{\gamma_2}[p_1 - p_2 + g_1(D_1^*) - g_2(D_2)]^+$, the second term and third term on the right-hand side of the equality represent the aggregate data demand of those users that do not subscribe to either WSP. Since $D_1^*(D_2)$ is increasing in $D_2 \in [0, D_{\max}]$ and $g_1(D_1)$ is increasing in D_1 in its domain, we see that $g_1(D_1^*) = g_1(D_1^*(D_2))$ is also increasing in $D_2 \in [0, D_{\max}]$. Therefore, (42) is decreasing in $D_2 \in [0, D_{\max}]$. This can also be intuitively expected. When the data demand of both WSPs increases, the congestion costs increase and hence fewer users will subscribe to the WSPs, which will in turn result in a decrease in the total data demand of these two WSPs. □

Recall that $D_1^*(D_2)$ is increasing in $D_2 \in [0, D_{\max}]$. Thus, following Lemma 1, it can be seen that $h_{d,2}(D_1^*(D_2), D_2) = h_t(D_2) - D_1^*(D_2)$ is a nonincreasing function of D_2 and $\tilde{h}_{d,2}(D_1^*, D_2) = \tilde{h}_{d,2}(D_2)$ is a strictly decreasing function of D_2 in $[0, D_{\max}]$. On the one hand, $h_{d,2}(D_{\max}) - D_{\max} \leq 0$ and, on the other hand, $h_{d,2}(0) - 0 \geq 0$. Thus, due to its continuity and strictly decreasing property, $\tilde{h}_{d,2}(D_1^*, D_2) = h_{d,2}(D_1^*(D_2), D_2) - D_2$ has a unique root in $D_2 \in [0, D_{\max}]$. This proves Proposition 4. ■

Appendix D
Proof of Proposition 5

First, define the mapping that specifies the data demand dynamics by $\mathbf{h}_d$

$$\mathbf{h}_d(D_1, D_2) = (h_{d,1}(D_1, D_2), h_{d,2}(D_1, D_2)) \quad (43)$$

where $h_{d,1}$ and $h_{d,2}$ are defined in (15)–(18). In order to establish the global convergence of the data demand dynamics, we shall show that the mapping $\mathbf{h}_d(\cdot)$ is a contraction on $\mathcal{D}$ with respect to a certain norm [26]. Unlike in a market with only one data plan offered, the mapping $\mathbf{h}_d$ is no longer a scalar function and hence the absolute value norm is not applicable. Instead, we apply L_1 norm and show that $\mathbf{h}_d(\cdot)$ is a contraction with respect to L_1 norm if (28) is satisfied.

Let $D_a = (D_{1,a}, D_{2,a}) \in \mathcal{D}$ and $D_b = (D_{1,b}, D_{2,b}) \in \mathcal{D}$ be two arbitrarily-chosen points such that $\tilde{d}_a \geq \tilde{d}_b \geq 0$, where $\tilde{d}_a = \frac{1}{\gamma_2}[p_1 - p_2 + g_1(D_{1,a}) - g_2(D_{2,a})]^+$ and $\tilde{d}_b = \frac{1}{\gamma_2}[p_1 - p_2 + g_1(D_{1,b}) - g_2(D_{2,b})]^+$. By the definition of L_1 norm, we have

$$\begin{aligned} &\|\mathbf{h}_d(D_{1,a}, D_{2,a}) - \mathbf{h}_d(D_{1,b}, D_{2,b})\|_1 \\ &= |h_{d,1}(D_{1,a}, D_{2,a}) - h_{d,1}(D_{1,b}, D_{2,b})| \\ &\quad + |h_{d,2}(D_{1,a}, D_{2,a}) - h_{d,2}(D_{1,b}, D_{2,b})|. \end{aligned} \quad (44)$$

Note that the term $|h_{d,1}(D_{1,a},D_{2,a}) - h_{d,1}(D_{1,b},D_{2,b})|$ can be expanded and rewritten as

$$|h_{d,1}(D_{1,a},D_{2,a}) - h_{d,1}(D_{1,b},D_{2,b})| \tag{45}$$

$$= \left| \int\limits_{y=\tilde{d}_a}^{d_{\max}} \int\limits_{x=g_1(D_{1,a})+p_1}^{\theta_{\max}} yf(x,y)dxdy - \int\limits_{y=\tilde{d}_b}^{\tilde{d}_a} \int\limits_{x=g_1(D_{1,b})+p_1}^{\theta_{\max}} yf(x,y)dxdy - \int\limits_{y=\tilde{d}_a}^{d_{\max}} \int\limits_{x=g_1(D_{1,b})+p_1}^{\theta_{\max}} yf(x,y)dxdy \right| \tag{46}$$

$$= \left| \int\limits_{y=\tilde{d}_a}^{d_{\max}} \int\limits_{x=g_1(D_{1,a})+p_1}^{g_1(D_{1,b})+p_1} yf(x,y)dxdy - \int\limits_{y=\tilde{d}_b}^{\tilde{d}_a} \int\limits_{x=g_1(D_{1,b})+p_1}^{\theta_{\max}} yf(x,y)dxdy \right|. \tag{47}$$

Denote $K = \max_{(\theta,d)\in\mathcal{U}} f(\theta,d)$. Next, we show that the following inequalities can be established:

$$\left| \int\limits_{y=\tilde{d}_a}^{d_{\max}} \int\limits_{x=g_1(D_{1,a})+p_1}^{g_1(D_{1,b})+p_1} yf(x,y)dxdy - \int\limits_{y=\tilde{d}_b}^{\tilde{d}_a} \int\limits_{x=g_1(D_{1,b})+p_1}^{\theta_{\max}} yf(x,y)dxdy \right| \tag{48}$$

$$\leq K\cdot|g_1(D_{1,a}) - g_1(D_{1,b})| \frac{d_{\max}^2 - \min\left\{d_{\max}^2, \tilde{d}_a^2\right\}}{2} + K\cdot[\theta_{\max} - (p_1 + g_1(D_{1,b}))]^+ \frac{\tilde{d}_a^2 - \tilde{d}_b^2}{2} \tag{49}$$

$$\leq K\cdot g_1'(D_{1,c})\cdot|D_{1,a} - D_{1,b}|\cdot\frac{d_{\max}^2}{2} + K\cdot[\theta_{\max} - p_1]^+\cdot\frac{(\tilde{d}_a + \tilde{d}_b)(\tilde{d}_a - \tilde{d}_b)}{2} \tag{50}$$

$$\leq K\cdot g_1'(D_{1,c})\cdot|D_{1,a} - D_{1,b}|\cdot\frac{d_{\max}^2}{2} + K\cdot[\theta_{\max} - p_1]^+\cdot d_{\max}\cdot[\tilde{d}_a - \tilde{d}_b]^+ \tag{51}$$

$$\leq K\cdot g_1'(D_{1,c})\cdot|D_{1,a} - D_{1,b}|\cdot\frac{d_{\max}^2}{2} + K\cdot[\theta_{\max} - p_1]^+\cdot\frac{d_{\max}}{\gamma_2}\cdot|g_1(D_{1,a}) - g_1(D_{1,b})| - K\cdot[\theta_{\max} - p_1]^+\cdot\frac{d_{\max}}{\gamma_2}|g_2(D_{2,a}) - g_2(D_{2,b})| \tag{52}$$

where $D_{1,c}$ is a number between $D_{1,a}$ and $D_{1,b}$, (50) follows (49) based on the intermediate value theorem, and (52) is due to the fact that $[[x_1]^+ - [x_2]^+]^+ \leq [x_1 - x_2]^+$. Thus, by combining (45)–(52), we see that (45) is less than or equal to (52). Similarly, $|h_{d,2}(D_{1,a},D_{2,a}) - h_{d,2}(D_{1,b},D_{2,b})|$ can also be upper bounded by

$$Kg_2'(D_{2,c'})|D_{2,a} - D_{2,b}|\cdot\frac{d_{\max}^2}{2} + K[\theta_{\max} - p_2]^+\cdot\frac{d_{\max}}{\gamma_2}|g_1(D_{1,a}) - g_1(D_{1,b})| - K[\theta_{\max} - p_2]^+\cdot\frac{d_{\max}}{\gamma_2}|g_2(D_{2,a}) - g_2(D_{2,b})|. \tag{53}$$

Thus, we can derive

$$\|\mathbf{h}_d(D_{1,a},D_{2,a}) - \mathbf{h}_d(D_{1,b},D_{2,b})\|_1 \leq K\left(\frac{d_{\max}^2}{2} + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_1]^+ + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_2]^+\right)\cdot g_1'(D_{1,c})\cdot|D_{1,a} - D_{1,b}| + K\left(\frac{d_{\max}^2}{2} + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_1]^+ + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_2]^+\right)\cdot g_2'(D_{2,c'})\cdot|D_{2,a} - D_{2,b}|. \tag{54}$$

Therefore, if (28) is satisfied, the mapping $\mathbf{h}_d(\cdot)$ is a contraction on $\mathcal{D}$ with respect to L_1 norm with a modulus $\kappa_d \in [0,1)$, where $\kappa_d = \max_{(D_1,D_2)\in[0,d_{max}]^2}\{g_1'(D_1), g_2'(D_2)\}\cdot K\cdot\left(\frac{d_{\max}^2}{2} + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_1]^+ + \frac{d_{\max}}{\gamma_2}[\theta_{\max} - p_2]^+\right)$, and the data demand dynamics specified by (15)–(18) converges, regardless of the initial points, to the unique equilibrium point. This proves Proposition 5. ■

REFERENCES

[1] S. Ren, J. Park, and M. van der Schaar, "User subscription dynamics and revenue maximization in communication markets," in *Proc. IEEE Infocom*, Apr. 2011.

[2] A. Zemlianov and G. de Veciana, "Cooperation and decision making in wireless multiprovider setting," in *Proc. IEEE Infocom*, Mar. 2005.

[3] L. He and J. Walrand, "Pricing and revenue sharing strategies for Internet service providers," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 942–951, May 2006.

[4] Y. Jin, S. Sen, R. Guerin, K. Hosanagar, and Z.-L. Zhang, "Dynamics of competition between incumbent and emerging network technologies," in *Proc. NetEcon*, Seattle, WA, Aug. 22, 2008.

[5] M. Manshaei, J. Freudiger, M. Felegyhazi, P. Marbach, and J. P. Hubaux, "On wireless social community networks," in *Proc. IEEE Infocom*, Apr. 2008.

[6] C. K. Chau, Q. Wang, and D. M. Chiu, "On the viability of Paris metro pricing for communication and service networks," in *Proc. IEEE Infocom*, Mar. 2010.

[7] N. Shetty, S. Parekh, and J. Walrand, "Economics of femtocells," in *Proc. IEEE Globecom*, Dec. 2009.

[8] N. Shetty, G. Schwartz, and J. Walrand, "Internet QoS and regulations," *IEEE/ACM Trans. Netw.*, vol. 18, no. 6, pp. 1725–1737, Dec. 2010.

[9] G. Kesidis, A. Das, and G. de Veciana, "On flat-rate and usage-based pricing for tiered commodity internet services," in *Proc. CISS*, 2008.

[10] J. Musacchio and D. Kim, "Network platform competition in a two-sided market: Implications to the net neutrality issue," in *Proc. TPRC: Conf. Commun., Inf., Internet Policy*, Sep. 2009.

[11] P. Hande, M. Chiang, A. R. Calderbank, and S. Rangan, "Network pricing and rate allocation with content provider participation," in *Proc. IEEE Infocom*, Apr. 2009.

[12] T. Basar and R. Srikant, "Revenue-maximizing pricing and capacity expansion in a many-users regime," in *Proc. IEEE Infocom*, Jun. 2002.

[13] L. Jiang, S. Parekh, and J. Walrand, "Time-dependent network pricing and bandwidth trading," in *Proc. IEEE Int. Workshop Bandwidth on Demand (BoD)*, 2008.

[14] D. Acemoglu and A. Ozdaglar, "Competition and efficiency in congested markets," *Math. Oper. Res.*, vol. 32, no. 1, pp. 1–31, 2007.

[15] Z. Kun, D. Niyato, and P. Wang, "Optimal bandwidth allocation with dynamic service selection in heterogeneous wireless networks," in *Proc. IEEE Globecom*, Dec. 2010, to be published.

[16] S. Shakkottai, E. Altman, and A. Kumar, "Multihoming of users to access points in WLANs: A population game perspective," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 6, pp. 1207–1215, Aug. 2007.

[17] V. Krishnamurthy, "Decentralized spectrum access amongst cognitive agents—An interacting multivariate global games approach," *IEEE Trans. Signal Process.*, vol. 57, no. 10, pp. 3999–4013, Oct. 2009.

[18] R. Gibbens, R. Mason, and R. Steinberg, "Internet service classes under competition," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 2, pp. 2490–2498, Feb. 2000.

[19] R. Jain, T. Mullen, and R. Hausman, "Analysis of Paris Metro pricing strategy for QoS with a single service provider," in *IEEE/IFIP Int. Workshop Quality of Serv.*, 2001.

[20] AT&T Wireless Data Plan [Online]. Available: http://www.wireless.att.com

[21] HSPA to LTE-Advanced: 3GPP Broadband Evolution to IMT-Advanced Rysavy Res. 3G Amer., Sep. 2009.

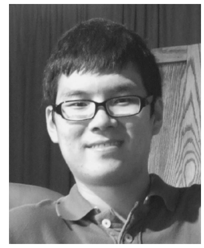
[22] "Mobile Users Wary of Usage-Based Pricing: Survey," Reuters, Nov. 18, 2010 [Online]. Available: http://www.reuters.com/article/idUSN1827905920101118

[23] R. Johari, G. Y. Weintraub, and B. Van Roy, "Investment and market structure in industries with congestion," *Oper. Res.*, vol. 58, no. 5, pp. 1303–1317, 2010.

[24] G. W. Evans and S. Honkapohja, *Learning and Expectations in Macroeconomics*. Princeton, NJ: Princeton Univ. Press, 2001.

[25] J. R. Munkres, *Elements of Algebraic Topology*. New York: Perseus, 1993.

[26] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA: Athena Scientific, 1989.



Shaolei Ren (S'06) received the B.E. and M.Phil. degrees, both in electronic engineering, from Tsinghua University, in July 2006 and the Hong Kong University of Science and Technology, August 2008, respectively.

He is currently pursuing the Ph.D. degree in electrical engineering at the University of California, Los Angeles (UCLA). Since September 2009, he has been affiliated with the Electrical Engineering Department, UCLA, as a Research Assistant. His research interests include signal processing, energy-efficient computing, social media, and network economics.

Mihaela van der Schaar (F'10) is the Chancellor's Professor in the Electrical Engineering Department, University of California, Los Angeles. Her research interests include multimedia systems, networking, communication, and processing, dynamic multiuser networks and system designs, online learning, network economics, and game theory. She holds 33 granted US patents.

Prof. van der Schaar is a Distinguished Lecturer of the Communications Society for 2011–2012, the Editor in Chief of the IEEE TRANSACTIONS ON MULTIMEDIA and a member of the Editorial Board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING. She received an NSF CAREER Award (2004), the Best Paper Award from the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (2005), the Okawa Foundation Award (2006), the IBM Faculty Award (2005, 2007, 2008), and the Most Cited Paper Award from EURASIP: Image Communications Journal (2006). She was formerly an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, SIGNAL PROCESSING LETTERS, CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and SIGNAL PROCESSING MAGAZINE. . She received three ISO awards for her contributions to the MPEG video compression and streaming international standardization activities.