The Role of Data Cap in Optimal Two-part Network Pricing

Xin Wang
University of Science and Technology of China
yixinxa@mail.ustc.edu.cn

Richard T. B. Ma
School of Computing, National University of Singapore
tbma@comp.nus.edu.sg

Yinlong Xu
University of Science and Technology of China
ylxu@ustc.edu.cn

ABSTRACT

Internet services are traditionally priced at flat rates; however, many Internet service providers (ISPs) have recently shifted towards two-part tariffs where a data cap is imposed to restrain data demand from heavy users and usage over the data cap is charged based on a per-unit fee. Although the two-part tariff could generally increase the revenue for ISPs and has been supported by the FCC chairman, the role of data cap and its revenue-optimal and welfare-optimal pricing structures are not well understood.

In this paper, we study the impact of data cap on the optimal two-part pricing schemes for congestion-prone service markets, e.g., broadband or cloud services. We model users' demand and preferences over pricing and congestion alternatives and derive the market share and congestion of service providers under a market equilibrium. Based on the equilibrium model, we characterize the two-part structures of the revenue-optimal and welfare-optimal pricing schemes. Our results reveal that 1) the data cap provides a mechanism for ISPs to transition from flat-rate to pay-as-you-go type of schemes, 2) with growing data demand and network capacity, the revenue-optimal pricing moves towards usage-based schemes with diminishing data caps, and 3) the structure of the welfare-optimal tariff comprises lower fees and data cap than those of the revenue-optimal counterpart, suggesting that regulators might want to promote usage-based pricing but regulate the per-unit fees. Our results could help providers design revenue-optimal pricing schemes and guide regulatory authorities to legislate desirable regulations.

Categories and Subject Descriptors

C.2.5 [**Computer-Communication Networks**]: Local and Wide-Area Networks—*Internet*; C.4 [**Performance of Systems**]: Modeling techniques

Keywords

Data Cap; Two-part Tariff; Revenue-optimal Pricing; Welfare-optimal Pricing

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741129.

1. INTRODUCTION

Traditionally, Internet service providers (ISPs) use flat-rate pricing [19] for network services, where users pay fixed monthly fees for unlimited data usage. Flat-rate pricing was widely adopted because it was easy for ISPs to implement and was preferred by users for its simplicity. However, with the rapid development and growing popularity of data intensity services, e.g., online video streaming and cloud-based applications, the Internet traffic keeps growing more than 50% per annum [13], which exposes some disadvantages of the flat-rate scheme. Because flat-rate does not count the users' data usage, "bandwidth hogs" [18] consume an unfair share of capacity and are subsidized by normal users, and ISPs cannot generate enough revenue to recoup their costs, especially for the mobile providers. As a consequence, most mobile LTE providers [17] and even broadband ISPs, e.g., Verizon [23] and AT&T [25], start to introduce a data cap and adopt a two-part tariff structure, a combination of flat-rate and usage-based pricing. Under such a two-part scheme, additional charges are imposed if a user's data demand exceeds the data cap and the exceeded amount is charged based on a per-unit fee.

Although prior work [5, 6, 8, 18] has shown that data-capped schemes could help ISPs generate higher revenue than that under the flat-rate pricing and the FCC chairman has recently backed usage-based pricing for broadband to penalize heavy Internet users [22], little is known about 1) the data cap's role and impact on a provider's optimal pricing structure, 2) the optimal two-part pricing structure and its dynamics under varying system parameters, e.g., the users' data values and demand, the capacity of providers and market competition, and 3) potential regulations on two-part pricing for protecting social welfare from monopoly providers. In this paper, we focus on a generic congestion-prone service market, e.g., mobile, broadband or cloud services, and study the data cap under two-part pricing schemes. Unlike physical commodities, the quality of network service is intricately influenced by a negative network effect (or network externality): the more users access the service simultaneously, the worse performance it provides. We model this service congestion as a function of providers' capacity and their data load. We characterize users by their demand and values on data usage, and analyze the market shares and congestion levels of the providers under varying pricing and market structures. Based on our model, we analyze the effect of data cap on the provider's optimal pricing structure and the resulting congestion and revenue. We also analyze and compare the revenue-optimal and welfare-optimal two-part schemes under varying system environments, and derive regulatory implications. Our main contributions and findings include the following.

- We model users' optimal data usages and preferences over various pricing schemes and exogenous levels of congestion.

We characterize the existence and uniqueness (Theorem 1) of a market equilibrium and show its monotonic dynamics (Theorem 2) under varying pricing and market structures.

- We analyze the impact of data cap on a provider's optimal pricing structure (Theorem 3) and find that data cap plays a transitional role between pay-as-you-go and flat-rate pricing and could increase the provider's revenue (Corollary 3).
- We characterize the dynamics of revenue-optimal two-part pricing (Theorem 4) under varying demand and values of users and capacities of providers. We find that with growing demand of users and capacity of providers, the structure of revenue-optimal solution moves closer to pay-as-you-go pricing with diminishing data cap, which provides a smooth transition from flat-rate to usage-based schemes. Although market competition drives pricing "cheaper", it does not necessarily change the structure of an optimal two-part tariff.
- We characterize the dynamics of welfare-optimal two-part pricing (Theorem 6) and find that welfare-optimal pricing imposes stricter data cap but lower fees than its revenue-optimal counterpart (Theorem 5). Our result implies that, to protect social welfare, regulators might want to encourage the use of two-part pricing with a limited data cap, while regulating the per-unit fee of the usage-based component.

We believe that our work provides new insights into the role of data cap in the optimal structure of two-part tariff. Our results could help service providers design revenue-optimal pricing schemes and guide regulatory authorities to legislate desirable regulations. The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 models the behavior of users and characterizes the market equilibrium under providers' two-part pricing schemes. We analyze the impact of data cap and the revenue-optimal pricing in Section 4, and study the welfare-optimal pricing in Section 5, respectively. We draw some conclusions in Section 6.

2. RELATED WORK

As the demarcation between the lump-sum and usage-based fees, data cap plays a crucial role in the two-part tariff structures. From the perspective of economic theory, prior work [5–8, 18] demonstrated that data-capped schemes can increase providers' revenue compared with the traditional flat-rate pricing [19]. Odlyzko *et al.* [20] showed that ISPs could also reduce network congestion by imposing data caps. Our analysis and results also confirm these observations. Furthermore, we focus on understanding the impact of data cap on the structures of revenue-optimal and welfare-optimal tariffs. The impact and optimal design of data cap have been empirically studied. In a qualitative study of households users under bandwidth caps, Chetty *et al.* [4] studied how the uncertainties of user types and demand would impact the setting of data cap and operator's revenue, and proposed new tools to help users manage their caps. Poularakis *et al.* [12] proposed a framework to calculate the optimal data caps and empirically evaluated the gains of ISPs when they adopt data caps based on traffic datasets. Unlike these efforts, our work adopts an analytical approach to characterize the desirable data cap and the optimal structure of the two-part pricing schemes.

More generally, there have been several works that study the usage-based Internet pricing. Hande *et al.* [10] characterized the economic loss due to ISPs' inability or unwillingness to price broadband access based on the time of use. Li *et al.* [14] studied the optimal price differentiation under complete and incomplete information. Basar *et al.* [1] and our related work [27] devised a revenue-maximizing pricing under varying user scale and network capacity. Shen *et al.* [24] investigated optimal nonlinear pricing policy design for a monopolistic service provider and showed that the introduction of nonlinear pricing provides a large profit improvement over linear pricing. In this paper, we focus on the two-part pricing. Besides optimizing the revenue from the provider's perspective, we also look into the welfare-optimal solution, through which we derive regulatory implications.

From a modeling perspective, Chander [2], Reitman [21], Ma [15] and our work all consider the service market with congestion externalities. Chander [2] studied the quality differentiation strategy of a monopoly provider and Reitman [21] studied a multi-provider price competition. Both of them modeled the market as a continuum of non-atomic users, each of which is characterized by a quality-sensitivity parameter. However, this one-dimensional model only applies for flat-rate pricing and the distribution of users was often assumed to be uniform for analytical tractability. To faithfully characterize the utility of users under two-part tiered pricing, we establish a novel two-dimensional model that describes users by their data demand and valuation on data usage. Furthermore, we analyze a class of distributions, including the uniform distribution, to understand the impact of user demand and value on the optimal pricing structures of the providers. Ma [15] also considered a two-dimensional user model; however, the author only focused on the pay-as-you-go pricing, a special case of the two-part tariff structure studied in this paper.

3. MODEL

3.1 Model of Users and Their Data Demand

We model each user by two orthogonal characteristics: her average value of per-unit data usage v and desirable data demand u. The user's data demand is measured by what she is billed, e.g., the number of bits transmitted or the amount of time being online.

We denote q as the congestion level of an ISP. Given the network congestion q, we denote $\rho(u, q)$ as the user's achievable demand.

Assumption 1. *$\rho(u,q)\colon \mathbb{R}_+ \times \mathbb{R}_+ \mapsto \mathbb{R}_+$ is a continuous function, increasing in u and decreasing in q. It has an upper-bound $\rho(u,0) = u$ and satisfies $\lim_{q\to+\infty} \rho(u,q) = 0$.*

Assumption 1 states that a user's achievable data demand equals its desirable demand u under no congestion and decreases monotonically when the network congestion q becomes more severe.

Beside network congestion, the user's actual data demand also depends on her ISP's pricing. We consider an ISP that adopts a two-part tiered pricing structure $\theta = (g, f, p)$, where g, f and p denote a data cap, a lump-sum service fee and a per-unit usage fee, respectively. Under this scheme, we denote $t(y, \theta)$ as the user's charge when y units of data are consumed, defined as

$$t(y,\theta) \triangleq f + p(y-g)^+.$$

Intuitively, if a user's usage is below the data cap g, the ISP only collects the lump-sum fee f; otherwise, extra charges are imposed on the usage above the data cap with the per-unit usage fee of p. This two-part structure is also a generalization of flat-rate, i.e., $g = +\infty$, and pay-as-you-go [15] pricing, i.e., $f = 0$ and $g = 0$.

We denote $\pi(y)$ as the user's utility when she consumes y units of data, defined by $\pi(y) \triangleq vy - t(y,\theta)$, i.e., the total traffic value minus the charge. We assume that users determine their optimal data demands that maximize their utilities. In other words, each user tries to solve the following optimization problem:

$$\begin{aligned} \text{Maximize} \quad & \pi(y) = vy - t(y,\theta) \\ \text{subject to} \quad & 0 \le y \le \rho(u,q), \end{aligned} \tag{1}$$

where a user's actual data demand is constrained by the achievable demand $\rho(u,q)$ under the network congestion q. We define $\phi = (u,v)$ as the type of the user and denote $y^*(\phi,\theta,q)$ as its optimal demand under ISP's pricing scheme θ and congestion q. Due to the space limitation, proofs of some lemmas, theorems and corollaries are omitted, but are available in the technical report [26].

Lemma 1. *Given an ISP with pricing scheme θ and congestion q, a user of type ϕ has a unique solution y^* that maximizes her utility. This optimal demand satisfies*

$$y^*(\phi,\theta,q) = \rho(u,q) - \big[\rho(u,q) - g\big]^+ \mathbf{1}_{\{v<p\}} \tag{2}$$

and is non-increasing in p and q and non-decreasing in u, v and g.

Lemma 1 states that if a user's achievable demand $\rho(u,q)$ is beyond the data cap g and her value v is lower than the extra per-unit usage fee p, she would avoid consuming extra usage above g which would result in reducing her utility, and thus the optimal data demand equals the data cap, i.e., $y^* = g$; otherwise, the optimal data demand equals her achievable demand, i.e., $y^* = \rho(u,q)$. Lemma 1 also intuitively states that the optimal demand would increase if the user's value v or desirable demand u increases, or the provider's per-unit fee p or data cap g or congestion q alleviates.

3.2 Users' Preferences over Providers

We consider a market that comprises of a set $\mathcal{N}$ of providers. We denote $\boldsymbol{\theta} = (\theta_i : i \in \mathcal{N})$ and $\mathbf{q} = (q_i : i \in \mathcal{N})$ as the pricing strategy and congestion vectors of the providers. We define $y_i^*(\phi) \triangleq y^*(\phi,\theta_i,q_i)$ as the optimal demand of user type ϕ when it chooses provider i. For any two providers $i,j \in \mathcal{N}$, we denote $i \succ_\phi j$ if users of type ϕ prefer i over j. We denote Φ_i as provider i's market share, i.e., the set of user types that choose to use i, defined as follows.

Definition 1. *We denote $\pi_i(y)$ as the user's utility function when using provider i. For any $i,j \in \mathcal{N}$, $i \succ_\phi j$ if and only if 1) $\pi_i(y_i^*(\phi)) > \pi_j(y_j^*(\phi))$ or 2) $\pi_i(y_i^*(\phi)) = \pi_j(y_j^*(\phi))$ and i is chosen over j by the user based on any arbitrary tie-breaking condition. Thus, the market share of provider i is defined as*

$$\Phi_i(\boldsymbol{\theta},\mathbf{q}) = \{\phi : i \succ_\phi j, \forall j \in \mathcal{N}\backslash\{i\}\}.$$

Definition 1 assumes that users would choose the provider that induces the highest utility under their optimal data demand. However, a user's best provider might still induce negative utility. Thus, we allow users not to use any of the providers if they all induce negative utility as follows.

Assumption 2. *There exists a dummy provider $0 \in \mathcal{N}$ with fees $f_0 = p_0 = 0$ and congestion $q_0 = +\infty$.*

Under Assumption 2, users can choose the dummy provider to obtain zero utility and $\Phi_0(\boldsymbol{\theta},\mathbf{q})$ conveniently defines the set of users that do not use any of the real providers. Next, we show how providers' market shares $\Phi_i(\boldsymbol{\theta},\mathbf{q})$ $(i \in \mathcal{N})$ vary when the set of competing providers $\mathcal{N}$ or the pricing strategies $\boldsymbol{\theta}$ change.

Lemma 2. *For a set $\mathcal{N}$ of providers, if two pricing strategies $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}$ satisfy $\hat{g}_i \ge g_i, \hat{f}_i \le f_i, \hat{p}_i \le p_i$ for some $i \in \mathcal{N}$ and $\hat{\theta}_j = \theta_j$ for all $j \ne i$, we have*

$$\Phi_i(\boldsymbol{\theta},\mathbf{q}) \subseteq \Phi_i(\hat{\boldsymbol{\theta}},\mathbf{q}) \quad \text{and} \quad \Phi_j(\boldsymbol{\theta},\mathbf{q}) \supseteq \Phi_j(\hat{\boldsymbol{\theta}},\mathbf{q}),\ \forall\, j \ne i.$$

For two sets $\mathcal{N}$ and $\mathcal{N}'$ of providers, if $\mathcal{N} \subseteq \mathcal{N}'$ and $(\theta_i',q_i') = (\theta_i,q_i)$ for all $i \in \mathcal{N}$, we have $\Phi_i(\boldsymbol{\theta}',\mathbf{q}') \subseteq \Phi_i(\boldsymbol{\theta},\mathbf{q})$, $\forall\, i \in \mathcal{N}$.

Lemma 2 states that under fixed levels of congestion $\mathbf{q}$, the market share of a provider would increase if the provider reduces its fees f or p, or raises its data cap g, unilaterally. Meanwhile, the market share of any other provider will decrease. It implies that monopolistic providers could use fees and data cap to trade off its market share and revenue; while, oligopolistic providers could compete for market shares by decreasing their fees and increasing data caps. Lemma 2 also implies that bringing new providers into the market will intensify market competition and existing providers' market shares will decrease, because some of their users might switch to the new providers.

Lemma 2 holds under the condition of fixed network congestion. However, a provider's congestion level depends on its market share and the data usage of its users. We discuss the dynamics of network congestion and equilibrium in the next subsection.

3.3 Network Congestion and Equilibrium

We denote U and V as the maximum desirable data demand and maximum per-unit data value of all users. Thus, the domain of users is defined as $\Phi = [0,U]\times[0,V]$. We model the set of all users by the measure space (Φ,μ), where μ denotes a product measure

$$\mu(E_1 \times E_2) = \mu_u(E_1) \times \mu_v(E_2), \quad \forall\, E_1 \subseteq [0,U], E_2 \subseteq [0,V],$$

where μ_u and μ_v are two continuous measures, defined by

$$\begin{aligned} &\mu_u\left((u_1,u_2]\right) = F_u(u_2) - F_u(u_1), \forall u_1 \le u_2, \text{ and} \\ &\mu_v\left((v_1,v_2]\right) = F_v(v_2) - F_v(v_1), \forall v_1 \le v_2, \end{aligned}$$

for some non-decreasing distribution functions F_u and F_v.

Based on the distribution of users, we denote d_i as provider i's data load, i.e., the aggregate data demand of users of i, defined by

$$d_i = D(\Phi_i(\boldsymbol{\theta},\mathbf{q});\theta_i,q_i) \triangleq \int_{\Phi_i(\boldsymbol{\theta},\mathbf{q})} y_i^*(\phi,\theta_i,q_i)d\mu \tag{3}$$

On the one hand, given congestion levels $\mathbf{q}$, provider i has an induced data load $d_i = D(\Phi_i;\theta_i,q_i)$. On the other hand, the provider's congestion level q_i is influenced by its data load d_i. We denote c_i as provider i's capacity and model its congestion q_i as a function $q_i = Q_i(d_i,c_i)$ of its data load d_i and capacity c_i.

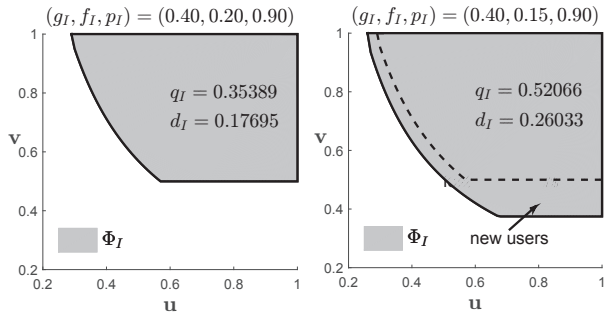
Assumption 3. *$Q_i(d_i,c_i) : \mathbb{R}_+^2 \mapsto \mathbb{R}_+$ is continuous, increasing in d_i, decreasing in c_i and satisfies $Q_i(0,c_i) = 0$.*

Different forms of the congestion function Q_i can be used to model the different technologies used by the provider. Assumption 3 implies that a provider i's congestion increases (decreases) when its data load d_i (capacity c_i) increases, and no congestion exists when no user consumes data from the provider.

We denote $Q_i^{-1}(q_i,c_i)$ as the inverse function of $Q_i(d_i,c_i)$ with respect to d_i, which defines the implied load under the capacity c_i and an observed congestion level of q_i. By Assumption 3, we know that $Q_i^{-1}(q_i,c_i)$ is continuous, increasing in both q_i and c_i, and satisfies $Q_i^{-1}(0,c_i) = 0$. We denote $\mathbf{c} = (c_i : i \in \mathcal{N})$ as the vector of the capacities of all the providers. When the providers make exogenous pricing decisions $\boldsymbol{\theta}$ and capacity planning decisions $\mathbf{c}$, the resulting congestion $\mathbf{q}$ of the providers can be determined endogenously when users choose their best providers. We define such a market equilibrium of the system as follows.

Definition 2. *For a set $\mathcal{N}$ of providers with any fixed pricing strategies $\boldsymbol{\theta}$ and capacities $\mathbf{c}$, $\mathbf{q}$ is an equilibrium if and only if*

$$q_i = Q_i\Big(D_i\big(\Phi_i(\boldsymbol{\theta},\mathbf{q});\theta_i,q_i\big),c_i\Big), \quad \forall i \in \mathcal{N}.$$



(a) equilibrium market share (b) f_I decreases to 0.15

(c) p_I decreases to 0.50 (d) g_I increases to 0.60

Figure 1: Shift of market share for a monopoly provider

To better understand the above definition, we can equivalently rephrase the above equality condition as $D_i\big(\Phi_i(\boldsymbol{\theta}, \mathbf{q}); \theta_i, q_i\big) = Q_i^{-1}(q_i, c_i)$, where the left-hand side is the induced data load of provider i given its market share $\Phi_i(\boldsymbol{\theta}, \mathbf{q})$, pricing strategy θ_i and congestion q_i and the right-hand side is its implied data load under capacity c_i. In equilibrium, both equal the actual aggregate user demand d_i. Because the equilibrium is depend on the pricing strategies $\boldsymbol{\theta}$ and capacities $\mathbf{c}$, we also denote it as $\mathbf{q}(\boldsymbol{\theta}, \mathbf{c})$.

Theorem 1. *Under Assumption 1-3, for any fixed pricing strategies $\boldsymbol{\theta}$ and capacities $\mathbf{c}$, there always exists a market equilibrium $\mathbf{q}$. In particular, when the market has only one (real monopoly) provider, the equilibrium is unique.*

Theorem 1 states that under minor assumptions of the demand (Assumption 1) and congestion (Assumption 3), the existence of a market equilibrium can be guaranteed. Besides, the equilibrium is unique in a market consisting of a monopoly provider. For this case, we denote I as the monopoly provider and define $\mathcal{I} = \{I\}$. The next theorem shows how a monopoly provider's congestion level varies when its pricing strategy θ_I and capacity c_I change, or other providers enter the market to compete.

Theorem 2. *In a monopoly market $\mathcal{I}$, provider I's unique level of congestion $q_I(\theta_I, c_I)$ in equilibrium is non-increasing in its fees f_I, p_I and capacity c_I, but is non-decreasing in its data cap g_I. When new providers enter and form a new market $\mathcal{N} \supseteq \mathcal{I} = \{I\}$, where provider I keeps its pricing θ_I and capacity c_I, if $q_I(\boldsymbol{\theta}, \mathbf{c})$ is I's congestion under an equilibrium, then $q_I(\boldsymbol{\theta}, \mathbf{c}) \leq q_I(\theta_I, c_I)$.*

Theorem 2 states that when the monopoly provider increases fees or decreases data cap, its induced congestion level will be alleviated, because its market share as well as data load will decrease. When the provider increases capacity, the congestion level will decrease, although the resulting market share and data load will increase. It also shows that with more competing providers, the existing monopoly's congestion will decrease, as users have more choices and may switch to other providers. This implies that competition could alleviate the congestion at the providers, because effectively more providers bring higher capacity to the entire market.

Under any market equilibrium $\mathbf{q}(\boldsymbol{\theta}, \mathbf{c})$, we denote R_i and S_i as the revenue and social welfare generated by provider i, defined as

$$R_i \triangleq \int_{\Phi_i} t\left(y_i^*(\phi), \theta_i\right) d\mu \quad \text{and} \quad S_i \triangleq \int_{\Phi_i} v y_i^*(\phi) d\mu, \qquad (4)$$

where $y_i^*(\phi) = y^*(\phi, \theta_i, q_i)$ and $\Phi_i = \Phi_i(\boldsymbol{\theta}, \mathbf{q})$ are evaluated at the equilibrium $\mathbf{q}(\boldsymbol{\theta}, \mathbf{c})$.

3.4 Model Parameters and Properties

Although our model is built upon generic assumptions (Assumption 1 and 3), it does not yet capture the characteristics of network services. To this end, we carefully choose the model parameters, i.e., the congestion function $Q_i(d_i, c_i)$, the achievable demand function $\rho(u, q)$ and the measure space (Φ, μ) of user domain. We discuss the rationales and implications of our choices as follows.

First, we adopt the congestion function $Q_i(d_i, c_i) = d_i/c_i$, which models the *capacity sharing* [3] nature of network services. This form has been used in much prior work such as [3, 9, 11].

Corollary 1. *Suppose $Q_i(d_i, c_i) = d_i/c_i$ for all $i \in \mathcal{N}$ and $\mathbf{q}$ is an equilibrium of a system with capacities $\mathbf{c}$ and a measure μ of the users. For any scaled system with capacities $\hat{\mathbf{c}} = k\mathbf{c}$ and $\hat{\mu}(E) = k\mu(E)$ for all $E \subseteq \Phi$ for some $k > 0$, $\hat{\mathbf{q}} = \mathbf{q}$ is also an equilibrium, under which $\hat{d}_i = k d_i$ for all $i \in \mathcal{N}$.*

Corollary 1 states that when the providers' capacities and the user size scale linearly at the same rate, the market equilibrium does not change. Thus, by appropriately scaling the capacities, we can normalize the measure of the users to be $\mu(\Phi) = 1$, i.e., F_u and F_v can be normalized to probability distribution functions $F_u(U) = F_v(V) = 1$ without loss of generality.

Next, we choose a quintessential form $\rho(u, q) = ue^{-q}$ for the achievable demand function, which is used by prior work [16, 21]. Under this form, the user's achievable demand decays exponentially at a rate of q, i.e., the level of congestion.

Corollary 2. *If $\rho(u, q) = ue^{-q}$ and $Q_i(d_i, c_i) = d_i/c_i$ for all $i \in \mathcal{N}$, let $\mathbf{q}$ be an equilibrium under parameters $\boldsymbol{\theta}, \mathbf{c}, \Phi$ and μ. For another market with $\hat{\mathbf{f}} = \mathbf{f}/(UV)$,$\hat{\mathbf{g}} = \mathbf{g}/U$,$\hat{\mathbf{p}} = \mathbf{p}/V$, $\hat{\mathbf{c}} = \mathbf{c}/U$, $\hat{\boldsymbol{\Phi}} = [0, 1] \times [0, 1]$ and $\hat{\mu}([0, \hat{u}] \times [0, \hat{v}]) = \mu([0, U\hat{u}] \times [0, V\hat{v}])$ for all $(\hat{u}, \hat{v}) \in \hat{\Phi}$, we must have $\hat{\mathbf{q}} = \mathbf{q}$ as an equilibrium under which $\hat{d}_i = d_i/U$ for all $i \in \mathcal{N}$.*

Corollary 2 states that under the exponential form of achievable demand, any domain of the users can be normalized onto the domain $[0, 1] \times [0, 1]$. In particular, the equilibrium does not change when 1) the lump-sum fees, data caps, capacities and the users' desirable demands are normalized by U, or 2) the lump-sum and per-unit fees and users' values are normalized by V. Based on this result, we can focus on $U = V = 1$ without loss of generality and we will consider the forms $F_u(x) = x^\alpha$ and $F_v(x) = x^\beta$ for $x \in [0, 1]$, where α and β model the distribution of users with respect to their desirable data demands and values, respectively. For instance, when $\beta = 1$, user values are uniformly distributed; otherwise, they are leaning toward the high ($\beta > 1$) or low ($\beta < 1$) values in the domain $[0, 1]$. In summary, we will analyze the two-part tiered pricing of the providers with the following assumption.

Assumption 4. *Any provider i's congestion satisfies $Q_i(d_i, c_i) = d_i/c_i$, the users are distributed by $F_u(x) = x^\alpha, F_v(x) = x^\beta$ for $x \in [0, 1]$, and their achievable demands satisfy $\rho(u, q) = ue^{-q}$.*

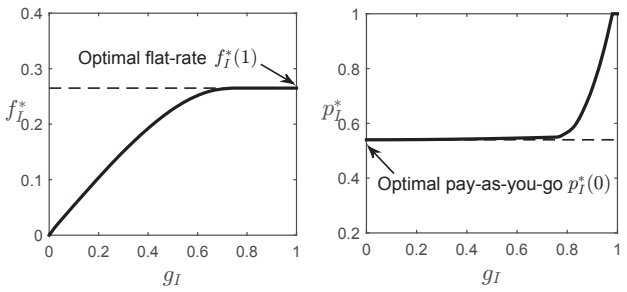
When the level of congestion is exogenously given, Lemma 2 implies that the market share of a monopoly provider will decrease when it raises fees, e.g., f_I and p_I, or reduces data cap g_I. However, the higher fees or lower data cap would alleviate the provider's congestion in equilibrium by Theorem 2, which results in attracting more congestion-sensitive users to join. As a result, the dynamics of the provider's market share combines both effects and is not monotonic. Under Assumption 4, we illustrate an example in Figure 1 that shows how a monopoly provider's market share shifts when it changes the pricing strategy under equilibrium. In this example, the user distribution is parameterized by $\alpha = 4.0$ and $\beta = 0.8$ and the provider has a capacity of $c_I = 0.5$. In each subfigure, x-axis and y-axis vary the desirable data demand u and value v of the user types. In other words, each point in the subfigures represents a unique user type. In subfigure (a), we illustrate the market share Φ_I as the shaded area when the provider uses strategy $\theta_I = (g_I, f_I, p_I) = (0.4, 0.2, 0.9)$ and induces congestion $q_I = 0.35389$ and data load $d_I = 0.17695$ in equilibrium. Notice that the region of market share Φ_I has a flat-bottom on the right side where the desirable demand u is large. This flat boundary corresponds to the value of $v = f_I/g_I$, because if a user's per-unit value is lower than the effective average per-unit value for the capped amount data g_I, she would not use the provider. From subfigure (a) to (b), the provider decreases f_I from 0.2 to 0.15 and the cheaper lump-sum fee attracts a larger market share for the provider and induces higher data load and congestion in equilibrium. From subfigure (a) to (c), the provider decreases p_I from 0.9 to 0.5 and the cheaper per-unit usage fee induces higher load and congestion. However, the market share Φ_I shrinks, because higher congestion reduces the utility of some users, forcing them to leave the provider. From subfigure (a) to (d), the provider increases g_I from 0.4 to 0.6 and the larger data cap again induces higher data load and congestion in equilibrium. However, the resulting market share attracts more low-value heavy users and loses some high-value light users. Notice that although the changes in market share from subfigure (a) to subfigures (b) to (d) are not monotonic, the increase in congestion and data load in these cases are consequences of cheap prices and larger data caps stated in Theorem 2.

4. REVENUE-OPTIMAL PRICING

In this section, we study the revenue-optimal two-part pricing of the providers and characterize its dynamics under changes of system parameters, i.e., the distribution of users' demand and values, and the capacity of the providers. Because data cap is the demarcation between the lump-sum (for demand below the data cap) and usage-based (for demand above the data cap) charges, we first analyze the impact of data cap on the provider's pricing decisions and the corresponding data load, congestion and revenue, and then identify the role of data cap in the two-part pricing structure.

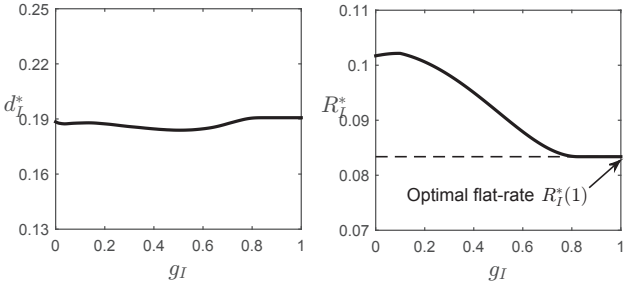
4.1 The Role of Data Cap

We start with a monopoly provider and assume that it chooses its pricing strategy, i.e., f_I and p_I, to maximize its revenue R_I. Given any fixed data cap g_I, we denote $f_I^*(g_I)$ and $p_I^*(g_I)$ as the optimal lump-sum and per-unit fee, respectively, and $d_I^*(g_I)$ and $R_I^*(g_I)$ as the resulting data load and maximum revenue. We study how various data caps influence its optimal pricing decisions, e.g., f_I^* and p_I^*, and the resulting optimal revenue R_I^*. In particular, we compare two-part schemes with the flat-rate pricing, a special case of an infinite data cap, i.e., $g_I = +\infty$. For simplification, we define the maximum revenue under the flat-rate as $R_\infty^* \triangleq \lim_{g_I \to \infty} R_I^*(g_I)$.



(a) optimal lump-sum fee (b) optimal per-unit fee

Figure 2: Optimal lump-sum fee $f_I^*(g_I)$ and per-unit fee $p_I^*(g_I)$ under varying data cap g_I.



(a) corresponding data load (b) maximum revenue

Figure 3: Corresponding data load $d_I^*(g_I)$ and maximum revenue $R_I^*(g_I)$ under varying data cap g_I.

We consider the models under Assumption 4 where the maximum desirable demand is normalized as $U = 1$. As a result, an effective data cap g that might influence the users' demand has to be less than or equal to 1. In other words, when g_I is larger than 1, the two-part pricing is equivalent to a flat-rate scheme, i.e., $R_I^*(g_I) = R_\infty^* = R_I^*(1)$ for any $g_I \geq 1$. Thus, we will focus on $g_I \in [0, 1]$ without loss of generality.

Figure 2 plots the optimal lump-sum fee $f_I^*(g_I)$ and the optimal per-unit fee $p_I^*(g_I)$ as a function of the data cap g_I varying along the x-axis, respectively, where users are uniformly distributed, i.e., $\alpha = \beta = 1$, and the provider has a unit capacity, i.e., $c = 1$. We observe that both fees increase with the data cap; however, when g_I is small, $f_I^*(g_I)$ increases steeper and $p_I^*(g_I)$ is almost flat; when g_I is large, $f_I^*(g_I)$ becomes flatter and $p_I^*(g_I)$ increases steeper. This observation also illustrates that when g_I is small, the pricing structure is close to the pay-as-you-go pricing, where the lump-sum component is close to zero and the per-unit charge is close to the optimal pay-as-you-go price $p_I^*(0)$; when g_I becomes large, the pricing structure converges to the flat-rate scheme, where the lump-sum component converges to the optimal flat-rate $f_I^*(1)$ and the per-unit fee further increases to capture revenue from high-value users with large demands. The following theorem shows that our observed trends in the optimal fees are not particular to the model parameters, i.e., $\alpha = \beta = c = 1$.

Theorem 3. *Under Assumption 4, for any monopoly market with parameters $\alpha, \beta, c > 0$, the optimal lump-sum fee $f_I^*(g_I)$ is non-decreasing and concave in the data cap g_I, and the optimal per-unit fee $p_I^*(g_I)$ is non-decreasing and convex in the data cap g_I.*

Theorem 3 intuitively states that when the data cap is relaxed, a monopoly provider could compensate by increasing its prices so as to maximize revenue. Without adapting prices, larger data cap

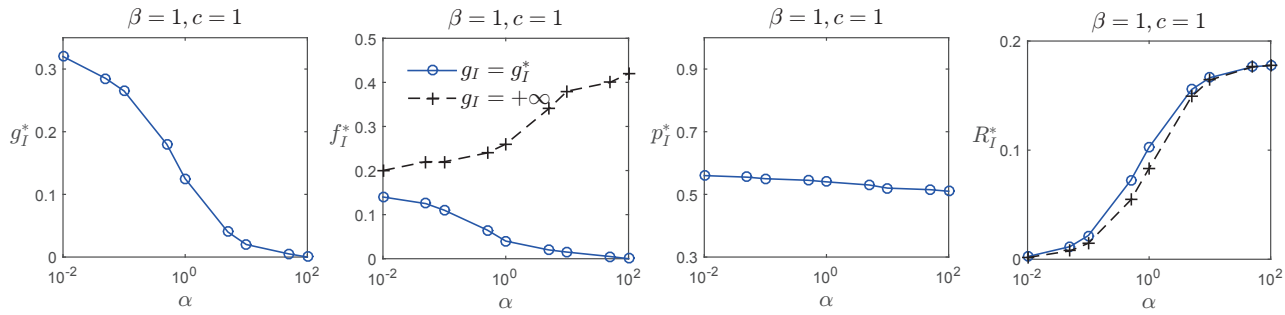


Figure 4: Optimal two-part pricing (g_I^*, f_I^*, p_I^*) and the resulting maximum revenue R_I^* under varying α.

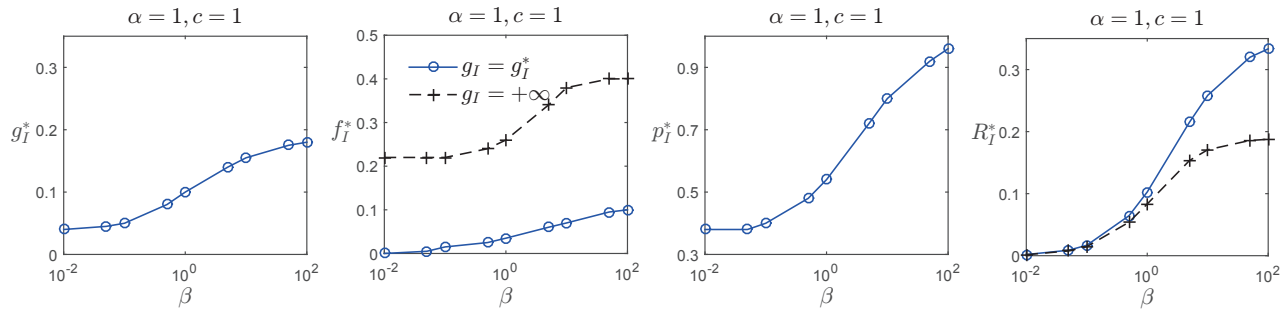


Figure 5: Optimal two-part pricing (g_I^*, f_I^*, p_I^*) and the resulting maximum revenue R_I^* under varying β.

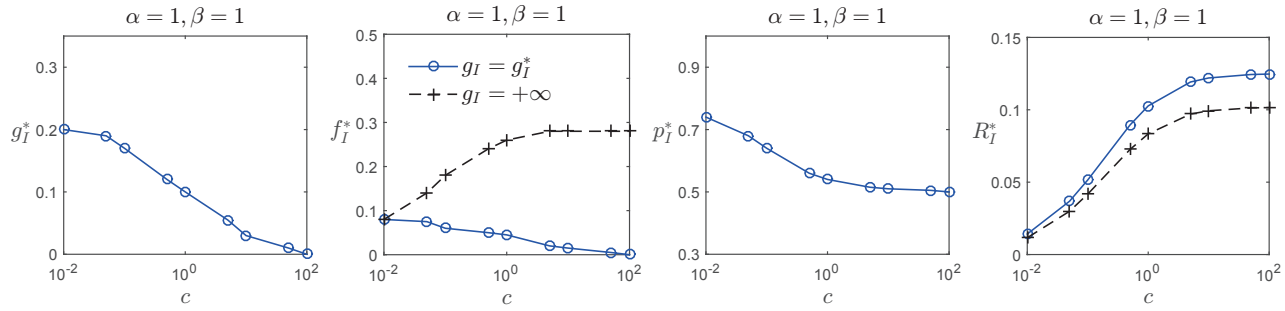


Figure 6: Optimal two-part pricing (g_I^*, f_I^*, p_I^*) and the resulting maximum revenue R_I^* under varying c.

will induce higher load and congestion in a new equilibrium by Theorem 2. However, the optimal tradeoff between prices and data load will induce an increase in prices, which oppositely results in a decrease in load as stated by Theorem 2. Under any fixed data cap g_I, if new providers enter the market, by Theorem 2, the existing provider's congestion and load will decrease. To increase load and maximize revenue under competition, it would need to reduce prices as compared to the prices in a monopoly market. Thus, under market competition, we would observe lower prices but similar pricing trends, e.g., when $g_I = 0$ and $g_I = 1$, the optimal per-unit and lump-sum prices will be lower than $p_I^*(0)$ and $f_I^*(1)$, respectively. The convexity of $p_I^*(g_I)$ and concavity of $f_I^*(g_I)$ shown in Theorem 3 also imply that the rate of change of the optimal per-unit and lump-sum fees are slow when the data cap is small and large, respectively. The reason is that in the regime of small (large) data cap, the two-part pricing structure is very close to that of the pay-as-you-go (flat-rate) scheme; and therefore, the revenue-optimal per-unit (lump-sum) fee is close to the optimal pay-as-you-go $p_I^*(0)$ (flat-rate $f_I^*(1)$) price. In general, the increase in data cap transitions a provider's pricing from a more pay-as-you-go structure to a more flat-rate structure regardless of the market structure.

Figure 3 plots the corresponding load $d_I^*(g_I)$ and maximum revenue $R_I^*(g_I)$, respectively. We observe that although the resulting data load under any data cap g_I does not vary much, the maximum revenue $R_I^*(g_I)$ varies significantly. In particular, when the pricing converges to the flat-rate structure as g_I goes to 1, the optimal revenue decreases to a minimum value. The following corollary shows that this observation is a very general result that does not depend on Assumption 4 of our model.

Corollary 3. *Under Assumption 1-3 and any monopoly market, the provider's revenue satisfies $R_I^*(g_I) \geq R_\infty^*$, $\forall g_I \geq 0$.*

Corollary 3 implies that the maximum revenue generated from any optimal two-part pricing under a fixed data cap $g_I \geq 0$ is no less than that under a flat-rate scheme. In other words, a provider could always be better off by switching from an optimal flat-rate scheme to an optimal two-part pricing scheme. Intuitively, under any fixed lump-sum f_I, the two-part structure generalizes the flat-rate scheme under which $p_I = 0$ and data cap does not play a role. Consequently, even without changing its flat-rate f_I, the provider could restrict g_I and increase p_I to trade off between higher usage revenue from high-value users' demand and its market share, which potentially lead to higher total revenue. This result is consistent with the views in prior work [5, 6, 8] that data cap could help providers extract higher revenue from the market. Although pure flat-rate is inferior to two-part schemes, Figure 3 shows that the maximum revenue does not always increase when the pricing struc-

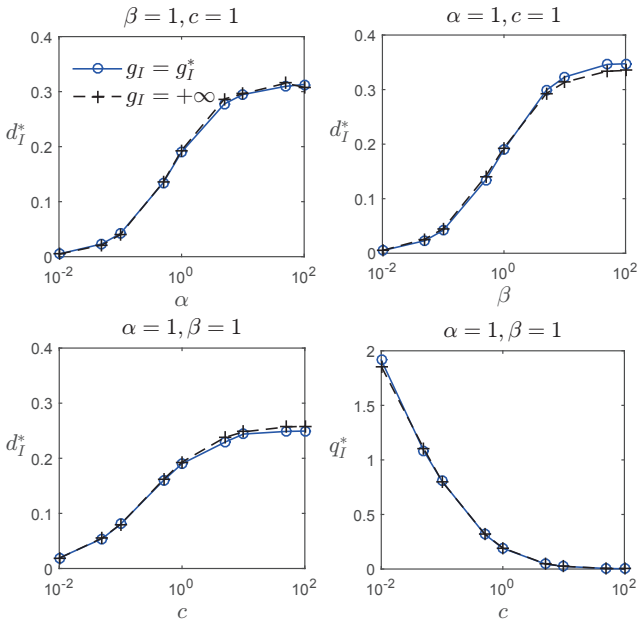


Figure 7: Resulting data load d_I^* under varying α, β and c, and congestion q_I^* under varying c.

ture moves towards the pure pay-as-you-go scheme, i.e., $g_I = 0$. In the next subsection, we further explore the optimal two-part pricing structure and its dynamics under various system parameters.

4.2 Optimal Two-Part Pricing Structure

In the previous subsection, we considered the revenue-optimal prices $f_I^*(g_I)$ and $p_I^*(g_I)$ under any fixed data cap g_I. In this subsection, we further explore revenue-optimal data cap g_I^*. Under any distribution of users specified by parameters α and β, and any fixed capacity c of the provider, we denote $\theta_I^* \triangleq (g_I^*, f_I^*, p_I^*)$ as the optimal two-part pricing that results in the maximum revenue R_I^*. We explore the dynamics of optimal two-part pricing θ_I^* when system parameters, i.e., α, β and c, change.

Figures 4 to 6 plot the optimal two-part pricing g_I^*, f_I^* and p_I^*, and the resulting maximum revenue R_I^* as functions of 1) parameter α of the distribution of users' demands, 2) parameter β of the distribution of users' values, and 3) parameter c of the provider's capacity, respectively. In each of the second and fourth subfigures, we also plot the optimal flat-rate price $f_I^*(1)$ and its corresponding revenue for comparison, respectively. We observe that as α, β and c increase, the maximum revenue increases under both the optimal flat-rate and optimal two-part schemes, where the latter induces higher revenue than the former as stated by Corollary 3. Intuitively, larger α, β and c imply higher users' demand, users' value and provider's capacity, respectively, where the provider is in a more advantageous position in extracting revenue. For example, we observe that the optimal flat-rate price $f_I^*(1)$ always increases under these changes. We also observe that when α or c increases, g_I^*, f_I^* and p_I^* all decrease; however, when β increases, g_I^*, f_I^* and p_I^* all increase. These monotonicities could be formally stated as follows.

Theorem 4. *Under Assumption 4, the data cap g_I^*, lump-sum fee f_I^* and per-unit fee p_I^* of the optimal two-part pricing are non-increasing in α and c, and are non-decreasing in β. The resulting maximum revenue R_I^* is non-decreasing in all α, β and c.*

By Theorem 3, we know that the optimal fees $f_I^*(g_I)$ and $p_I^*(g_I)$ increase with the data cap g_I when the provider's pricing structure transitions from pure pay-as-you-go to pure flat-rate pricing. Theorem 4 implies that when the demand of the users or/and the capacity of the provider increase, the optimal two-part pricing moves closer to a more pay-as-you-go type of structure; however, when the values of the users increase, it moves closer to a structure that has a bigger lump-sum component with a larger data cap. When the capacity increases, it reduces the network congestion and increases the achievable demand of the users. Because flat-rate pricing does not restrict the users' demand, it cannot effectively optimize the provider's revenue when the user demand increases; therefore, the provider will transition to impose a data cap and a per-unit fee so as to restrict demands from "bandwidth hogs". On the contrary, when the values of users tend to be concentrated towards the high-end, the provider could impose a high per-unit fee and also introduce a lump-sum component to capture users in the low-value and low-demand regime. The high per-unit fee does not affect these users as their demand is lower than the data cap.

Under market competition, if the provider keeps the same data cap g_I^*, it has to reduce the lump-sum and per-unit prices to capture market share and optimize revenue; however, it might be better off to increase data cap from g_I^* to attract more users rather than decreasing the fees too much. In general, all three components of the optimal two-part pricing under market competition have to be "cheaper" than those under a monopoly market. However, it does not alter the structure of the optimal two-part pricing, i.e., whether it is more like pay-as-you-go or flat-rate, in an obvious manner.

The implications of Theorem 4 also provide compelling explanations of why Internet service providers have shifted their pricing from the traditional flat-rate to the two-part structure recently. In the early years of the Internet, data demands were mostly for texts and used by small groups of advanced users with high-value tasks, e.g., scientific and business purposes. The network capacities were scarce before the emergence of fiber optics backbones. Under such conditions, flat-rate pricing could be used to effectively recoup costs and the maximum revenue under two-part pricing would not be much higher than that under an optimal flat-rate scheme. As a result, most providers adopted flat-rate pricing because of its simplicity. As mobile devices become a key means to access the Internet for end-users and multimedia content become more pervasive, together with the increase in network capacities, Internet traffic keeps growing more than 50% per annum [13]. As a result, providers now feel that flat-rate schemes cannot effectively influence users' demand and optimize their revenues; and therefore, start to adopt two-part pricing schemes. Furthermore, implied by Theorem 4, if these trends continue, we would expect that providers will further reduce the lump-sum component and data cap and move closer to pure pay-as-you-go pricing in the near future. In practice, the introduction of data cap in the two-part pricing also provides a means for providers to transition from flat-rate to pay-as-you-go smoothly, so that the changes will not be too abrupt for users and providers will not lose users due to the structural change of pricing schemes.

Figure 7 plots the corresponding data load under the optimal flat-rate and two-part pricing schemes when α, β and c increase, respectively. As the system congestion has the same trends as data load under fixed capacities, we only plot the corresponding network congestion in the lower-right subfigure when capacity c varies. We observe that the induced data load and network congestion do not differ much under the optimal two-part and flat-rate schemes. In general, increasing in users' desirable demand, users' values or provider's capacity will induce higher data load in the system; however, the congestion will decrease when the capacity becomes more abundant.

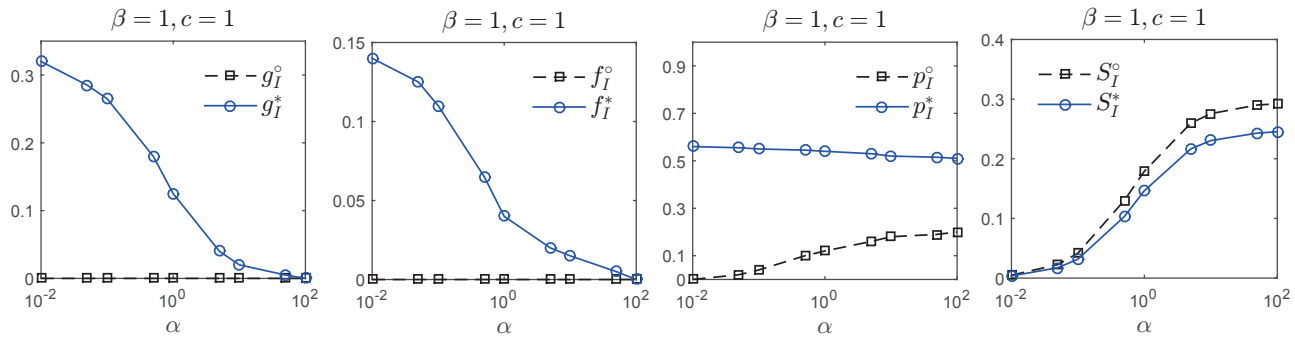


Figure 8: Welfare-optimal and revenue-optimal pricing schemes under varying α.

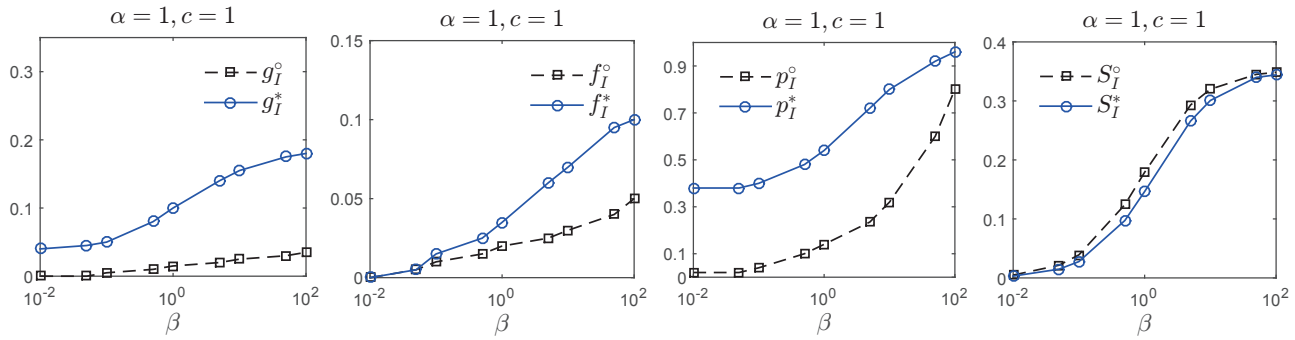


Figure 9: Welfare-optimal and revenue-optimal pricing schemes under varying β.

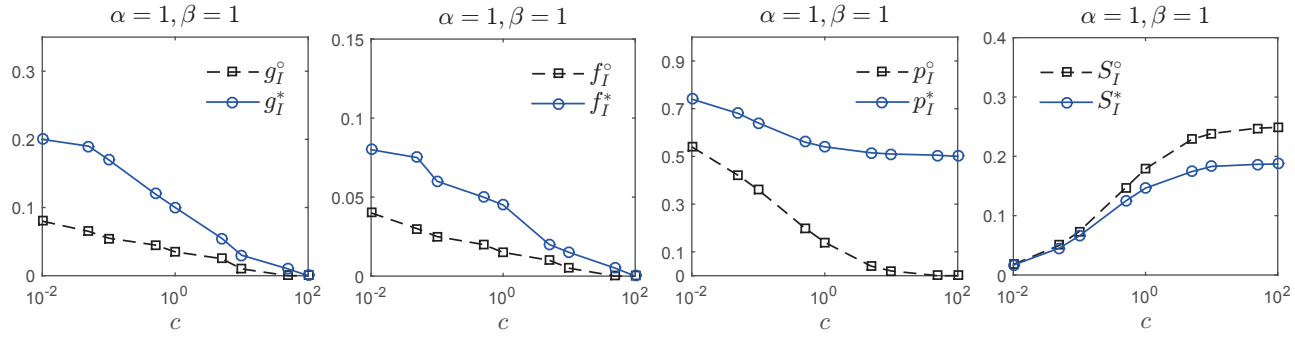


Figure 10: Welfare-optimal and revenue-optimal pricing schemes under varying c.

Corollary 4. *Under Assumption 4 and an optimal two-part pricing $\theta_I^* = (g_I^*, f_I^*, p_I^*)$, the provider's data load d_I^* is non-decreasing in α, β and c, and the corresponding network congestion q_I^* is non-decreasing in α and β, but is non-increasing in c.*

Corollary 4 implies that when the demand or values of users, or the capacity of the provider increases, under the optimal two-part pricing, the system will accommodate higher data load. The system congestion will decrease if the capacity increases; otherwise, it will increase. The increase in data load shows that revenue-optimal pricing is partially aligned with the users' welfare where the provider would adaptively serve more data demand. In the next section, we will explore how different the revenue-optimal pricing is from the welfare-optimal solution, through which we will obtain some insights on how monopolistic providers should be regulated for social welfare maximization.

In summary, the structure of revenue-optimal two-part pricing is largely influenced by the data cap, which plays a transitional role (mechanism) between flat-rate and pay-as-you-go pricing schemes. As users' demand and providers' capacity grow in the Internet, revenue objectives will drive providers to shift from flat-rate towards usage-based schemes, where data cap and both lump-sum and per-unit fees would decrease. Although market competition would force the providers to use "cheaper" schemes, i.e., higher data cap and lower lump-sum and per-unit fees, it does not necessarily change the structure of the optimal two-part pricing.

5. WELFARE-OPTIMAL PRICING

In the previous section, we studied the revenue-optimal pricing and showed that it accommodates higher data load when the users' demand or values, or the providers' capacities increase. However, the revenue-optimal solution does not maximize the social welfare, i.e., the total utility of the providers and their users. Although market competition could improve the social welfare, which would be maximized under a perfect competitive market, a large deviation from the maximum welfare would more likely happen in a monopoly market. In the section, we focus on a monopoly provider and compare its welfare-optimal and revenue-optimal solutions.

Under any distribution of users specified by parameters α and β, and any capacity c of the provider, we denote $\theta_I^\circ \triangleq (g_I^\circ, f_I^\circ, p_I^\circ)$ as the optimal two-part pricing that maximizes the social welfare S_I

(defined in Equation 4) and results in the maximum welfare S_I°. We denote the corresponding network congestion and data load as q_I° and d_I°, respectively. To make a comparison, we denote S_I^* as the social welfare achieved under the provider's revenue-optimal pricing θ_I^*. Similarly, we explore the dynamics of the welfare-optimal pricing θ_I° when system parameters, i.e., α, β and c, change.

Figures 8 to 10 plot the welfare-optimal solution g_I°, f_I° and p_I°, and the resulting maximum welfare S_I° as functions of 1) parameter α of the distribution of users' demands, 2) parameter β of the distribution of users' values, and 3) parameter c of the provider's capacity, respectively. As a comparison, we also plot the corresponding revenue-optimal solution g_I^*, f_I^* and p_I^*, and the resulting social welfare S_I^* in the four subfigures, respectively. We observe that the curves of the welfare-optimal pricing g_I°, f_I° and p_I° are always lower than those of the revenue-optimal pricing g_I^*, f_I^* and p_I^*, respectively. This observation can be formally shown as follows.

Theorem 5. *Under Assumption 4, for any monopoly market with parameters $\alpha, \beta, c > 0$, we have $g_I^\circ \le g_I^*$, $f_I^\circ \le f_I^*$ and $p_I^\circ \le p_I^*$.*

Theorem 5 implies that to shift the provider's two-part pricing from revenue-optimal to welfare-optimal, one should lower its fees, i.e., f_I and p_I; however, allow the provider to further throttle the data cap and move towards a usage-based structure. Besides revenue maximization, this result provides justifications for regulators, e.g., the US FCC, to encourage usage-based pricing for the Internet service providers. On the one hand, exorbitant fees reduce users' utilities and data demand, resulting lower social welfare; on the other hand, limiting the data cap will self-regulate the demand of low-value, and therefore, increase the social welfare.

When comparing the trends between the revenue and welfare-optimal solutions, we further observe that the welfare-optimal solution has the same trend as that of the revenue-optimal solution when β or c increases; however, this trend is reversed when α increases. These monotonicities could be formally stated as follows.

Theorem 6. *Under Assumption 4, the data cap g_I°, lump-sum fee f_I° and per-unit fee p_I° of the welfare-optimal two-part pricing are non-increasing in c, and are non-decreasing in α and β. The resulting revenue S_I° is non-decreasing in all α, β and c.*

Similar to the result of Theorem 4, Theorem 6 shows that when c increases, the welfare-optimal solution moves towards usage-based, but also reduces the per-unit fee to allow more user demand. However, when users' demand becomes more concentrated at the high-end as α increases, although the welfare-optimal solution is close to the pay-as-you-go scheme, the optimal per-unit fee increases so as to discourage the data demand from low-value users.

By comparing the maximum social welfare S_I° with S_I^*, we observe that the difference widens when α and c increases, but reaches largest as the values of users are uniformly distributed, i.e., $\beta = 1$. This implies that regulations are most in need when the capacity and users' demand become large. In both cases, we observe that both revenue and welfare-optimal solution tend to be pure pay-as-you-go, i.e., $g_I = 0$; however, the per-unit price for maximum welfare is lower than that of the revenue-optimal solution. This further implies that regulators could encourage usage-based pricing for the providers; however, price regulation on the per-unit fee might be needed to guarantee higher social welfare.

Add on to Figure 7, Figure 11 compares the data load and congestion under the revenue and welfare-optimal pricing schemes. We observe that the network congestion drops as the capacity c increases; however, when α or c becomes large, the achieved load from welfare-optimal pricing is much higher. This coincides with the large gaps in social welfare, where regulation is most in need.

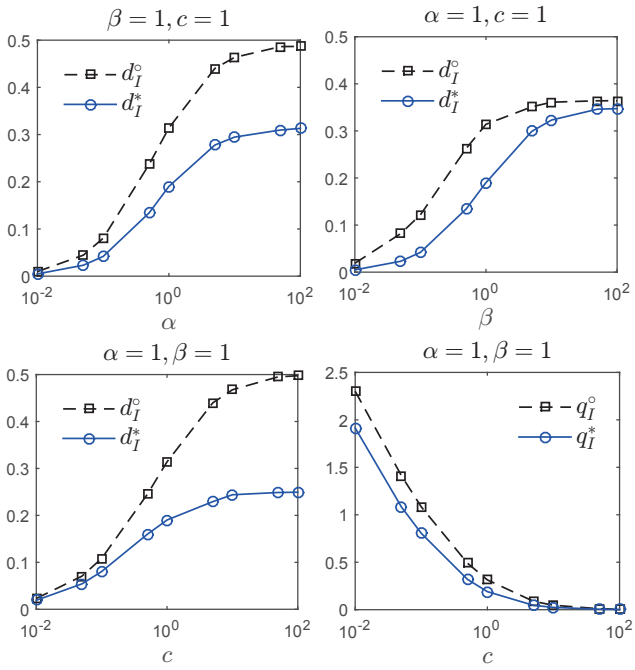


Figure 11: Resulting data load d_I^* and d_I° under varying α, β and c, and congestion q_I^* and q_I° under varying c.

The trends of d_I° and q_I° under the welfare-optimal pricing θ_I° are similar to those under the revenue-optimal pricing θ_I^* shown in Corollary 4, and can be shown as follows.

Corollary 5. *Under Assumption 4 and a welfare-optimal two-part pricing $\theta_I^\circ = (g_I^\circ, f_I^\circ, p_I^\circ)$, the data load d_I° is non-decreasing in α, β and c, and the corresponding network congestion q_I° is non-decreasing in α and β, but is non-increasing in c.*

Compared to the result of Corollary 4, Corollary 5 is even more intuitive: higher provider's capacity and users' demand and values will lead to higher social welfare when it is optimized.

In summary, our results suggest that regulatory authorities might want to regulate a monopoly market under heavy demand of users or abundant capacity of the provider. Under these scenarios, welfare-optimal solution will have lower fees and data cap than the revenue-optimal solution, forcing the pricing structure of the provider moving further towards usage-based. As the growth of users' demand and providers' capacities in the current Internet, our result further justifies from a welfare perspective that the regulators should encourage ISPs to shift towards two-part pricing with limited data caps; however, it also suggests that the per-unit fees might need to be regulated when market competition does not exist.

6. CONCLUSIONS

In this paper, we study the role of data cap in the optimal structure of two-part tariffs. We present a novel model of users' demand and preferences over pricing and congestion alternatives and derive the market share and congestion of the service providers under a market equilibrium. Based on the equilibrium model, we characterize the two-part structure of the revenue-optimal and welfare-optimal pricing schemes. We identify that the data cap provides a mechanism for service providers to transition from flat-rate to pay-as-you-go type of usage-based schemes. Our results reveal that with growing data demand and network capacity in the current Internet, revenue-optimal pricing will move towards usage-based

schemes with diminishing data caps. Although market competition might drive pricing "cheaper", it does not necessarily change the optimal structure of the two-part tariff. From a perspective of social welfare, our results suggest that regulators might want to promote usage-based pricing but regulate the per-unit fees, because the structure of the welfare-optimal tariff comprises lower fees and data cap than those of the revenue-optimal counterpart.

Acknowledgments

This work is partially supported by the research grants for the HCCS project at ADSC from Singapore's Agency for Science, Technology and Research (A*STAR), Ministry of Education of Singapore AcRF grants T1 R-252-000-526-112 and R-252-000-572-112. This work is also supported in part by National Nature Science Foundation of China under Grant No. 61379038.

APPENDIX

A. PROOFS OF SELECTED RESULTS

Proof of Lemma 1: For any user type ϕ that uses the ISP with pricing scheme θ and congestion q, her utility is $\pi(y) = vy - f - p(y-g)^+$ from Equation 1. When $v \geq p$, the utility $\pi(y)$ is non-decreasing in $y \in \big[0, \rho(u,q)\big]$ and thus the optimal demand is $y^*(\phi,\theta,q) = \rho(u,q)$. When $v < p$, the utility $\pi(y)$ is non-decreasing in $y \in \big[0, g\big]$ and non-increasing in $y \in \big(g, \rho(u,q)\big]$, and thus the optimal demand is $y^*(\phi,\theta,q) = g$. In summary, we have $y^*(\phi,\theta,q) = \rho(u,q) - \big[\rho(u,q)-g\big]^+ \mathbf{1}_{\{v<p\}}$, and therefore, the optimal usage $y^*(\phi,\theta,q)$ is non-increasing in p and q and non-decreasing in u,v and g. ■

Proof of Lemma 2: For a set $\mathcal{N}$ of providers and any user type $\phi \in \Phi_i(\boldsymbol{\theta},\mathbf{q})$, we have $y^*(\phi,\hat{\theta}_i,q_i) \geq y^*(\phi,\theta_i,q_i)$. Because the optimal data usage y^* is non-increasing in p and non-decreasing in g from Lemma 1. Further, the utility function $\pi(y)$ is non-decreasing in the data usage y, thus it satisfies $\pi\big(y^*(\phi,\hat{\theta}_i,q_i)\big) \geq \pi\big(y^*(\phi,\theta_i,q_i)\big) \geq \pi\big(y^*(\phi,\theta_j,q_j)\big) = \pi\big(y^*(\phi,\hat{\theta}_j,q_j)\big)$ for $\forall j \neq i$. Then we have $\phi \in \Phi_i(\hat{\boldsymbol{\theta}},\mathbf{q})$ from Definition 1 and therefore $\Phi_i(\boldsymbol{\theta},\mathbf{q}) \subseteq \Phi_i(\hat{\boldsymbol{\theta}},\mathbf{q})$. Similarly, we can show that $\Phi_j(\boldsymbol{\theta},\mathbf{q}) \supseteq \Phi_j(\hat{\boldsymbol{\theta}},\mathbf{q}), \forall\, j \neq i$.

For two sets $\mathcal{N}$ and $\mathcal{N}'$ of providers and any provider $i \in \mathcal{N}$, for any user type $\phi \in \Phi_i(\boldsymbol{\theta}',\mathbf{q}')$, based on Definition 1, it satisfies that $\pi\big(y^*(\phi,\theta_i,q_i)\big) = \pi\big(y^*(\phi,\theta_i',q_i')\big) \geq \pi\big(y^*(\phi,\theta_j',q_j')\big) = \pi\big(y^*(\phi,\theta_j,q_j)\big), \forall j \in \mathcal{N}\backslash\{i\}$. Then we have $\phi \in \Phi_i(\boldsymbol{\theta},\mathbf{q})$ and therefore $\Phi_i(\boldsymbol{\theta}',\mathbf{q}') \subseteq \Phi_i(\boldsymbol{\theta},\mathbf{q}), \forall i \in \mathcal{N}$. ■

Proof of Theorem 1: We first prove the existence of equilibrium. By Definition 2, $\mathbf{q}$ is an equilibrium if and only if for all $i \in \mathcal{N}$,

$$q_i = Q_i\Big(D_i\big(\Phi_i(\boldsymbol{\theta},\mathbf{q});\theta_i,q_i\big),c_i\Big) = Q_i\big(D_i(\boldsymbol{\theta},\mathbf{q}),c_i\big).$$

Since $\boldsymbol{\theta}$ and $\mathbf{c}$ are constants, we omit them in the notation and write the above in a matrix form as $\mathbf{q} = Q\big(D(\mathbf{q})\big) = Q \circ D(\mathbf{q})$. Thus, we can view the composite function $Q \circ D$ as a mapping from the convex set $\mathbb{R}_+^{|\mathcal{N}|}$ to itself. By Assumption 3, we know that each $Q_i(d_i,c_i)$ is continuous in d_i and thus each $Q_i\big(D_i(\boldsymbol{\theta},\mathbf{q}),c_i\big)$ is continuous in $\mathbf{q}$. To this end, we know that $Q \circ D(\mathbf{q})$ is continuous in $\mathbf{q}$. By Brouwer fixed-point theorem, there always exists a fixed point that satisfies $Q \circ D(\mathbf{q}) = \mathbf{q}$ and is also an equilibrium.

We then prove the uniqueness of the equilibrium in a monopoly market by contradiction. Suppose there exist two equilibriums q_I and q_I' under fixed pricing strategy θ_I and capacity c_I. Without loss of generality, we assume that $q_I' > q_I$. For any user type $\phi \in \Phi_I\big(\theta_I,q_I'\big)$, we have $y^*(\phi,\theta_I,q_I) \geq y^*(\phi,\theta_I,q_I')$. Because the optimal data usage y^* is non-increasing in the congestion q_I from Lemma 1. Further, the utility function $\pi(y)$ is non-decreasing in the data usage y, thus it satisfies $\pi\big(y^*(\phi,\theta_I,q_I)\big) \geq \pi\big(y^*(\phi,\theta_I,q_I')\big) \geq 0$ implying that $\phi \in \Phi_I\big(\theta_I,q_I\big)$ from Definition 1. Thus we have $\Phi_I(\theta_I,q_I') \subseteq \Phi_I(\theta_I,q_I)$. By Equation 3, we deduce $D_I\big(\theta_I,q_I\big) = \int_{\Phi_I(\theta_I,q_I)} y^*(\phi,\theta_I,q_I)d\mu \geq \int_{\Phi_I(\theta_I,q_I')} y^*(\phi,\theta_I,q_I')d\mu = D_I\big(\theta_I,q_I'\big)$. Because $Q_I(d_I,c_I)$ is non-decreasing in d_I from Assumption 3, we have

$$q_I = Q_I\Big(D_I\big(\theta_I,q_I\big),c_I\Big) \geq Q_I\Big(D_I\big(\theta_I,q_I'\big),c_I\Big) = q_I'$$

based on Definition 2, which is contradictory with the supposition that $q_I' > q_I$. Therefore, the equilibrium is unique in the market with only one real provider. ■

Proof of Theorem 2: We first prove the congestion $q_I(\theta_I,c_I)$ is non-increasing in the lump-sum fee f_I by contradiction. Suppose $q_I(\theta_I,c_I)$ is not non-increasing in f_I, there must exist pricing strategies $\theta_I' = (g_I,f_I',p_I)$ and $\theta_I = (g_I,f_I,p_I)$ satisfying $f_I' > f_I$ and $q_I(\theta_I',c_I) > q_I(\theta_I,c_I)$. For any user type $\phi \in \Phi_I\big(\theta_I',q_I(\theta_I',c_I)\big)$, because the optimal data usage y^* is non-increasing in the congestion q and the lump-sum fee f from Lemma 1, we have $y^*\big(\phi,\theta_I',q_I(\theta_I',c_I)\big) \leq y^*\big(\phi,\theta_I,q_I(\theta_I,c_I)\big)$. Furthermore, because the utility function π is non-decreasing in y and non-increasing in f, we have

$$\pi\Big(y^*\big(\phi,\theta_I,q_I(\theta_I,c_I)\big)\Big) \geq \pi\Big(y^*\big(\phi,\theta_I',q_I(\theta_I',c_I)\big)\Big) \geq 0$$

implying that $\phi \in \Phi_I\big(\theta_I,q_I(\theta_I,c_I)\big)$ from Definition 1. Thus we have $\Phi_I\big(\theta_I',q_I(\theta_I',c_I)\big) \subseteq \Phi_I\big(\theta_I,q_I(\theta_I,c_I)\big)$. From Equation 3, it satisfies $D_I\big(\theta_I',q_I(\theta_I',c_I)\big) \leq D_I\big(\theta_I,q_I(\theta_I,c_I)\big)$. Because $Q_I(d_I,c_I)$ is non-decreasing in d_I from Assumption 3, we have

$$\begin{aligned} q_I(\theta_I',c_I) &= Q_I\Big(D_I\big(\theta_I',q_I(\theta_I',c_I)\big),c_I\Big) \\ &\leq Q_I\Big(D_I\big(\theta_I,q_I(\theta_I,c_I)\big),c_I\Big) = q_I(\theta_I,c_I) \end{aligned}$$

by Definition 2, which is contradictory with the supposition that $q_I(\theta_I',c_I) > q_I(\theta_I,c_I)$. Therefore, $q_I(\theta_I,c_I)$ is non-increasing in f_I. Similarly, we can prove that $q_I(\theta_I,c_I)$ is non-increasing in the per-unit fee p_I, the capacity c_I and non-decreasing in the data cap g_I, and when new providers enter the market, the existing provider's congestion must be improved, i.e., $q_I(\boldsymbol{\theta},\mathbf{c}) \leq q_I(\theta_I,c_I)$. ■

Proof of Corollary 1: Given $\hat{\mathbf{c}} = k\mathbf{c}$, $Q_i(d_i,c_i) = d_i/c_i$ and $\hat{d}_i = kd_i$ for all $i \in \mathcal{N}$, we know that $\hat{q}_i = Q(\hat{d}_i,\hat{c}_i) = \hat{d}_i/\hat{c}_i = kd_i/kc_i = Q_i(d_i,c_i) = q_i$ and therefore $\hat{\mathbf{q}} = \mathbf{q}$. Thus, we only need to show that under $\hat{\mathbf{q}} = \mathbf{q}$, we must have $\hat{d}_i = kd_i$ for all $i \in \mathcal{N}$. Since $\boldsymbol{\theta}$ does not change and $\hat{\mathbf{q}} = \mathbf{q}$, we know that $\hat{\Phi}_i = \Phi_i$ and $y_i^*(\phi,\theta_i,\hat{q}_i) = y_i^*(\phi,\theta_i,q_i)$ for all $i \in \mathcal{N}$. Thus, we have $\hat{d}_i = \int_{\hat{\Phi}_i} y_i^*(\phi,\theta_i,\hat{q}_i)d\hat{\mu} = \int_{\Phi_i} y_i^*(\phi,\theta_i,q_i)dk\mu = kd_i$ for all $i \in \mathcal{N}$. ■

References

[1] T. Basar and R. Srikant. Revenue-maximizing pricing and capacity expansion in a many-users regime. *Proceedings of IEEE INFOCOM*, pages 294–301, 2002.

[2] P. Chander and L. Leruth. The optimal product mix for a monopolist in the presence of congestion effects: A model and some results. *International Journal of Industrial Organization*, 7(4):437–449, 1989.

[3] C.-K. Chau, Q. Wang, and D.-M. Chiu. On the viability of Paris Metro pricing for communication and service networks. *Proceedings of IEEE INFOCOM*, pages 1–9, 2010.

[4] M. Chetty, R. Banks, A. Brush, J. Donner, and R. Grinter. You're capped: understanding the effects of bandwidth caps on broadband use in the home. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3021–3030, 2012.

[5] F. O. I. A. Committee. Policy issues in data caps and usage-based pricing. *Annual Report*, 2013.

[6] W. Dai and S. Jordan. Design and impact of data caps. *IEEE Global Communications Conference*, pages 1650–1656, 2013.

[7] W. Dai and S. Jordan. How do ISP data caps affect subscribers? *Telecommunications Policy Research Conference TPRC*, 2013.

[8] R. Edell and P. Varaiya. Providing Internet access: What we learn from INDEX. *IEEE Network*, 13(5):18–25, 1999.

[9] R. Gibbens, R. Mason, and R. Steinberg. Internet service classes under competition. *IEEE Journal on Selected Areas in Communications*, 18(12):2490–2498, 2000.

[10] P. Hande, M. Chiang, R. Calderbank, and J. Zhang. Pricing under constraints in access networks: Revenue maximization and congestion management. *Proceedings of IEEE INFOCOM*, pages 1–9, 2010.

[11] R. Jain, T. Mullen, and R. Hausman. Analysis of Paris Metro pricing strategy for QoS with a single service provider. *Quality of Service IWQoS*, pages 44–58, 2001.

[12] J. C. L. T. K. Poularakis, I. Pefkianakis. Pricing the last mile: Data capping for residential broadband. *ACM SIGCOMM CoNEXT*, pages 295–306, 2014.

[13] C. Labovitz, D. McPherson, S. Iekel-Johnson, J. Oberheide, and F. Jahanian. Internet inter-domain traffic. *ACM SIGCOMM Computer Communication Review*, 41(4):75–86, 2011.

[14] S. Li, J. Huang, and S.-Y. Li. Revenue maximization for communication networks with usage-based pricing. *Global Telecommunications Conference*, pages 1–6, 2009.

[15] R. T. Ma. Pay-as-you-go pricing and competition in congested network service markets. *Proceedings of the 22nd IEEE International Conference on Network Protocols (ICNP)*, pages 257–268, 2014.

[16] R. T. Ma, J. C. Lui, and V. Misra. On the evolution of the Internet economic ecosystem. *Proceedings of the 22nd International World Wide Web Conference*, pages 849–860, 2013.

[17] M. Morgan. Pricing schemes key in LTE future. *Telecomasia. net. September 12*, 2011.

[18] P. Nabipay, A. Odlyzko, and Z.-L. Zhang. Flat versus metered rates, bundling, and "bandwidth hog". *6th Workshop on the Economics of Networks, Systems, and Computation*, 2011.

[19] A. Odlyzko. Internet pricing and the history of communications. *Computer networks*, 36(5):493–517, 2001.

[20] A. Odlyzko, B. S. Arnaud, E. Stallman, and M. Weinberg. Know your limits: Considering the role of data caps and usage based billing in Internet access service. *Public Knowledge*, April 23, 2012.

[21] D. Reitman. Endogenous quality differentiation in congested markets. *The Journal of Industrial Economics*, 39(6):621–647, 1991.

[22] A. Schatz and S. E. Ante. FCC chief backs usage-based broadband pricing. *Wall Street Journal*, December 2, 2010.

[23] L. Segall. Verizon ends unlimited data plan. *CNN Money. July*, 6, 2011.

[24] H. Shen and T. Basar. Optimal nonlinear pricing for a monopolistic network service provider with complete and incomplete information. *IEEE Journal on Selected Areas in Communications*, 25(6):1216–1223, 2007.

[25] P. Taylor. AT&T imposes usage caps on fixed-line broadband. *Financial Times. March*, 14, 2011.

[26] X. Wang, R. T. Ma, and Y. Xu. The role of data cap in optimal two-part network pricing. *Technical report*, 2015. `http://arxiv.org/pdf/1503.01514.pdf`.

[27] X. Wang, R. T. Ma, and Y. Xu. The role of data cap in two-part pricing under market competition. *4th Workshop on Smart Data Pricing*, 2015.