Exploring Time Flexibility in Wireless Data Plans

Zhiyuan Wang, Lin Gao, *Senior Member, IEEE,* and Jianwei Huang, *Fellow, IEEE*

Abstract—Recently, the mobile network operators (MNOs) are exploring more time flexibility with the rollover data plan, which allows the unused data from the previous month to be used in the current month. Motivated by this industry trend, we propose a general framework for designing and optimizing the mobile data plan with time flexibility. Such a framework includes the traditional data plan, two existing rollover data plans, and a new credit data plan as special cases. Under this framework, we formulate a monopoly MNO's optimal data plan design as a three-stage Stackelberg game: In Stage I, the MNO decides the data mechanism; In Stage II, the MNO further decides the corresponding data cap, subscription fee, and the per-unit fee; Finally in Stage III, users make subscription decisions based on their own characteristics. Through backward induction, we analytically characterize the MNO's profit-maximizing data plan and the corresponding users' subscriptions. Furthermore, we conduct a market survey to estimate the distribution of users' two-dimensional characteristics, and evaluate the performance of different data mechanisms using the real data. We find that a more time-flexible data mechanism increases MNO's profit and users' payoffs, hence improves the social welfare.

Index Terms—Rollover data plan, Three-part tariff, Time flexibility, Game theory.

✦

1 INTRODUCTION

1.1 Background and Motivations

DUE to the increasing competition in the telecommunications market, Mobile Network Operators (MNOs) are under an increasing pressure to increase market shares and improve profits [2]. One approach is to adopt various novel wireless technologies to improve the quality of service (QoS) to attract more subscribers. However, the technology upgrade is often costly and time-consuming. A complementary economical approach is to explore various innovative data pricing schemes to better address heterogeneous user requirements.

Traditionally, most network operators used the **flat-rate** data plans for wireless data services [2], where users pay a fixed fee for unlimited monthly data usage. Later in 2010, the Federal Communications Commission (FCC) and Cisco backed usage-based pricing to penalize those heavy users and manage the traffic. Therefore, MNOs started to adopt the **usage-based** pricing scheme, where the subscribers are charged based on their actual data consumptions. A widely used form of usage-based plan adopted by many MNOs today is the **three-part tariff** plan, which consists of a monthly subscription fee, a data cap within which there is no additional cost of usage, and a linear unit price for any data consumption exceeding the data cap.

Recent years have witnessed many MNOs exploring the **time flexibility** in their data plans to further increase their market competitiveness. For example, the *rollover data plan*,

which allows the unused part of the data cap from the previous month to be used in the current month, has been implemented by many MNOs, e.g., AT&T [3], T-Mobile [4], and China Mobile [5]. Such a rollover plan reduces the users' uncertainty due to stochastic data demands, and offers users more time flexibility. This motivates us to ask the first key question in this paper:

Question 1. Who will benefit more from the introduced time flexibility, the MNO or users?

Although centering around the same core idea, various rollover data plans in practice can be quite different in terms of the consumption priority and the expiration time. For example, AT&T specifies that the rollover data from the previous month will be used after the current monthly data cap is fully consumed [3], while China Mobile specifies that the rollover data will be used before consuming the current monthly data cap [5]. As for the expiration time, both AT&T and China Mobile require the rollover data to expire after one month, while T-Mobile allows subscribers to accumulate their rollover data over several months [4].

We observe that a common feature of various rollover data plans is that a user can only use the "remaining" data cap from the previous month(s). This motivates us to propose a *credit data plan* that inversely allows users to "borrow" their data quota from future months.[1] In fact, the rollover and credit data mechanisms represent two different ways of exploring the data dynamics across the time dimension: backward and forward. This motivates us to ask the second and the third key questions in this paper:

Question 2. Which data mechanism is the most time-flexible?

Question 3. Which data mechanism should the MNO adopt?

---

1. The telecom market in many countries is based on the real-name registration, hence it is difficult for a user to keep borrowing data and then stop the subscription without paying back. Moreover, the Data Bank platform [6] implemented by China Unicom allows users to borrow data from the MNO, which exhibits a similar idea to the credit data plan.

Furthermore, the MNOs profit from mobile data services through carefully choosing and optimizing their mobile data plans. Even though the MNOs have implemented different versions of rollover data plans in practice, there is no systematical understanding on how the time flexibility affects the optimal data plan design. This motivates us to ask the fourth key question in this paper:

Question 4. What is the impact of time flexibility on the MNO's optimal data cap, subscription fee, and per-unit fee?

In this paper we will study and evaluate these innovative data mechanisms and reveal the impact of time flexibility in a comprehensible way. We hope that our results in this paper could pave the way for the MNOs to better implement and the public to better understand these time-flexible mobile data plans.

1.2 Solutions and Contributions

In this paper, we study the optimization of the three-part tariff data plan with time flexibility in a monopoly market, and consider four data mechanisms which are different in the *special data* (rollover or credit data) and the *consumption priority*.

We formulate the monopoly MNO's optimal data plan design as a three-stage Stackelberg game, with the MNO as the leader and users as followers. Specifically, the MNO decides which kind of data mechanism to adopt in Stage I, then further decides its profit-maximizing data cap and the corresponding subscription fee and per-unit fee in Stage II. Finally, users make their subscription decisions to maximize their payoffs in Stage III.

To the best of our knowledge, this is the first paper that systematically studies the MNO's optimal three-part tariff plan with time flexibility. The main results and contributions of this paper are summarized as follows:

- *Systematic Study of Data Mechanisms with Time Flexibility:* We propose a general framework that includes the traditional data mechanism and three innovative data mechanisms with rollover or credit data as special cases. Based on such a unified framework, we further study the optimal design for mobile data plan with time flexibility.
- *Three-Stage Decision Model:* We model and analyze the interactions between the MNO and users as a three-stage Stackelberg game. Despite the complexity of the model, we are able to fully characterize the user subscription in Stage III, the MNO's optimal data cap and pricing strategy in Stage II, and the optimal data mechanism choice in Stage I.
- *User Subscription:* We consider users' heterogeneity in the data valuation and the network substitutability. We find that under the optimal data plan, the profit-maximizing MNO admits subscribers based only on their data valuations, while treating users of different network substitutability identically.
- *Optimal Data Plan:* We study the impact of MNO's quality of service (QoS), operational cost, and capacity cost on its optimal data plan. Our analysis reveals a counter-intuitive insight: a better time flexibility dose not necessarily lead to a smaller data cap; the

data cap can be larger if the MNO is weak with a poor QoS and large costs.
- *Performance Evaluation:* We conduct a market survey to estimate the statistical distribution of users' data valuation and network substitutability. The simulations based on the empirical data further reveal that both MNO and users can benefit from the time flexibility. The MNO benefits more than users if the MNO provides good services and experiences small costs. Otherwise, users will benefit more from the time flexibility than the MNO.

The remainder of this paper is organized as follows. In Section 2, we review the related works. Section 3 introduces our system model and the three-stage game. Section 4 presents the four data mechanisms with time flexibility in detail. In Section 5, we analyze the three-stage decision model through backward induction. Section 6 presents the numerical results and Section 7 concludes this paper.

2 LITERATURE REVIEW

The optimal design of mobile data plan has been extensively studied in the literature. The early studies focused on the debate between flat-rate and usage-based schemes [7]. After introducing the data cap, Dai *et al.* in [8] demonstrated that heavy users would pay for their usage, while light users would benefit from it. Then Wang *et al.* in [9] studied the optimization of the three-part-tariff in the congestion-prone network. Xiong *et al.* in [10] focused on the sponsored content and analyzed a Stackelberg game pricing model on the MNO's data plan optimization. Zheng *et al.* in [11] studied the dynamics of users' data consumption through a dynamic programming formulation. However, the above studies did not consider various forms of flexibility introduced in recent mobile data plans, including the time dimension [12], [13], user dimension [14], [15], [16], and location dimension [17], [18].

Time flexibility corresponds to the rollover data plan in practice. Despite of the increasing popularity of the rollover data plan, the related theoretical study just emerged very recently. As far as we know, there existed only two related works before this work. Specifically, Zheng *et al.* in [12] compared the rollover data plan with a traditional three-part tariff, and found that the moderately price-sensitive users can benefit from subscribing to the rollover data plan. Wei *et al.* in [13] further looked at the choice of expiration time of the rollover data and analyzed the impact of the rollover period lengths through a contract-theoretic approach.

User flexibility corresponds to the shared data plan and data trading. Sen *et al.* in [14] introduced an analytical framework for studying the economics of shared data plans. Zheng *et al.* in [15] examined the "2CM" data trading market launched by China Mobile Hong Kong, and Yu *et al.* in [16] further analyzed users' realistic trading behaviors using prospect theory.

Location flexibility corresponds to the global data services, e.g., Skype Wi-Fi and Uroaming. Duan *et al.* in [17] studied how the global providers work with many local providers to promote the global mobile data services, and examined the flat-rate and usage-based schemes. Ma *et al.* in [18] proposed an optimal design of time and location aware mobile data

TABLE 1: $\mathcal{T} \triangleq \{Q, \Pi, \pi, \kappa\}, \ \kappa \in \{0, 1, 2, 3\}$

Data Mechanism	Special Data	Surplus or Deficit	Consumption Priority	Effective Cap $Q_\kappa^e(\tau)$
$\kappa = 0$	None	$\tau = 0$	Cap	Q
$\kappa = 1$	Rollover data	$\tau \in [0, Q]$	Cap$\Rightarrow$Rollover	$Q + \tau \in [Q, 2Q]$
$\kappa = 2$	Rollover data	$\tau \in [0, Q]$	Rollover$\Rightarrow$Cap	$Q + \tau \in [Q, 2Q]$
$\kappa = 3$	Credit data	$\tau \in [-Q, 0]$	Cap$\Rightarrow$Credit	$2Q + \tau \in [Q, 2Q]$

Stage I
The MNO decides the data mechanism κ to implement in the three-part tariff data plan.

$\Downarrow$

Stage II
The MNO decides the data cap Q, subscription fee Π and the per-unit fee π to maximize its profit.

$\Downarrow$

Stage III
Each type-(θ, β) user makes his subscription choice based on his expected payoff under $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$.

Fig. 1: Three-Stage Stackelberg Game.

pricing, incentivizing users to smoothen their traffic and reduce network congestion.

Note that each of the above three dimensions deserves substantial studies and further explorations. Our focus is the optimization of the three-part tariff mobile data plan with time flexibility (which was not studied in [12], [13]).

3 SYSTEM MODEL

In this paper, we consider a *monopoly* market with a single MNO who provides mobile data services for heterogeneous users.[2] The MNO designs a mobile data plan to maximize its profit, and each user decides whether to subscribe to the MNO to maximize his payoff.

We formulate the economic interactions between the MNO and the mobile users as a three-stage Stackelberg game, as shown in Fig. 1. The MNO is the Stackelberg leader: it first decides the data mechanism to be implemented within its three-part tariff data plan in Stage I, then decides the data cap, the subscription fee, and the per-unit fee in Stage II.[3] Finally, the users make their subscription decisions to maximize their payoffs in Stage III.

Next we first present a unifying framework of different mobile data plans, then introduce the model in more details from the perspectives of users and the MNO, respectively.

3.1 Mobile Data Plans

The *three-part tariff data plan with time flexibility* can be characterized by the tuple $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$, where the subscriber pays a fixed lump-sum subscription fee Π for the

---

2. There are multiple competitive MNOs in the practical market. The analysis for the competitive market requires a comprehensive understanding on each MNO's optimal data plan design. Due to space limit, in this paper we focus on the monopoly case. We have reported some preliminary results on the duopoly competition in [19].

3. In practice, an MNO adopts a data mechanism (i.e., rollover, credit, or the traditional one) for a relatively long time period (e.g., three or five years), while has the flexibility of updating data mechanism parameters (i.e., the data cap, monthly subscription fee, and the per-unit fee) more often (e.g., on a yearly basis). The three-stage formulation captures the MNO's different decisions at different time scales.

data usage up to the cap Q, beyond which the subscriber pays an overage fee π for each unit of additional data consumption. Here κ represents different *data mechanisms* that gives the subscriber different degrees of time flexibility on their data consumption over time.

Next we first introduce the four data mechanisms, then demonstrate that the pure usage-based data plan and the flat-rate data plan are special cases of the tuple $\mathcal{T}$.

3.1.1 Four Data Mechanisms

We consider four data mechanisms indexed by $\kappa \in \{0, 1, 2, 3\}$. The key differences among the different data mechanisms are the *special data* and the *consumption priority*. To be more specific, the special data could be the rollover data inherited from the previous month or the credit data that can be borrowed from the next month, both of which can enlarge a subscriber's *effective data cap* (within which no overage fee involved) in the current month. Moreover, the consumption priority of the special data and the current monthly data cap further affects how much the effective data cap can be enlarged.

We summarize the key differences of the four data mechanisms in Table 1. Here we use τ to denote a user's data surplus ($\tau > 0$) or data deficit ($\tau < 0$) at the beginning of a month, and use $Q_\kappa^e(\tau)$ to denote the corresponding effective data cap. More specifically,

- The case of $\kappa = 0$ denotes the **traditional data mechanism** without time flexibility. The subscriber has no data surplus or deficit, and the effective cap of each month is $Q_0^e(\tau) = Q$;
- The case of $\kappa = 1$ denotes the **rollover data mechanism** offered by AT&T. The rollover data τ from the previous month is consumed *after* the current monthly data cap and expires at the end of the current month. Thus the effective cap of the current month is $Q_1^e(\tau) = Q + \tau$;
- The case of $\kappa = 2$ denotes the **rollover data mechanism** offered by China Mobile. The rollover data τ from the previous month is consumed *prior* to the current monthly data cap Q and expires at the end of the current month. Thus the effective cap of the current month is $Q_2^e(\tau) = Q + \tau$;
- The case of $\kappa = 3$ denotes the **credit data mechanism** proposed in this paper. The credit data is from the next month's data cap Q,[4] which is consumed *after* the current monthly data cap Q (with a data deficit τ from the previous month). Thus the effective cap of the current month is $Q_3^e(\tau) = 2Q + \tau$;

As mentioned above, the time flexibility can enlarge the subscriber's effective data cap. According to Table 1, the

---

4. Since users can only borrow data from the next one month's data cap, Q is the maximal amount that can be borrowed.

effective data cap of the traditional data mechanism (i.e., $\kappa = 0$) is always Q; while the potential maximal value of the effective data cap is $2Q$ for the rollover and credit data mechanisms (i.e., $\kappa = 1, 2, 3$). The larger the effective data cap is, the less overage usage is incurred, which will further change the users' subscription decisions.

Table 2 provides a numerical example with five months' data consumptions and the corresponding payments under the four data mechanisms. Among the four schemes, the last two schemes (i.e., $\kappa = 2, 3$) lead to the same least total payment of \$250, and the first scheme (i.e., $\kappa = 0$) leads to the maximum total payment of \$280. Is this a coincidence? We will further discuss it in Section 4.

3.1.2 Special Cases of Three-Part Tariff

Next we illustrate that the pure usage-based data plan and the flat-rate data plan are two special cases of the tuple $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$, $\kappa \in \{0, 1, 2, 3\}$.

Under the *pure usage-based* data plan, the subscriber has a zero monthly data cap and only needs to pay a usage-based fee for each unit of data consumption. Therefore, we can denote the pure usage-based data plan as $\mathcal{T}_\mathrm{p} = \{0, 0, \pi_\mathrm{p}, \mathtt{Na}\}$, where $\mathtt{Na}$ represents that the data mechanism index κ has no impact under the pure usage-based data plan (as there is no data cap).

Under the *flat-rate* data plan, the subscriber pays a fixed monthly subscription fee for unlimited data usage without any additional overage fee, thus the per-unit fee π has no impact. Strictly speaking, the data cap of the flat-rate data plan is infinity. However, in this paper we model the user's data demand as a random variable $d \in \{0, 1, 2, ..., D\}$ (to be defined in Section 3.2), where D is the maximal data demand. Therefore, in practice it is enough to set the data cap to be no less than D to achieve the effect of a flat-rate plan. Accordingly, we denote the flat-rate data plan as $\mathcal{T}_\mathrm{f} = \{Q_\mathrm{f}, \Pi_\mathrm{f}, \mathtt{Na}, \mathtt{Na}\}$, where $Q_\mathrm{f} \geq D$, and $\mathtt{Na}$ represents that the per-unit fee π and the data mechanism κ have no impact.

Our later analysis on the MNO's optimal data plan is based on the general tuple $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$. Meanwhile, we will characterize the conditions under which the MNO's optimal three-part tariff with time flexibility would degenerate into the pure usage-based data plan $\mathcal{T}_\mathrm{p}$ or the flat-rate data plan $\mathcal{T}_\mathrm{f}$.

3.2 User Model

Next we introduce a user's three characterizations: the data demand d, the data valuation θ, and the network substitutability β. Based on these, we derive a user's expected payoff.

First, we model the user's monthly data demand d as a discrete random variable with a probability mass function $f(d)$, a mean value of $\bar{d}$, and a finite integer support $\{0, 1, 2, ..., D\}$. Here the data demand is measured in the minimum data unit (e.g, 1KB or 1MB according to the MNO's billing practice).

Second, we follow [7] by denoting θ as a user's utility from consuming one unit of data, i.e., the user's data valuation. According to our market survey conducted in mainland China, θ falls into the range between 10 RMB/GB and 60 RMB/GB with a large probability. We will further discuss it in Section 6.1.

TABLE 2: Numerical Example for $\kappa = 0, 1, 2, 3$.

Month		Jan.	Feb.	Mar.	Apr.	May	Total
Data Consumption		2GB	2GB	4GB	4GB	1GB	13GB
$\kappa = 0$	Payment	\$50	\$50	\$65	\$65	\$50	**\$280**
$\kappa = 1$	τ	0	1GB	1GB	0	0	**\$265**
	Payment	\$50	\$50	\$50	\$65	\$50	
$\kappa = 2$	τ	0	1GB	2GB	1GB	0	**\$250**
	Payment	\$50	\$50	\$50	\$50	\$50	
$\kappa = 3$	τ	0	0	0	-1GB	-2GB	**\$250**
	Payment	\$50	\$50	\$50	\$50	\$50	

Here the data cap is 3GB, the subscription fee is \$50, the overage fee is \$15/GB, and τ denotes the data surplus or deficit of each month.

Third, we further explore a user's behavior change when he reaches the effective data cap $Q_\kappa^e(\tau)$, since further data consumption leads to an additional payment. Although the user will still continue to consume data in this case, he will rely more heavily on other alternative networks (such as Wi-Fi). As in [14], we use the network substitutability $\beta \in [0, 1]$ to denote the fraction of overage usage shrink. A larger β value represents more overage usage cut (thus a better network substitutability).

A user's mobility pattern can significantly influence the availability of alternative networks. Hence, different users usually have heterogeneous network substitutabilities. For example, a businessman who is always on the road may have a poor network substitutability (hence has a small value of β); while a student can take the advantage of the school Wi-Fi network (hence has a large value of β). According to our market survey, β falls into the range between 0.7 and 1 with a large probability. We will further discuss it in Section 6.1.

Without loss of generality, we normalize the total user population size to one in the rest of the paper. We assume that users are homogeneous in the data demand distribution $f(d)$, [5] and investigate the heterogeneity in the data valuation θ and network substitutability β. Therefore, we model each user by the two-dimensional characteristics (θ, β), and define the whole user market as $\mathcal{M} = \{(\theta, \beta) : 0 \leq \theta \leq \theta_{\max}, 0 \leq \beta \leq 1\}$ with probability density functions $h(\theta)$ and $g(\beta)$, since our market survey shows that θ and β are independent with the Pearson correlation coefficient less than 0.05. Furthermore, we denote $\Psi(\mathcal{T}) \subseteq \mathcal{M}$ as the subscriber set, i.e., the type-(θ, β) user subscribes to the MNO if and only if $(\theta, \beta) \in \Psi(\mathcal{T})$.

A subscriber's payoff is the difference between his utility and total payment. More specifically, for a type-(θ, β) subscriber with d units of data demand and an effective data cap $Q_\kappa^e(\tau)$, his actual data usage is $d - \beta[d - Q_\kappa^e(\tau)]^+$, where $[x]^+ = \max\{0, x\}$. Moreover, we use ρ to represent the MNO's *average* quality of service (QoS).[6] Mathematically, ρ is a utility multiplicative coefficient, thus the subscriber's utility is $\rho\theta(d - \beta[d - Q_\kappa^e(\tau)]^+)$. In addition, the subscriber's

5. According to the statistical analysis in [20], [21], users' monthly demand can be estimated by a log-normal distribution. For analysis tractability, we consider a homogeneous demand distribution. In the future, we will consider the heterogeneous case and collect users' data usage records to estimate the demand distributions as in [11], [15].

6. In practice, an MNO's wireless data service depends on the network congestion, which has been studied before (e.g., [2], [7]). In this work, instead of modeling the detailed congestion-aware control, we are more interested in the *long-term average quality* of the MNO's wireless data service.

total payment consists of the subscription fee Π and the overage payment $\pi(1-\beta)[d-Q_\kappa^e(\tau)]^+$. Therefore, *the payoff of a type-(θ,β) subscriber* with d units of data demand and an effective cap $Q_\kappa^e(\tau)$ is given by

$$S(\mathcal{T},\theta,\beta,d,\tau)$$
$$=\rho\theta\left(d-\beta[d-Q_\kappa^e(\tau)]^+\right)-\pi(1-\beta)[d-Q_\kappa^e(\tau)]^+-\Pi, \tag{1}$$

where the data demand d and the data surplus (or deficit) τ are two random variables that change in each month. After taking the expectation over d and τ, we obtain a *type-(θ,β) subscriber's expected monthly payoff* as

$$\bar{S}(\mathcal{T},\theta,\beta)=\mathbb{E}_{d,\tau}\left[S\left(\mathcal{T},\theta,\beta,d,\tau\right)\right]$$
$$=\rho\theta\left[\bar{d}-\beta A_\kappa(Q)\right]-\pi(1-\beta)A_\kappa(Q)-\Pi. \tag{2}$$

Here $A_\kappa(Q)$ is the type-$(\theta,0)$ subscriber's *expected monthly overage data consumption* under the data mechanism κ, defined as follows:

$$A_\kappa(Q)=\mathbb{E}_{d,\tau}\left\{[d-Q_\kappa^e(\tau)]^+\right\}$$
$$=\sum_\tau\sum_d[d-Q_\kappa^e(\tau)]^+f(d)p_\kappa(\tau), \tag{3}$$

where the summation range of d is from 0 to the maximal demand D, while the range of τ depends on the data mechanism κ, which is given in (14), (17), and (20). The $p_\kappa(\tau)$ represents the probability mass function of τ under the data mechanism κ. Moreover, $p_\kappa(\cdot)$ is the key difference among the four data mechanisms, since the data mechanism κ affects a subscriber's payoff through $A_\kappa(Q)$ in (2). In Section 4, we will further explain how to compute $p_\kappa(\tau)$ and $A_\kappa(Q)$ in detail.

Furthermore, the *expected total payoff of the whole market* under $\mathcal{T}$ is the integration over all the subscribers in $\Psi(\mathcal{T})$, as follows:

$$\tilde{S}(\mathcal{T})=\iint_{\Psi(\mathcal{T})}\bar{S}(\mathcal{T},\theta,\beta)h(\theta)g(\beta)d\theta d\beta. \tag{4}$$

So far, we have introduced users' characteristics and derived their payoffs. Next we move on to model the profit-maximizing MNO.

3.3 MNO Model

In the following we formulate the MNO's revenue, cost, and profit, respectively.

3.3.1 *MNO's Revenue*

The MNO's revenue obtained from a subscriber consists of the subscription fee and the possibly overage fee. Therefore, the *MNO's revenue from a type-(θ,β) subscriber* with d units of data demand and an effective cap $Q_\kappa^e(\tau)$ is

$$R(\mathcal{T},\theta,\beta,d,\tau)=\pi(1-\beta)\left[d-Q_\kappa^e(\tau)\right]^++\Pi. \tag{5}$$

Since d and τ are two random variables that change in each month, we take the expectation and obtain the *MNO's expected monthly revenue from a type-(θ,β) subscriber* as follows:

$$\bar{R}(\mathcal{T},\theta,\beta)=\mathbb{E}_{d,\tau}\left[R\left(\mathcal{T},\theta,\beta,d,\tau\right)\right]$$
$$=\pi(1-\beta)A_\kappa(Q)+\Pi, \tag{6}$$

where $(1-\beta)A_\kappa(Q)$ is the type-(θ,β) subscriber's expected overage usage. Again we will provide more details on

$A_\kappa(Q)$ in Section 4. Moreover, the *MNO's expected total revenue* from the entire market under $\mathcal{T}$ is

$$\tilde{R}(\mathcal{T})=\iint_{\Psi(\mathcal{T})}\bar{R}(\mathcal{T},\theta,\beta)h(\theta)g(\beta)d\theta d\beta. \tag{7}$$

3.3.2 *MNO's Cost*

In reality, the MNO's cost is quite a complicated function that is related to many factors [2]. In this paper, we consider two kinds of costs experienced by the MNO, i.e., the capacity cost and the operational cost.

The MNO's capacity cost mainly arises from its capital expenditure (CapEx), the investment on its network capacity. In reality, the data cap helps the MNO manage the network congestion and ration the scarce network capacity [22], and most MNOs imposed the data cap to alleviate the network congestion [23]. Therefore, once the MNO decides a data cap to be offered in the market, it should make sure a corresponding network capacity is in place to support the traffic. Motivated by this phenomenon, we model the *MNO's capacity cost* as an increasing function $J(Q)$ on the data cap Q. Intuitively, a larger data cap corresponds to a more severe network congestion, which requires more investment on the network capacity in advance. The MNO's capacity investment affects the network congestion, which will change its QoS and eventually affect the user utility of consuming data. Here we will not incorporate the congestion-aware formulation in this work, but refer interested readers to the related studies in [2], [7].

Furthermore, the MNO's operational expense (OpEx) mainly arises from the system management. After the MNO decides the mobile data plan to implement in the market, the subscribers' total data consumption will affect the MNO's operational expense. Specifically, the *expected total data consumption* from the whole market is

$$L(\mathcal{T})=\iint_{\Psi(\mathcal{T})}\left[\bar{d}-\beta A_\kappa(Q)\right]h(\theta)g(\beta)d\theta d\beta. \tag{8}$$

For analysis tractability, we follow [24] by considering a linear operational cost, and denote c as the marginal operational cost from unit data consumption.[7] Accordingly, the *MNO's operational cost* is $L(\mathcal{T})\cdot c$.

Putting the capacity cost and the operational cost together, we compute the *MNO's expected total cost* as follows:

$$\tilde{C}(\mathcal{T})=L(\mathcal{T})\cdot c+J(Q). \tag{9}$$

3.3.3 *MNO's Profit*

The MNO's profit is the difference between its revenue and cost. Hence the *MNO's expected total profit* under $\mathcal{T}$ is

$$\tilde{W}(\mathcal{T})=\tilde{R}(\mathcal{T})-\tilde{C}(\mathcal{T}). \tag{10}$$

So far we have introduced the four data mechanisms, users' payoffs, and MNO's profit. In the following, we first compare the degrees of time flexibility offered by the four data mechanisms in Section 4, then study the three-stage game in Section 5.

---

7. Such a linear-form cost has been widely used to model an operator's operational cost (e.g., [24], [25]).

Fig. 2: Transition from τ (data surplus or deficit from the previous month) to τ' (data surplus or deficit of the next month).

4 DATA MECHANISMS AND TIME FLEXIBILITY

Recall that the numerical example in Table 2 shows that the total payments for $\kappa = 2, 3$ are the same and the least, while the payment for $\kappa = 0$ is the most. In this section we will demonstrate that *it is not a coincidence, but a general conclusion reflecting the data mechanisms' degrees of time flexibility*.

In the following, we introduce how to compute $A_\kappa(Q)$ in Section 4.1, then answer **Question 2** (i.e., which data mechanism is the most time-flexible) in Section 4.2.

4.1 Data Mechanisms

As mentioned in Section 3.2, the key difference among the four data mechanisms is the distribution of the subscriber's data surplus or deficit $p_\kappa(\tau)$, which further determines $A_\kappa(Q)$ according to (3). Particularly, for $\kappa = 0$, the data surplus or deficit is always zero, i.e., $\tau = 0$, since it does not offer subscribers any special data. However, for $\kappa \in \{1, 2, 3\}$, we need to consider the data demand dynamic between successive months.

We illustrate the transition of users' data surplus or deficit τ between two successive months in Fig. 2. Specifically, the horizontal axis corresponds to users' random data demand $d \in [0, D]$, and the vertical axis represents users' data surplus or deficit τ' for the next month (given his data surplus or deficit τ in the current month and the data demand d). The differences among the three red curves in Fig. 2 indicate the differences among the three data mechanisms $\kappa \in \{1, 2, 3\}$.

In the following, we analyze the distribution $p_\kappa(\tau)$ and compute $A_\kappa(Q)$ under the four data mechanisms.

4.1.1 Traditional Data Mechanism $\kappa = 0$

For a $\mathcal{T} = \{Q, \Pi, \pi, 0\}$ subscriber, there is no special data to use, i.e., $\tau = 0$ and $Q_0^e(\tau) = Q$. Thus we only need to take the expectation over the data demand d. Thus $A_0(Q)$ is

$$A_0(Q) = \sum_{d=0}^{D} [d - Q]^+ f(d). \quad (11)$$

4.1.2 Rollover Data Mechanism $\kappa = 1$

For a $\mathcal{T} = \{Q, \Pi, \pi, 1\}$ subscriber, the special data is the rollover data from the previous month, which is consumed *after* the current monthly data cap. Therefore, the effective data cap consists of the monthly data cap and the rollover data surplus $\tau \in [0, Q]$, i.e., $Q_1^e(\tau) = Q + \tau$. Fig. 2(a) plots the rollover data to the next month, denoted by τ', versus the subscriber's data demand d in the current month. Thus we know

$$\tau' = \begin{cases} Q - d, & \text{if } d < Q, \\ 0, & \text{if } d \geq Q. \end{cases} \quad (12)$$

Note that the rollover data to the next month τ' only depends on the subscriber's monthly data cap Q and the data demand d, and is independent of the data surplus τ from the previous month. Therefore, the probability mass function $p_1(\tau)$ is

$$p_1(\tau) = \begin{cases} f(Q - \tau), & \text{if } \tau \in (0, Q], \\ \sum_{d=Q}^{D} f(d), & \text{if } \tau = 0. \end{cases} \quad (13)$$

Then we need to take the expectation over the data demand d and the rollover data surplus τ to compute the expected overage usage $A_1(Q)$, as follows:

$$A_1(Q) = \sum_{\tau=0}^{Q} \sum_{d=0}^{D} [d - Q_1^e(\tau)]^+ f(d) p_1(\tau). \quad (14)$$

4.1.3 Rollover Data Mechanism $\kappa = 2$

For a $\mathcal{T} = \{Q, \Pi, \pi, 2\}$ subscriber, the special data is the rollover data from the previous month, which is consumed *prior* to the current monthly data cap. Therefore, the effective data cap is the same as that for $\kappa = 1$, i.e., $Q_2^e(\tau) = Q + \tau$. However, the rollover data is consumed *prior* to the monthly cap. As showed in Fig. 2(b), we know

$$\tau' = \begin{cases} Q, & \text{if } d \in [0, \tau], \\ Q + \tau - d, & \text{if } d \in (\tau, Q + \tau), \\ 0, & \text{if } d \in [Q + \tau, D]. \end{cases} \quad (15)$$

It is notable that the rollover data to the next month τ' depends on the monthly data cap Q, data demand d, and the rollover data surplus τ from the previous month, resulting in a Markov property on the rollover data surplus τ. The one-step transition probability of the rollover data surplus τ is given by

$$p_2(\tau, \tau') = \begin{cases} \sum_{d=0}^{\tau} f(d), & \text{if } \tau' = Q, \\ f(Q + \tau - \tau'), & \text{if } \tau' \in (0, Q), \\ \sum_{d=Q+\tau}^{D} f(d), & \text{if } \tau' = 0. \end{cases} \quad (16)$$

Then we can derive the stationary distribution of the rollover data surplus τ, denoted by $p_2(\tau)$, according to the above transition probability [26]. Thus $A_2(Q)$ is given by

$$A_2(Q) = \sum_{\tau=0}^{Q} \sum_{d=0}^{D} [d - Q_2^e(\tau)]^+ f(d) p_2(\tau). \quad (17)$$

4.1.4 Credit Data Mechanism $\kappa = 3$

For a $\mathcal{T} = \{Q, \Pi, \pi, 3\}$ subscriber, the special data is the credit data borrowed from the next month, which is used after the current monthly data cap. Therefore, the effective data cap of a subscriber with a data deficit $\tau \in [-Q_3, 0]$ consists of the remaining current monthly data cap (with a deficit τ) and the maximum credit data that he can borrow from the next month (which is Q), i.e., $Q_3^e(\tau) = 2Q + \tau$.

According to Fig. 2(c), the data deficit in the next month, denoted by τ', is given by

$$\tau' = \begin{cases} 0, & \text{if } d \in [0, Q+\tau], \\ Q+\tau-d, & \text{if } d \in (Q+\tau, 2Q+\tau), \\ -Q, & \text{if } d \in [2Q+\tau, D]. \end{cases} \quad (18)$$

Similar to the case when $\kappa = 2$, we note that the data deficit τ' in the next month depends on the monthly data cap Q, the data demand d, and the data deficit τ in the current month, which indicates a Markov property on the data deficit τ for $\kappa = 3$. The corresponding one-step transition probability of the data deficit τ is

$$p_3(\tau, \tau') = \begin{cases} \sum_{d=0}^{Q+\tau} f(d), & \text{if } \tau' = 0, \\ f(Q+\tau-\tau'), & \text{if } \tau' \in (-Q, 0), \\ \sum_{d=2Q+\tau}^{D} f(d), & \text{if } \tau' = -Q. \end{cases} \quad (19)$$

Similarly we can derive the stationary distribution $p_3(\tau)$ of the data deficit τ and compute $A_3(Q)$ as follows

$$A_3(Q) = \sum_{\tau=-Q}^{0} \sum_{d=0}^{D} [d - Q_3^e(\tau)]^+ f(d) p_3(\tau). \quad (20)$$

Now that we have demonstrated how to compute $A_\kappa(Q)$ under the four data mechanisms, next we will compare the degree of time flexibility based on $A_\kappa(Q)$.

4.2 Time Flexibility

In the following we compare the degrees of time flexibility among the four data mechanisms and answer *Question 2*.

Definition 1 (Time Flexibility). Consider two data mechanisms $i, j \in \{0, 1, 2, 3\}$. The data mechanism i has a better time flexibility than the data mechanism j, denoted by $\mathcal{F}_i > \mathcal{F}_j$, if and only if for an arbitrary data demand distribution $f(d)$, we have $A_i(Q) < A_j(Q)$ for all $Q \in (0, D)$.

Definition 1 uses the type-$(\theta, 0)$ subscriber's expected overage data consumption $A_\kappa(Q)$ to indicate the time flexibility of the data mechanism κ. Intuitively, the better time flexibility the data mechanism offers, the less overage usage is incurred by its subscribers under the same data cap Q for an arbitrary data demand distribution $f(d)$. Note that we require $Q \in (0, D)$ in the definition, this is because that the three-part tariff with time flexibility degenerates into the pure usage-based data plan if $Q = 0$ or the flat-rate data plan if $Q \geq D$. In these two extreme cases, any data mechanism κ has no impact.

Lemma 1 summarizes the time flexibility of the four data mechanisms. The proof is given in Appendix A.

Lemma 1. For an arbitrary data demand distribution $f(d)$, we have $A_0(Q) > A_1(Q) > A_2(Q) = A_3(Q)$ for all $Q \in (0, D)$. Therefore, the four data mechanisms' degrees of time flexibility satisfy $\mathcal{F}_0 < \mathcal{F}_1 < \mathcal{F}_2 = \mathcal{F}_3$.

Lemma 1 provides the answer to *Question 2* that we mentioned in Section 1. The rollover data mechanism offered by China Mobile (i.e., $\kappa = 2$) and our proposed credit data mechanism (i.e., $\kappa = 3$) are both the most time-flexible. The traditional data mechanism (i.e., $\kappa = 0$) is the least time-flexible.

The reason why $\mathcal{F}_2 = \mathcal{F}_3$ is twofolds. First, the two data mechanisms can expand the effective data cap with the same intensity, i.e., $Q_\kappa^e(\tau) \in [Q, 2Q]$ for $\kappa \in \{2, 3\}$. Second, the consumption priorities of the two data mechanisms specify that subscribers should first consume the earlier data, i.e., the rollover data prior to the current monthly data cap for $\kappa = 2$, and the current monthly data cap prior to the credit data for $\kappa = 3$.

The reason why $\mathcal{F}_1 < \mathcal{F}_2$ is because of the irregular consumption priority for $\kappa = 1$. Recall that the rollover data mechanism offered by AT&T (i.e., $\kappa = 1$) requires that the current monthly data cap is consumed prior to the rollover data from the previous month, which means the later data (i.e., current monthly data cap) would be consumed prior to the earlier data (i.e., rollover data from the previous month). Such an irregular consumption priority prevents subscribers from fully utilizing their data quota in the long run, hence reduces the degree of time flexibility. Nevertheless, it is still time flexible than the traditional data mechanism, i.e., $\mathcal{F}_0 < \mathcal{F}_1$.

So far we have compared the degree of time flexibility among the four data mechanisms. Next in Section 5, we analyze the three-stage game.

5 BACKWARD INDUCTION OF THE THREE-STAGE STACKELBERG GAME

In this section, we study the Subgame Perfect Equilibrium (SPE, or simply referred to as equilibrium in this paper) of the three-stage Stackelberg game by backward induction.

5.1 User's Subscription in Stage III

In Stage III, each user makes his subscription choice given the data plan $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$ offered by the MNO.

The type-(θ, β) user has two choices. If he does not subscribe, his payoff will be zero.[8] Hence the user will subscribe to the MNO if and only if $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$ brings him a non-negative expected monthly payoff, i.e., $\bar{S}(\mathcal{T}, \theta, \beta) \geq 0$. Theorem 1 presents the MNO's market share under the data plan $\mathcal{T}$. The proof is given in Appendix B.

Theorem 1 (Market Share). The MNO's market share under data plan $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$ is

$$\Psi(\mathcal{T}) = \{(\theta, \beta) : \Theta(\mathcal{T}, \beta) \leq \theta \leq \theta_{\max}, 0 \leq \beta \leq 1\}, \quad (21)$$

where $\Theta(\mathcal{T}, \beta)$ is referred to as the *threshold valuation* under $\mathcal{T}$ and β, which is given by

$$\Theta(\mathcal{T}, \beta) \triangleq \frac{1}{\rho} \left[\pi + \frac{\pi \left[\bar{d} - A_\kappa(Q) \right] - \Pi}{\beta A_\kappa(Q) - \bar{d}} \right]. \quad (22)$$

Based on Theorem 1, we further summarize the impacts of users' characteristics θ and β on their subscription decisions in *Corollary 1* and *Corollary 2*, respectively.

Corollary 1. [Impact of Data Valuation] Given the network substitutability β, it is more likely for a user to subscribe to the MNO as his data valuation θ increases.

8. Here we normalize the non-subscription payoff to be zero. If the user has other options, for example, relying purely on Wi-Fi networks, the non-subscription payoff can be positive. Our analysis will still go through in that case with a simple constant shift.

Fig. 3: Illustration of different market partitions. Gray region: the subscribers $\Psi(\mathcal{T})$.

Corollary 1 indicates that higher valuation users are more likely to subscribe to the MNO. However, *Corollary 2* shows that the impact of network substitutability is more complicated.

Corollary 2. [Impact of Network Substitutability] Given the data valuation θ, there are three possibilities for the impact of network substitutability β:

- Case 1 ($\pi[\bar{d} - A_\kappa(Q)] > \Pi$): As the network substitutability improves, users are *more likely* to subscribe to the MNO.
- Case 2 ($\pi[\bar{d} - A_\kappa(Q)] < \Pi$): As the network substitutability improves, users are *less likely* to subscribe to the MNO.
- Case 3 ($\pi[\bar{d} - A_\kappa(Q)] = \Pi$): The network substitutability *does not affect* the subscription decision.

In *Corollary 2*, $\bar{d} - A_\kappa(Q)$ represents the total data consumption of a type-$(\theta, 1)$ subscriber. This type of subscriber has so good a network substitutability that he stops using mobile data after his effective data cap is used up. Thus he only needs to pay the subscription fee Π in each month. Accordingly, $\Pi / [\bar{d} - A_\kappa(Q)]$ represents the average payment per unit data that he uses in each month.

Therefore, Case 1 in *Corollary 2* represents that the per-unit fee π is more expensive compared with the subscription fee Π in terms of the average rate of a type-$(\theta, 1)$ subscriber; Case 2 is the opposite of Case 1; Case 3 represents that the per-unit fee π, and the subscription fee Π are comparable for type-$(\theta, 1)$ subscribers.

Now we illustrate *Corollary 2* in Fig. 3, where the two axes correspond to the user's network substitutability β and data valuation θ, respectively. The gray region denotes the subscriber set $\Psi(\mathcal{T})$. The red arrow represents the direction where the network substitutability β increases.

- Fig. 3(a): If the per-unit fee π is more expensive than the average payment per unit data, i.e., $\pi[\bar{d} - A_\kappa(Q)] > \Pi$, then the red arrow shows that the users with a better network substitutability are more likely to become subscribers, since they incur less overage usage and thus less additional payment.
- Fig. 3(b): If the average payment for unit data is more expensive than the per-unit fee π, i.e., $\pi[\bar{d} - A_\kappa(Q)] < \Pi$, then the red arrow shows that the users with a better network substitutability are less likely to become subscribers. This is because that they are not willing to pay for an expensive subscription fee,

considering the cheap per-unit fee and their good alternative networks.

- Fig. 3(c): If the average payment for unit data and the per-unit fee π are comparable and satisfy $\pi[\bar{d} - A_\kappa(Q)] = \Pi$, then the network substitutability does not change users' subscription decisions.

Later on we will show that under the MNO's optimal pricing strategy, the market partition is Fig. 3(c).

Considering the subscription choice derived from Theorem 1, the MNO's expected total profit is given by

$$\tilde{W}(\mathcal{T}) = \int_0^1 \int_{\Theta(\mathcal{T},\beta)}^{\theta_{\max}} \Big\{ \underbrace{\pi(1-\beta)A_\kappa(Q) + \Pi}_{\text{Revenue}}$$
$$- \underbrace{c\left[\bar{d} - \beta A_\kappa(Q)\right]}_{\text{Operational cost}} \Big\} h(\theta)g(\beta)d\theta d\beta - \underbrace{J(Q)}_{\text{Capacity cost}}. \tag{23}$$

Next we will further analyze the MNO's optimal data cap and pricing strategy in Stage II.

5.2 Optimal Data Cap and Pricing Strategy in Stage II

In Stage II, the MNO determines the profit-maximizing data cap Q^* and the pricing strategy $\{\Pi^*, \pi^*\}$ considering users' subscription decisions from Stage III, given the data mechanism κ obtained in Stage I.

To make the presentation clear and reveal the key insights, we first present the MNO's optimal pricing strategy $\{\Pi^*, \pi^*\}$ given the data cap Q in Section 5.2.1, and then we introduce the MNO's optimal data cap Q^* in Section 5.2.2. [9]

5.2.1 Optimal Pricing Strategy

Given the data cap Q, the objective of MNO is to find the optimal subscription fee Π^* and per-unit fee π^* that maximize its expected total profit, that is,

Problem 1 (Optimal Pricing Strategy).

$$\{\Pi^*, \pi^*\} = \arg \max_{\Pi, \pi \geq 0} \tilde{W}(Q, \Pi, \pi, \kappa). \tag{24}$$

Before analyzing *Problem 1*, we need to introduce the following assumption on the MNO's QoS ρ and marginal operational cost c.

---

9. The two-step presentation enables us to illustrate the key insights of the optimal pricing strategy for a particularly data cap. This is practically important, since the MNO usually adopt integral data caps for simplicity, e.g., 1GB, 2GB, and 3GB.

Assumption 1. The MNO's QoS ρ and marginal operational cost c satisfy $c < \rho \cdot \theta_{\max}$.

Assumption 1 is made to avoid a trivial case, where the MNO offers very poor wireless service such that it cannot benefit from the market. It is not a technical assumption that limits our analysis and results.

Now we characterize the MNO's profit-maximizing subscription fee Π^* and per-unit fee π^* in Theorem 2.

Theorem 2 (Optimal Pricing Strategy). Given the data cap $Q \geq 0$ and the data mechanism κ, the MNO's profit-maximizing subscription fee Π^* and per-unit fee π^* satisfying the following conditions:

$$\begin{cases} H\left(\dfrac{\pi^*}{\rho}\right) + \dfrac{\pi^* - c}{\rho} \cdot h\left(\dfrac{\pi^*}{\rho}\right) = 1, \\ \Pi^* = \pi^* \left[\bar{d} - A_\kappa(Q)\right], \end{cases} \quad (25)$$

where $h(\cdot)$ and $H(\cdot)$ are the PDF and CDF of the data valuation θ. Furthermore, π^* is unique for an arbitrary θ distribution with an increasing failure rate (IFR).[10]

The proof of Theorem 2 is given in Appendix C.

Based on Theorem 2, we summarize the impact of several parameters on the optimal per-unit fee π^* and the optimal subscription fee Π^* in Proposition 1 and Proposition 2, respectively.

Proposition 1. The optimal per-unit fee π^* increases in the MNO's QoS ρ and marginal operational cost c. It dose not depend on the data mechanism κ or how large the data cap Q is.

Proposition 2. The optimal subscription fee $\Pi^*(Q, \kappa)$ is related to the data cap Q and the data mechanism κ in the following ways,

- $\Pi^*(Q, \kappa)$ increases in the data cap Q,
- a better time flexibility corresponds to a higher subscription fee, i.e., $\Pi^*(Q, 0) < \Pi^*(Q, 1) < \Pi^*(Q, 2) = \Pi^*(Q, 3)$ for all $Q \in (0, D)$.

Recall that Corollary 2 summarizes three possibilities for the impact of network substitutability β. Moreover, the optimal subscription fee Π^* and the per-unit fee π^* derived in Theorem 2 satisfy $\Pi^* = \pi^* \left[\bar{d} - A_\kappa(Q)\right]$, which is the same as Case 3 discussed in Corollary 2. In this case, the threshold valuation (defined in (22)) is

$$\Theta(Q, \Pi^*, \pi^*, \kappa, \beta) = \frac{\pi^*}{\rho}, \quad (26)$$

which is independent of β. This naturally leads to the following corollary on the market partition under the optimal pricing strategy.

Corollary 3. Under the optimal pricing strategy specified in Theorem 2, the network substitutability β does not affect the subscription decision. That is, the type-(θ, β) user will subscribe to the MNO if and only if $\theta > \pi^*/\rho$. The MNO obtains the market of

$$\Psi(Q, \Pi^*, \pi^*, \kappa) = \left\{ (\theta, \beta) : \frac{\pi^*}{\rho} \leq \theta \leq \theta_{\max}, 0 \leq \beta \leq 1 \right\}. \quad (27)$$

---

10. The increasing failure rate condition refers to $h(\theta)/\left[1 - H(\theta)\right]$ increasing in θ. Many commonly used distributions, such as uniform distribution, gamma distribution, and normal distribution, satisfy this condition [27].

Corollary 3 reveals that the MNO tends to select its subscribers based only on their data valuations, while ignoring the network substitutability. The intuition behind such a pricing strategy is that the MNO can benefit from good network substitutability users' subscription fee and poor network substitutability users' overage fee. Therefore, there is no incentive for the MNO to exclude either type of users.

Furthermore, according to Corollary 3, the MNO's market share under the optimal pricing strategy is fixed for any data cap Q and data mechanism κ. However, different data caps and data mechanisms bring the MNO different profit. Substitute $\Pi^*(Q, \kappa)$ and π^* into the MNO's expected total profit, then we obtain

$$\begin{aligned} &\tilde{W}(Q, \Pi^*(Q, \kappa), \pi^*, \kappa) \\ &= \frac{\left[\bar{d} - \bar{\beta} A_\kappa(Q)\right] (\pi^* - c)^2}{\rho} h\left(\frac{\pi^*}{\rho}\right) - J(Q), \end{aligned} \quad (28)$$

where $\bar{\beta}$ is the mean of the network substitutability among the whole user market, which is given by

$$\bar{\beta} = \int_0^1 \beta g(\beta) d\beta. \quad (29)$$

Next we move on to analyze the MNO's optimal data cap, considering the pricing strategy derived in Theorem 2.

5.2.2 Optimal Data Cap

The MNO needs to select a data cap Q to maximize its expected total profit $\tilde{W}(Q, \Pi^*(Q, \kappa), \pi^*, \kappa)$. That is, the MNO needs to solve the following problem

Problem 2 (Optimal Data Cap).

$$Q^* = \arg \max_{Q \geq 0} \tilde{W}(Q, \Pi^*(Q, \kappa), \pi^*, \kappa), \quad (30)$$

where $\Pi^*(Q, \kappa)$ and π^* are the MNO's profit-maximizing subscription fee and per-unit fee obtained from Theorem 2, respectively.

Problem 2 is not difficult to solve, since it is a single variable problem and is convex if $J(Q)$ is convex. To illustrate the key insights of the optimal data cap, hereafter, we follow [8] by making the following assumption on the MNO's capacity cost $J(Q)$ throughout the rest of the paper.

Assumption 2. The MNO's capacity cost takes a linear form, i.e., $J(Q) = z \cdot Q$ where z is the marginal capacity cost.

Before analyzing the MNO's optimal data cap, recall that the differences among the four data mechanisms $\kappa \in \{0, 1, 2, 3\}$ is entirely captured in $A_\kappa(Q)$, the expected overage data consumption. To facilitate our later analysis, we refer to $|A'_\kappa(Q)|$ as the **marginal overage data consumption**, which is the absolute value of the derivative of $A_\kappa(Q)$ with respect to Q. Specifically, the marginal overage data consumption $|A'_\kappa(Q)|$ measures the overage data usage decrement for a unit data cap increment on the data cap Q under the data mechanism κ.

Now we characterize the MNO's optimal data cap in Theorem 3. The proof of Theorem 3 is given in Appendix D.

Theorem 3 (Optimal Data Cap). Given the data mechanism κ, the MNO's optimal data cap $Q^*(\kappa)$ satisfies

$$|A'_\kappa(Q^*(\kappa))| = \Omega(\rho, c, z), \quad (31)$$

where $\Omega(\rho, c, z)$ is given by

$$\Omega(\rho, c, z) = \frac{z \cdot h\left(\frac{\pi^*}{\rho}\right)}{\bar{\beta}\rho\left[1 - H\left(\frac{\pi^*}{\rho}\right)\right]^2}. \tag{32}$$

Theorem 3 indicates that no matter which data mechanism κ that the MNO adopts, the MNO should always choose a data cap such that the corresponding *marginal overage data consumption* equals to $\Omega(\rho, c, z)$. Therefore, we refer to $\Omega(\rho, c, z)$ as the **target marginal overage data consumption** that the MNO must achieve to maximize its profit. Note that the target marginal overage data consumption $\Omega(\rho, c, z)$ is related to the MNO's QoS ρ, marginal operational cost c, and the marginal capacity cost z. Therefore, we further summarize how the three parameters affect the optimal data cap in Proposition 3.

Proposition 3. The MNO's optimal data cap Q^* increases in the QoS ρ, meanwhile decreases in the MNO's marginal operational cost c and marginal capacity cost z.

Now we have characterized the optimal data cap in Theorem 3, and revealed how the MNO's QoS (i.e., ρ) and marginal costs (i.e., c and z) affect it. Next we analyze the impact of the data mechanism κ on the optimal data cap, which is related to *Question 4* mentioned in Section 1.

Intuitively, we would think that the MNO can set a smaller cap under a data mechanism with a better time flexibility (to reduce the capacity cost), since it is more time-flexible for subscribers. In the following, however, we will reveal a counter-intuitive insight, i.e., *a better time flexibility does not necessarily correspond to a smaller data cap*.

To fully reveal this counter-intuitive insight, we need to know the detail mathematical expression of $A_\kappa(\cdot)$ according to (31) in Theorem 3. Even though we cannot analytically compute $A_\kappa(\cdot)$ due to the complexity of the Markov transition matrix, we are able to characterize some properties of $A_\kappa(\cdot)$ in Lemma 2. The proof is given in Appendix E.

Lemma 2. For an arbitrary data demand distribution $f(d)$, $A_\kappa(Q)$ is decreasing and convex in Q. Moreover, $A_\kappa(0) = \bar{d}$, $A_\kappa(D) = 0$, $A'_\kappa(0) = -1$, and $A'_\kappa(D) = 0$ for all $\kappa \in \{0, 1, 2, 3\}$.

To illustrate the counter-intuitive insight, let us consider two data mechanisms $i, j \in \{0, 1, 2, 3\}$, where j offers a better time flexibility, i.e., $\mathcal{F}_i < \mathcal{F}_j$.

Based on Lemma 2 and *Definition 1*, we can plot $A_i(Q)$ and $A_j(Q)$ versus the data cap Q in Fig. 4(a), and the corresponding marginal overage data consumptions $|A'_i(Q)|$ and $|A'_j(Q)|$ in Fig. 4(b).[11]

According to Theorem 3, the optimal data cap must make the corresponding marginal overage data consumption equal to the target marginal overage data consumption, i.e., $|A'_\kappa(Q^*(\kappa))| = \Omega(\rho, c, z)$. In Fig. 4(b), we consider two different target marginal overage data consumptions Ω_1 and Ω_2, which correspond to different values of ρ, c, and z.

Fig. 4: Illustration for $A_\kappa(Q)$ and $|A'_\kappa(Q)|$.

(a) $A_\kappa(Q)$ vs Q. (b) $|A'_\kappa(Q)|$ vs Q.

- To achieve a small target marginal overage data consumption Ω_1, the corresponding optimal data caps satisfy $Q^*(j) < Q^*(i)$, indicating that the data mechanism j (with a better time flexibility) leads to a smaller data cap;
- To achieve a large target marginal overage data consumption Ω_2, the corresponding optimal data caps satisfy $Q^*(j) > Q^*(i)$, indicating that the data mechanism j (with a better time flexibility) leads to a larger data cap.

Later we will further illustrate this counter-intuitive insight in Section 6.2.

So far we have fully characterized the optimal data cap Q^* together with the impact of the data mechanism as well as the MNO's QoS and costs. Next we will move on to study the MNO's optimal data mechanism in Stage I.

5.3 Optimal Data Mechanism in Stage I

In Stage I, the MNO determines the optimal data mechanism κ^* to maximize its expected total profit, which answers *Question 3* mentioned in Section 1.

Problem 3 (Optimal Data Mechanism).

$$\kappa^* = \arg\max_{\kappa \in \{0,1,2,3\}} \tilde{W}(Q^*(\kappa), \Pi^*(Q^*(\kappa)), \pi^*). \tag{33}$$

Lemma 3 reveals that a better time flexibility can always bring the MNO a higher profit under the optimal pricing strategy specified in Theorem 2 and the optimal data cap specified in Theorem 3, which naturally leads to Theorem 4.

Lemma 3. Consider two data mechanism $i, j \in \{0, 1, 2, 3\}$, and the data mechanism i has a better time flexibility than j, i.e., $\mathcal{F}_i > \mathcal{F}_j$. Then the following is true

$$\tilde{W}(Q^*(i), \Pi^*(Q^*(i)), \pi^*) \geq \tilde{W}(Q^*(j), \Pi^*(Q^*(j)), \pi^*), \tag{34}$$

where the equality holds if and only if $Q^*(i) = Q^*(j) = 0$ or $Q^*(i) = Q^*(j) = D$.

Theorem 4 (Optimal Data Mechanism). Among the four data mechanisms $\{0, 1, 2, 3\}$, the MNO's optimal data mechanism is $\kappa^* = 2$ and 3.

Theorem 4 shows that the MNO should adopt the rollover mechanism offered by China Mobile or the credit mechanism proposed in this paper to maximize its profit.

Next, we examine the conditions of the system parameters under which the optimal three-part tariff data plan

---

11. Note that in Fig. 4(b) there is only one crossing point for $|A'_i(Q)|$ and $|A'_j(Q)|$ when $Q \in (0, D)$. Basically, Lemma 2 cannot imply the uniqueness of the crossing point for an arbitrary demand distribution $f(d)$. Nevertheless, it does not affect the counter-intuitive insight, i.e., *a better time flexibility does not necessarily correspond to a smaller data cap*.

(a) $\theta \sim \mathrm{Gamma}(4.5, 0.11)$ (b) $\beta \sim \mathrm{TN}(0.91, 0.22, 0, 1)$

Fig. 5: Fitting the PDF of θ and β.

Fig. 6: $A_\kappa(Q)$ vs Q and $f(d)$ vs d. Due to space limit, we only show the part on the interval $[0, D/2]$.

degenerates into the pure usage-based plan or the flat-rate plan, in which case the choice of data mechanism $\kappa \in \{0, 1, 2, 3\}$ has no effect on the subscribers.

Corollary 4 (Pure Usage-Based Plan). If the target marginal overage data consumption $\Omega(\rho, c, z) \geq 1$, then the MNO maximizes its expected total profit by offering the pure usage-based data plan $\mathcal{T}_{\mathrm{p}} = \{0, 0, \pi^*, \mathtt{Na}\}$.

The condition in Corollary 4 is satisfied when the MNO

- provides poor services, i.e., ρ is small, or
- experiences a large cost, i.e., c or z is large.

The above insight is consistent with the reality. From the users' perspective, they are not willing to pay for any cap if the MNO's QoS is poor. From the MNO's perspective, it is not beneficial for it to incentivize more data consumption through a large data cap, if it experiences a large cost from network investment or system management.

As we mentioned in Section 1, the flat-rate data plan appears earliest in the telecommunication market. However, most MNOs do not offer the flat-rate data plan anymore. The following corollary can provide an explanation for this phenomenon.

Corollary 5 (Flat-Rate Plan). If the marginal capacity cost $z = 0$, then the MNO maximizes its expected total profit by offering the flat-rate data plan $\mathcal{T}_{\mathrm{f}} = \{D, \bar{d}\pi^*, \mathtt{Na}, \mathtt{Na}\}$.

Obviously, the condition in Corollary 5 corresponds to an extreme case that does not approximate the current reality well [28], which explains why the flat-rate data plan has ended in the past.

6 NUMERICAL RESULTS

To examine the performance of different data mechanisms, we collected some real data from the telecommunication market in mainland China[12]. In Section 6.1, we analyze the distributions of the data valuation and the network substitutability. In Section 6.2, we simulate the optimal data plan and investigate the effect of the time flexibility on users' payoffs and the MNO's profit.

---

12. The related data is obtained through a survey, and the questionnaire is available through https://www.wjx.cn/jq/16895923.aspx.

6.1 Empirical Results

To estimate the statistic information of users' data valuation θ and network substitutability β, we collected some mobile users' behavioral data from the telecommunication market in mainland China. The PDFs of these two parameters θ and β are shown by the green bars in Fig. 5(a) and Fig. 5(b), respectively. We observe that a large proportion of users' data valuations θ falls into the range between 10 RMB/GB to 60 RMB/GB, and most people would like to shrink $70\% \sim 100\%$ of their overage data demand. Moreover, we also find that the Pearson correlation coefficient between θ and β is less than 0.05, which allows us to fit the two distributions independently.

Next we estimate the data valuation θ PDF by assuming a gamma distribution, which satisfies the increasing failure rate (IFR). The PDF of a gamma distribution in the shape-rate parametrization is

$$h(\theta, k, r) = \frac{r^k \theta^{k-1} e^{-r\theta}}{\Gamma(k)}, \tag{35}$$

where $\Gamma(k)$ is a complete gamma function. We choose the parameters $k = 4.5$ and $r = 0.11$ by minimizing the least-squares divergence between the estimated and empirical PDF. In Fig. 5(a), the black curve is the estimated PDF. Visually, it is qualitatively similar to the empirical PDF. To further investigate the goodness-of-fit statistically, we use the Kolmogorov-Smirnov test on the null hypothesis that the data valuation comes from the gamma distribution, at the significance level of 0.05 (i.e., if the Kolmogorov-Smirnov test returns a p-value less than 0.05, then we need to reject this hypothesis) [29]. The test shows a p-value of 0.31, hence the hypothesis that the data valuation follows the gamma distribution is valid.

Next we estimate the network substitutability β PDF by assuming a truncated normal distribution, since the network substitutability β has a concrete upper bound and lower bound, i.e., $0 \leq \beta \leq 1$. The PDF of a normal distribution $N(\mu, \sigma^2)$ truncated between $[a, b]$ is

$$g(\beta; \mu, \sigma, a, b) = \frac{\phi\left(\frac{\beta - \mu}{\sigma}\right)}{\sigma\left[\Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right)\right]}, \tag{36}$$

where $\phi(\cdot)$ is the probability density function of the standard normal distribution, and $\Phi(\cdot)$ is its cumulative distribution function. Similarly, we find the truncated normal distribution $\beta \sim \mathrm{TN}(0.91, 0.22, 0, 1)$ by minimizing the least-squares divergence between the estimated and empirical

(a) Optimal cap $Q^*(\kappa)$ vs ρ (b) Optimal cap $Q^*(\kappa)$ vs c (c) Optimal cap $Q^*(\kappa)$ vs z

Fig. 7: The optimal data cap $Q^*(\kappa)$ under different data mechanisms $\kappa \in \{0, 1, 2\}$.

(a) $\Pi^*(\kappa)$ vs ρ (b) $\Pi^*(\kappa)$ vs c (c) $\Pi^*(\kappa)$ vs z

Fig. 8: The optimal subscription fee $\Pi^*(\kappa)$ under different data mechanisms $\kappa \in \{0, 1, 2\}$.

PDFs. The corresponding Kolmogorov-Smirnov test shows a p-value of 0.38, hence the hypothesis of the truncated normal distribution is valid.

6.2 Performance Evaluation

Next we will use the fitted market distribution to investigate how the MNO's QoS and marginal costs affect its optimal data plan, and examine the impact of time flexibility on the users' payoff and the MNO's profit.

We set the minimum data unit to 1MB. Following the data analysis results in [20], [21], we assume that users' monthly data demand follows a truncated log-normal distribution with a mean $\bar{d} = 10^3$ on the interval $[0, 10^4]$, i.e., the mean value is $\bar{d} = 1$GB and the maximal usage is $D = 10$GB. Fig. 6 shows the PMF $f(d)$ and the expected monthly overage usage $A_\kappa(Q)$ under the four data mechanisms, which indicates that $A_0(Q) > A_1(Q) > A_2(Q) = A_3(Q)$ for all $Q \in (0, D)$. Since the degree of time flexibility under $\kappa = 2$ and $\kappa = 3$ is equivalent, in the following we will neglect $\kappa = 3$ and only plot the results for $\kappa = 0, 1, 2$.

6.2.1 Optimal Data Plan

We investigate the impact of QoS ρ, marginal operational cost c, and the marginal capacity cost z individually.

In Fig. 7, we use three sub-figures to plot the optimal data cap $Q^*(\kappa)$ versus the MNO's QoS ρ, marginal operational cost c, and the marginal capacity cost z, respectively. In addition, the three curves in each sub-figure correspond to the three data mechanisms $\kappa \in \{0, 1, 2\}$.

Overall, the optimal data cap $Q^*(\kappa)$ increases (from zero) as the MNO becomes stronger, in terms of

- a better QoS ρ, as shown in Fig. 7(a),
- a smaller operational cost c, as shown in Fig. 7(b),
- a smaller capacity cost z, as shown in Fig. 7(c).

Particularly, the pure usage-based data plan appears when the MNO's QoS $\rho < 0.35$ in Fig. 7(a), marginal operational cost $c > 62$ RMB/GB in Fig. 7(b), or marginal capacity cost $z > 8.5 \times 10^{-3}$ RMB/GB in Fig. 7(c). In addition, Fig. 7(c) shows that the flat-rate data plan appears when $z = 0$, since the corresponding data cap is the maximal data demand 10GB.

As we mentioned in Section 5.2.2, a better time flexibility does not necessarily correspond to a smaller data cap. By comparing the three curves in each sub-figure of Fig. 7, we find that a *better* time flexibility would lead to an even *larger* data cap if the MNO is weak in terms of

- a poor QoS, e.g., $\rho = 0.4$ in Fig. 7(a),
- a large marginal operational cost, e.g., $c = 50$ RMB/GB in Fig. 7(b),
- a large marginal capacity cost, e.g., $z = 6 \times 10^{-3}$ RMB/GB in Fig. 7(c).

The intuitions behind the counter-intuitive result include

- When the MNO is *weak*, it chooses a small data cap, under which the main revenue comes from users' overage payments. In this case, offering a better time flexibility can significantly reduce its revenue from

Fig. 9: MNO's profit gain and users' payoff gain with $\kappa = 0$ as the benchmark.

users' overage payments. Therefore, the MNO will increase the data cap and the subscription fee to compensate its revenue loss.

- When the MNO is *strong*, it chooses a large data cap, under which its main revenue comes from users' subscription fee. In this case, offering a better time flexibility does not reduce its revenue too much, and the MNO will decrease the data cap to further reduce its cost.

Next we examine the impact of the time flexibility on the monthly subscription fee.

Fig. 8 plots the optimal subscription fee $\Pi^*(\kappa)$ under different data mechanisms. By comparing the three curves in each sub-figure, we observe that a higher subscription fee is always associated with a data mechanism with better time flexibility, even though it may correspond to a smaller or larger data cap as shown in Fig. 7.

The above discussions together with Proposition 1 provide answers to *Question 4* mentioned in Section 1: a better time flexibility corresponds to a smaller data cap Q^* if the MNO is strong or a larger data cap if the MNO is weak. Meanwhile, a better time flexibility always leads to a higher subscription fee Π^*. Finally, it does not affect the optimal per-unit fee π^*.

6.2.2 Users' Payoffs and MNO's Profit

We investigate the performance of different data mechanisms in terms of all users' payoffs and the MNO's profit. Specifically, we set the traditional data mechanism $\kappa = 0$ as the benchmark, and plot the performance gain of other schemes comparing to the benchmark. The three sub-figures in Fig. 9 plot the performance gains versus the MNO's QoS ρ, marginal operational cost c, and the marginal capacity cost z, respectively. The two *solid curves* in each sub-figure correspond to the MNO's profit gains for $\kappa \in \{1, 2\}$, the other two *dash curves* represent users' payoff gains for $\kappa \in \{1, 2\}$.

Overall, Fig. 9 show that the users' payoff gains (i.e., the dash curves) decrease as the MNO's QoS ρ increases as in Fig. 9(a) or the MNO's marginal costs c and z decrease as in Fig. 9(b) and Fig. 9(c). In this process, however, the MNO's profit gains (i.e., the solid curves) first increase then decrease in all three sub-figures.

From the two *solid curves* in each sub-figure of Fig. 9, we find that the MNO's profit gains under $\kappa = 1$ and $\kappa = 2$ are both non-negative, thus the time flexibility can increase the profit of MNO compared with the benchmark $\kappa = 0$. From the two *dash curves* in each sub-figure of Fig. 9, we find that the users' payoffs gains under $\kappa = 1$ and $\kappa = 2$ are both non-negative, thus the time flexibility increases the users' payoffs as well. Moreover, we also note that $\kappa = 2$ always outperforms $\kappa = 1$ in terms of both MNO's profit gain and users' payoffs gain, which indicates that a better time flexibility leads to a larger improvement.

Now we know that both the MNO and the users can benefit from the time flexibility, a natural question is who will benefit more? By comparing the two red curves with squares (or the two blue curves with triangles) in each sub-figure, we find that the MNO benefits more from the time flexibility than users if the MNO is strong, in terms of

- a good QoS, e.g., $\rho > 0.43$ in Fig. 9(a),
- a small marginal operational cost, e.g., $c < 50$ RMB/GB in Fig. 9(b),
- a small marginal capacity cost, e.g., $z < 4 \times 10^{-4}$ RMB/GB in Fig. 9(c).

Intuitively, a stronger MNO has a larger pricing power, thus the strong MNO can benefit more than users from adding time flexibility. However, a weaker MNO has to leave users more benefits to maintain its market, thus users benefit more than a weaker MNO from time flexibility.

Furthermore, we also investigate the impact of the variances of the demand distribution, which indicates that the performance gain of the MNO's profit and users' payoff will increase in the variance. Due to the space limit, please refer to Appendix F for more detailed discussions.

The above discussions answer *Question 1* in Section 1. In a monopoly market, both the MNO and users can benefit from a better time flexibility. Moreover, the MNO benefits more if the MNO offers good services and experiences small costs. Otherwise, the users benefit more.

7 CONCLUSIONS AND FUTURE WORKS

In this paper, we study the MNO's optimal three-part tariff plan with time flexibility. Specifically, we consider four data mechanisms, and formulate the MNO's optimal data plan design problem as a three-stage Stackelberg Game. Through backward induction, we analytically characterize the users' subscription choices in Stage III, the MNO's

optimal data cap and corresponding pricing strategy in Stage II, and the MNO's optimal data mechanism in Stage I. Moreover, we conduct a market survey to estimate the distribution of users' data valuation and the network substitutability. Then we evaluate the performance of different data mechanisms using the real data.

In the future work, we want to collect more empirical data to estimate the MNO's cost and users' data demand distributions, and validate the insights obtained based on the current linear costs model and homogeneous data demand distribution. Moreover, we will consider a more general competitive market, and analyze the impact of time flexibility on multiple MNOs' competition.

REFERENCES

[1] Z. Wang, L. Gao, and J. Huang, "Pricing optimization of rollover data plan," in *15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2017.

[2] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang, "Smart data pricing: using economics to manage network congestion," *Communications of the ACM*, vol. 58, no. 12, pp. 86–93, 2015.

[3] AT&T. [Online]. Available: https://www.att.com

[4] T-mobile. [Online]. Available: https://support.t-mobile.com

[5] China mobile. [Online]. Available: http://www.10086.cn

[6] Data bank. [Online]. Available: http://bank.wo.cn/pc/home.html

[7] R. T. Ma, "Usage-based pricing and competition in congestible network service markets," *IEEE/ACM Transactions on Networking*, vol. 24, no. 5, pp. 3084–3097, 2016.

[8] W. Dai and S. Jordan, "The effect of data caps upon isp service tier design and users," *ACM Transactions on Internet Technology (TOIT)*, vol. 15, no. 2, p. 8, 2015.

[9] X. Wang, R. T. Ma, and Y. Xu, "The role of data cap in optimal two-part network pricing," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3602–3615, 2017.

[10] Z. Xiong, S. Feng, D. Niyato, P. Wang, and Y. Zhang, "Economic analysis of network effects on sponsored content: a hierarchical game theoretic approach," in *IEEE Global Communications Conference (GLOBECOM)*, 2017.

[11] L. Zheng, C. Joe-Wong, M. Andrews, and M. Chiang, "Optimizing data plans: Usage dynamics in mobile data networks," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2018.

[12] L. Zheng and C. Joe-Wong, "Understanding rollover data," in *Smart Data Pricing Workshop*, 2016.

[13] Y. Wei, J. Yu, T. M. Lok, and L. Gao, "A novel mobile data plan design from the perspective of data period," in *IEEE International Conference on Communication Systems (ICCS)*, 2016.

[14] S. Sen, C. Joe-Wong, and S. Ha, "The economics of shared data plans," in *Annual Workshop on Information Technologies and Systems*, 2012.

[15] L. Zheng, C. Joe-Wong, C. W. Tan, S. Ha, and M. Chiang, "Customized data plans for mobile users: Feasibility and benefits of data trading," *IEEE journal on selected areas in communications*, vol. 35, no. 4, pp. 949–963, 2017.

[16] J. Yu, M. H. Cheung, J. Huang, and H. V. Poor, "Mobile data trading: Behavioral economics analysis and algorithm design," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 4, pp. 994–1005, 2017.

[17] L. Duan, J. Huang, and B. Shou, "Pricing for local and global wi-fi markets," *IEEE Transactions on Mobile Computing*, vol. 14, no. 5, pp. 1056–1070, 2015.

[18] Q. Ma, Y.-F. Liu, and J. Huang, "Time and location aware mobile data pricing," *IEEE Transactions on Mobile Computing*, vol. 15, no. 10, pp. 2599–2613, 2016.

[19] Z. Wang, L. Gao, and J. Huang, "Pricing competition of rollover data plan," *16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, 2018.

[20] A. Lambrecht, K. Seim, and B. Skiera, "Does uncertainty matter? consumer behavior under three-part tariffs," *Marketing Science*, vol. 26, no. 5, pp. 698–710, 2007.

[21] A. Nevo, J. L. Turner, and J. W. Williams, "Usage-based pricing and demand for residential broadband," *Econometrica*, vol. 84, no. 2, pp. 411–443, 2016.

[22] W. P. Rogerson and E. Charles, "The economics of data caps and free data services in mobile broadband," *CTIA report*, 2016.

[23] D. A. Lyons, "The impact of data caps and other forms of usage-based pricing for broadband access," *Mercatus Center Working Papers*, 2012.

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

[25] L. Duan, J. Huang, and B. Shou, "Duopoly competition in dynamic spectrum leasing and pricing," *IEEE Transactions on Mobile Computing*, vol. 11, no. 11, pp. 1706–1719, 2012.

[26] R. F. Bass, *Stochastic processes*. Cambridge University Press, 2011.

[27] X. Brusset, "Properties of distributions with increasing failure rate," *MPRA Paper 18299, University Library of Munich*, 2009.

[28] E. J. Oughton and Z. Frias, "The cost, coverage and rollout implications of 5G infrastructure in britain," *Telecommunications Policy*, 2017.

[29] F. J. Massey Jr, "The kolmogorov-smirnov test for goodness of fit," *Journal of the American statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

Zhiyuan Wang received the B.S. degree from Southeast University, Nanjing, China, in 2016. He is currently working toward the Ph.D. degree with the Department of Information Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong. His research interests include the field of network economics and game theory, with current emphasis on smart data pricing and fog computing. He is the recipient of the Hong Kong PhD Fellowship.

Lin Gao (S'08-M'10-SM'16) received the PhD degree in electronic engineering from Shanghai Jiao Tong University in 2010 and worked as a postdoctoral research fellow with the Chinese University of Hong Kong from 2010 to 2015. He is an associate professor in the School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, China. He received the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2016. His main research interests are in the area of network economics and games, with applications in wireless communications and networking. He has co-authored 3 books including the textbook Wireless Network Pricing by Morgan & Claypool (2013). He is the co-recipient of three Best (Student) Paper Awards from WiOpt 2013, 2014, 2015, and is a Best Paper Award Finalist from IEEE INFOCOM 2016.

Jianwei Huang (F'16) is a Professor and Director of the Network Communications and Economics Lab (ncel.ie.cuhk.edu.hk), in the Department of Information Engineering at the Chinese University of Hong Kong. He received the Ph.D. degree from Northwestern University in 2005, and worked as a Postdoc Research Associate at Princeton University during 2005-2007. He is an IEEE Fellow, a Distinguished Lecturer of IEEE Communications Society, and a Clarivate Analytics Highly Cited Researcher in Computer Science. Jianwei Huang is the co-author of 9 Best Paper Awards, including IEEE Marconi Prize Paper Award in Wireless Communications in 2011. He has co-authored six books, including the textbook on "Wireless Network Pricing." He received the CUHK Young Researcher Award in 2014 and IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2009. He has served as an Associate Editor of IEEE Transactions on Mobile Computing, IEEE/ACM Transactions on Networking, IEEE Transactions on Network Science and Engineering, IEEE Transactions on Wireless Communications, IEEE Journal on Selected Areas in Communications - Cognitive Radio Series, and IEEE Transactions on Cognitive Communications and Networking. He has served as an Editor of Wiley Information and Communication Technology Series, Springer Encyclopedia of Wireless Networks, and Springer Handbook of Cognitive Radio.

APPENDIX A
PROOF OF LEMMA 1

In order to prove Lemma 1, in the following three subsections, we prove $A_0(Q) \geq A_1(Q)$, $A_2(Q) = A_3(Q)$, and $A_1(Q) \geq A_2(Q)$, respectively.

A.1 Proof of $A_0(Q) \geq A_1(Q)$

We prove $A_0(Q) \geq A_1(Q)$ by computing the weighted summations in two steps. Specifically, $A_0(Q)$ and $A_1(Q)$ are given by

$$A_0(Q) = \sum_{d=0}^{D} [d - Q]^+ f(d),$$
$$A_1(Q) = \sum_{d=0}^{D} \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_1(\tau) f(d). \tag{37}$$

It is obvious that the following inequality is true

$$[d - Q]^+ \geq [d - Q - \tau]^+, \ \forall \, \tau \in \{0, 1, 2, ..., Q\}. \tag{38}$$

Now we compute the weighted summation over the weight $p_1(\tau)$ for (38), and we obtain

$$\sum_{\tau=0}^{Q} [d - Q]^+ p_1(\tau) \geq \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_1(\tau), \tag{39}$$

which is equivalent to

$$[d - Q]^+ \geq \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_1(\tau), \ \forall \, d \in \{0, 1, ..., D\}. \tag{40}$$

Then we further compute the weighted summation over the weight $f(d)$ for (40), and we obtain

$$\sum_{d=0}^{D} [d - Q]^+ f(d) \geq \sum_{d=0}^{D} \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_1(\tau) f(d), \tag{41}$$

which implies $A_0(Q) \geq A_1(Q)$.

A.2 Proof of $A_2(Q) = A_3(Q)$

Second, we prove $A_2(Q) = A_3(Q)$ through transforming their mathematical expressions. Specifically, $A_2(Q)$ and $A_3(Q)$ are given by

$$A_2(Q) = \sum_{d=0}^{D} \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_2(\tau) f(d),$$
$$A_3(Q) = \sum_{d=0}^{D} \sum_{\tau=-Q}^{0} [d - 2Q - \tau]^+ p_3(\tau) f(d). \tag{42}$$

As mentioned in Section 4.1, for $\kappa = 3$, the data deficit of the next month τ_3' is

$$\tau_3' = \begin{cases} 0, & \text{if } d \in [0, Q + \tau_3], \\ Q + \tau_3 - d, & \text{if } d \in (Q + \tau_3, 2Q + \tau_3), \\ -Q, & \text{if } d \in [2Q + \tau_3, D]. \end{cases} \tag{43}$$

Here we further define $\mu = \tau_3 + Q$ and $\mu' = \tau_3' + Q$, thus

$$\mu' = \begin{cases} Q, & \text{if } d \in [0, Q + \tau_3], \\ 2Q + \tau_3 - d, & \text{if } d \in (Q + \tau_3, 2Q + \tau_3), \\ 0, & \text{if } d \in [2Q + \tau_3, D]. \end{cases} \tag{44}$$

Now we substitute $\tau_3 = \mu - Q$ into (44), and get

$$\mu' = \begin{cases} Q, & \text{if } d \in [0, \mu], \\ Q + \mu - d, & \text{if } d \in (\mu, Q + \mu), \\ 0, & \text{if } d \in [Q + \mu, D]. \end{cases} \tag{45}$$

Therefore, the transition matrix from μ to μ' is

$$p_\mu(\mu, \mu') = \begin{cases} \sum_{d=0}^{\mu} f(d), & \text{if } \mu' = Q, \\ f(Q + \mu - \mu'), & \text{if } \mu' \in (0, Q), \\ \sum_{d=Q+\mu}^{D} f(d), & \text{if } \mu' = 0, \end{cases} \tag{46}$$

which is the same as the transition matrix of the data mechanism $\kappa = 2$ derived in (16). Thus they have the same stationary distribution, i.e., $p_\mu(\mu) = p_2(\tau)$ for any $\mu = \tau$.

Now we substitute $\tau_3 = \mu - Q$ into (42), and obtain

$$A_3(Q) = \sum_{\mu=0}^{Q} \sum_{d=0}^{D} [d - Q - \mu]^+ f(d) p_\mu(\mu) = A_2(Q). \tag{47}$$

A.3 Proof of $A_1(Q) \geq A_2(Q)$

Now we prove $A_1(Q) \geq A_2(Q)$. Specifically, $A_1(Q)$ and $A_2(Q)$ are given by

$$A_1(Q) = \sum_{d=0}^{D} \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_1(\tau) f(d). \tag{48}$$

$$A_2(Q) = \sum_{d=0}^{D} \sum_{\tau=0}^{Q} [d - Q - \tau]^+ p_2(\tau) f(d). \tag{49}$$

The only difference between (48) and (48) lies in the stationary distribution $p_1(\tau)$ and $p_2(\tau)$. Due to the complexity of the Markov chain specified in (16), we are not able to compute the closed-form expression for $A_1(Q)$ and $A_2(Q)$. Nevertheless, we will prove $A_1(Q) \geq A_2(Q)$ by showing that for an arbitrary realized data demand sequence, the overage data consumption under $\kappa = 1$ is larger than that under $\kappa = 2$.

We consider T months, i.e., $t = 1, 2, ..., T$. We denote $\boldsymbol{d} = \{d^1, d^2, ..., d^T\}$ as the realized data demand sequence, where d^t denotes the realized data demand in month t. Moreover, we denote τ_1^t and τ_2^t as the subscriber's data surplus under the data mechanism $\kappa = 1$ and $\kappa = 2$ at the beginning of month t.

Therefore, given the realized data usage sequence $\boldsymbol{d}$, the subscriber's overage data consumption in month t is

$$A_\kappa^t(Q, \boldsymbol{d}) = [d^t - Q - \tau_\kappa^t]^+. \tag{50}$$

Next we first introduce Lemma 4.

Lemma 4. For any realized data demand sequence $\boldsymbol{d}$, we have $A_1^t(Q, \boldsymbol{d}) \geq A_2^t(Q, \boldsymbol{d})$ and $\tau_1^t \leq \tau_2^t$ for any $t = 1, 2, ..., T$.

Proof of Lemma 4: We prove Lemma 4 by induction.

First, if $t = 1$, then $\tau_1^1 = \tau_2^1 = 0$, thus we know $A_1^1(Q, \boldsymbol{d}) = A_2^1(Q, \boldsymbol{d}) = [d^1 - Q]^+$. Hence, Lemma 4 is true when $t = 1$.

Second, we assume that Lemma 4 is true for $t = k$, and consider the case of $t = k + 1$.

We first show that $\tau_0^{k+1} \leq \tau_1^{k+1}$. According to we consumption priority, we know that τ_1^{k+1} is

$$\tau_1^{k+1} = [Q - d^k]^+, \tag{51}$$

and τ_2^{k+1} is

$$\tau_2^{k+1} = \begin{cases} Q, & \text{if } d^k < \tau_2^k, \\ \left[Q + \tau_2^k - d^k\right]^+, & \text{if } d^k \geq \tau_2^k. \end{cases} \quad (52)$$

Based on (51) and (52), we know that

- If $\tau_2^k = 0$, then $\tau_1^{k+1} = \tau_2^{k+1}$ for any d^k.
- If $\tau_2^k > 0$, then we have the following relation between τ_1^{k+1} and τ_2^{k+1},

$$\begin{cases} \tau_1^{k+1} < \tau_2^{k+1}, & \text{if } d^k \in (0, Q + \tau_2^k), \\ \tau_1^{k+1} = \tau_2^{k+1}, & \text{if } d^k \in \{0\} \cup [Q + \tau_2^k, +\infty). \end{cases} \quad (53)$$

Therefore, we find that $\tau_0^{k+1} \leq \tau_1^{k+1}$. Next we show that $A_1^{k+1}(Q, \boldsymbol{d}) \geq A_2^{k+1}(Q, \boldsymbol{d})$. According to the following formulation

$$\begin{cases} A_1^{k+1}(Q, \boldsymbol{d}) = \left[d^{k+1} - Q - \tau_1^{k+1}\right]^+, \\ A_2^{k+1}(Q, \boldsymbol{d}) = \left[d^{k+1} - Q - \tau_2^{k+1}\right]^+, \end{cases} \quad (54)$$

we know that

- If $\tau_1^{k+1} = \tau_2^{k+1}$, then we have $A_1^{k+1}(Q, \boldsymbol{d}) = A_2^{k+1}(Q, \boldsymbol{d})$ for any realized data demand sequence $\boldsymbol{d}$.
- If $\tau_1^{k+1} < \tau_2^{k+1}$, then we have the following relation between $A_1^{k+1}(Q, \boldsymbol{d})$ and $A_2^{k+1}(Q, \boldsymbol{d})$,

$$\begin{cases} A_1^{k+1}(Q, \boldsymbol{d}) = A_2^{k+1}(Q, \boldsymbol{d}) & \text{if } d^{k+1} \leq Q + \tau_1^{k+1}, \\ A_1^{k+1}(Q, \boldsymbol{d}) > A_2^{k+1}(Q, \boldsymbol{d}) & \text{if } d^{k+1} > Q + \tau_1^{k+1}. \end{cases} \quad (55)$$

Therefore, we obtain $A_1^{k+1}(Q, \boldsymbol{d}) \geq A_2^{k+1}(Q, \boldsymbol{d})$, which implies that Lemma 4 is true for $t = k + 1$. Hence, Lemma 4 is true. ∎

According to Lemma 4, we can conclude that

$$\sum_{t=1}^{T} A_1^{k+1}(Q, \boldsymbol{d}) \geq \sum_{t=2}^{T} A_1^{k+1}(Q, \boldsymbol{d}), \quad (56)$$

which implies that for any realized data demand sequence $\boldsymbol{d}$, the overage data consumption under $\kappa = 1$ is larger than that under $\kappa = 2$. Hence, it also holds when we take the expectation over the stochastic data demand d, i.e., $A_1(Q) \geq A_2(Q)$. This completes the proof of Lemma 1. ∎

APPENDIX B
PROOF OF THEOREM 1

A type-(θ, β) subscriber's expected monthly payoff is

$$\bar{S}(\mathcal{T}, \theta, \beta) = \rho\theta \left[\bar{d} - \beta A_\kappa(Q)\right] - \pi(1 - \beta)A_\kappa(Q) - \Pi, \quad (57)$$

by solving $\bar{S}(\mathcal{T}, \theta, \beta) \geq 0$, we obtain

$$\theta \geq \Theta(\mathcal{T}, \beta) \triangleq \frac{1}{\rho}\left[\pi + \frac{\pi\left[\bar{d} - A_\kappa(Q)\right] - \Pi}{\beta A_\kappa(Q) - \bar{d}}\right], \quad (58)$$

thus the market share of the MNO under the data plan $\mathcal{T} = \{Q, \Pi, \pi, \kappa\}$ is

$$\Psi(\mathcal{T}) = \{(\theta, \beta) : \Theta(\mathcal{T}, \beta) \leq \theta \leq \theta_{\max}, 0 \leq \beta \leq 1\}. \quad (59)$$

∎

APPENDIX C
PROOF OF THEOREM 2

We prove Theorem 2 by deriving the MNO's profit-maximizing subscription fee and per-unit fee. Recall that the profit of the MNO is given by

$$\tilde{W}(\mathcal{T}) = \int_0^1 \int_{\Theta(\mathcal{T}, \beta)}^{\theta_{\max}} \left\{\pi(1 - \beta)A_\kappa(Q) + \Pi \atop - c \cdot \left[\bar{d} - \beta A_\kappa(Q)\right]\right\} h(\theta)g(\beta)d\theta d\beta - J(Q). \quad (60)$$

Since we consider a fixed Q in Theorem 2, then $A_\kappa(Q)$ is a constant, and we denote it as A_κ. Accordingly, we obtain

$$\begin{aligned} \tilde{W}(\Pi, \pi) &= \int_0^1 \int_{\Theta(\mathcal{T}, \beta)}^{\theta_{\max}} \left\{\pi(1 - \beta)A_\kappa + \Pi \atop - c \cdot \left[\bar{d} - \beta A_\kappa\right]\right\} h(\theta)g(\beta)d\theta d\beta \\ &= \int_0^1 \left\{\pi(1 - \beta)A_\kappa + \Pi - c \cdot \left[\bar{d} - \beta A_\kappa\right]\right\} \\ &\qquad \left[1 - H\left(\Theta(\mathcal{T}, \beta)\right)\right]g(\beta)d\beta, \end{aligned} \quad (61)$$

where $\Theta(\mathcal{T}, \beta)$ is given by

$$\Theta(\mathcal{T}, \beta) = \frac{\pi(1 - \beta)A_\kappa + \Pi}{\rho\left(\bar{d} - \beta A_\kappa\right)}. \quad (62)$$

Therefore, we can write $\pi(1 - \beta)A + \Pi$ as

$$\pi(1 - \beta)A_\kappa + \Pi = \Theta(\mathcal{T}, \beta) \cdot \rho\left(\bar{d} - \beta A_\kappa\right). \quad (63)$$

When substituting (63) into (61), we can write the MNO's profit as follows:

$$\tilde{W}(\Pi, \pi) = \int_0^1 \rho\left(\bar{d} - \beta A_\kappa\right)\left[\Theta(\mathcal{T}, \beta) - \frac{c}{\rho}\right] \atop \cdot \left[1 - H\left(\Theta(\mathcal{T}, \beta)\right)\right]g(\beta)d\beta, \quad (64)$$

Note that, under the assumption that θ distribution satisfies the increasing failure rate (i.e., $h(x)/\left[1 - H(x)\right]$ increases in x), we have the following inequality

$$\left(x - \frac{c}{\rho}\right)\left[1 - H(x)\right] \leq \frac{\left[1 - H(x^*)\right]^2}{h(x^*)}, \quad (65)$$

where x^* is *uniquely* determined by

$$\frac{1 - H(x^*)}{h(x^*)} = x^* - \frac{c}{\rho}, \quad (66)$$

since $\left[1 - H(x)\right]/h(x)$ decreases in x and $x - c/\rho$ increases in x.

Based on (65) and (66), we know that when the subscription fee Π^* and the per-unit fee π^* satisfy

$$\begin{cases} H\left(\frac{\pi^*}{\rho}\right) + \frac{\pi^* - c}{\rho} \cdot h\left(\frac{\pi^*}{\rho}\right) = 1, \\ \Pi^* = \pi^*\left[\bar{d} - A\right], \end{cases} \quad (67)$$

we always have the following inequality

$$\rho\left(\bar{d} - \beta A\right)\left[\Theta(\mathcal{T}, \beta) - \frac{c}{\rho}\right]\left[1 - H\left(\Theta(\mathcal{T}, \beta)\right)\right]$$

$$\leq \rho\left(\bar{d} - \beta A\right) \cdot \frac{\left[1 - H\left(\Theta(Q, \Pi^*, \pi^*, \kappa, \beta)\right)\right]^2}{h\left(\Theta(Q, \Pi^*, \pi^*, \kappa, \beta)\right)}, \forall \beta \in [0, 1]. \quad (68)$$

By computing the integration over $\beta \in [0,1]$ on (68), we have

$$\int_0^1 \rho\left(\bar{d} - \beta A\right)\left[\Theta(\mathcal{T}, \beta) - \frac{c}{\rho}\right]\left[1 - H\left(\Theta(\mathcal{T}, \beta)\right)\right]g(\beta)d\beta$$

$$\leq \int_0^1 \rho\left(\bar{d} - \beta A\right) \cdot \frac{\left[1 - H\left(\Theta(Q, \Pi^*, \pi^*, \kappa, \beta)\right)\right]^2}{h\left(\Theta(Q, \Pi^*, \pi^*, \kappa, \beta)\right)}g(\beta)d\beta, \tag{69}$$

which indicates $\tilde{W}(\Pi, \pi) \leq \tilde{W}(\Pi^*, \pi^*)$. Moreover, π^* is unique according to (65) and (66). $\blacksquare$

APPENDIX D
PROOF OF THEOREM 3

We prove Theorem 3 by deriving the MNO's profit-maximizing data cap. Under the optimal pricing strategy, the MNO's expected total profit is given by

$$\tilde{W}(Q, \Pi^*(Q, \kappa), \pi^*, \kappa)$$
$$= \frac{\left[\bar{d} - \bar{\beta}A_\kappa(Q)\right](\pi^* - c)^2}{\rho}h\left(\frac{\pi^*}{\rho}\right) - z \cdot Q, \tag{70}$$

which is concave on the data cap Q, since $A_\kappa(Q)$ is convex on Q and $z \cdot Q$ is linear on Q. We take the derivative of (70) with respect to the data cap Q, and obtain the following optimality condition

$$\frac{-\bar{\beta}A'_\kappa(Q^*)(\pi^* - c)^2}{\rho}h\left(\frac{\pi^*}{\rho}\right) - z = 0, \tag{71}$$

which is equivalent to

$$-A'_\kappa(Q^*) = \frac{z \cdot \rho}{\bar{\beta} \cdot (\pi^* - c)^2 \cdot h\left(\frac{\pi^*}{\rho}\right)}. \tag{72}$$

Recall that the optimal per-unit fee π^* satisfies

$$H\left(\frac{\pi^*}{\rho}\right) + \frac{\pi^* - c}{\rho} \cdot h\left(\frac{\pi^*}{\rho}\right) = 1, \tag{73}$$

which means that

$$\pi^* - c = \rho \cdot \frac{1 - H\left(\frac{\pi^*}{\rho}\right)}{h\left(\frac{\pi^*}{\rho}\right)}. \tag{74}$$

We substitute (74) into (72), then obtain

$$A'_\kappa(Q^*) = \Omega(\rho, c, z), \tag{75}$$

where $\Omega(\rho, c, z)$ is given by

$$\Omega(\rho, c, z) = \frac{z \cdot h\left(\frac{\pi^*}{\rho}\right)}{\bar{\beta}\rho\left[1 - H\left(\frac{\pi^*}{\rho}\right)\right]^2}. \tag{76}$$

$\blacksquare$

APPENDIX E
PROOF OF LEMMA 2

We prove Lemma 2 by computing $A_\kappa(0)$, $A_\kappa(D)$, $A'_\kappa(0)$, and $A'_\kappa(D)$, and showing the convexity of $A_\kappa(Q)$ on Q. Recall that the mathematic expression of $A_\kappa(Q)$ is

$$A_\kappa(Q) = \mathbb{E}_{d,\tau}\left\{[d - Q^e_\kappa(\tau)]^+\right\}$$
$$= \sum_\tau \sum_d [d - Q^e_\kappa(\tau)]^+ f(d)p_\kappa(\tau). \tag{77}$$

E.1 Compute $A_\kappa(0)$ and $A_\kappa(D)$

When substituting $Q = 0$ and $Q = \bar{d}$ into $A_\kappa(Q)$, respectively, we obtain

$$A_\kappa(0) = \sum_d [d - 0]^+ f(d) = \bar{d},$$
$$A_\kappa(D) = \sum_\tau \sum_d 0 \cdot f(d)p_\kappa(\tau) = 0. \tag{78}$$

E.2 Compute $A'_\kappa(0)$, and $A'_\kappa(D)$

We compute the derivative of $A_\kappa(Q)$ with respect to Q as follows:

$$A'_\kappa(Q) = \lim_{\Delta Q \to 0^+} \frac{A_\kappa(Q + \Delta Q) - A_\kappa(Q)}{\Delta Q}. \tag{79}$$

When substituting $Q = 0$ into (79), we obtain $A'_\kappa(0)$ as following

$$A'_\kappa(0) = \lim_{\Delta Q \to 0^+} \frac{A_\kappa(\Delta Q) - \bar{d}}{\Delta Q}. \tag{80}$$

Now we can compute $A'_0(0)$ by substituting $A_0(\Delta Q)$ into (80), as follows

$$A'_0(0) = \lim_{\Delta Q \to 0^+} \frac{\sum_{d=0}^D [d - \Delta Q]^+ f(d) - \bar{d}}{\Delta Q} = -1. \tag{81}$$

Similarly, we are able to show that $A_\kappa(0) = -1$ for any $\kappa \in \{0, 1, 2, 3\}$.

Furthermore, when substituting $Q = D$ into (79), we obtain $A'_\kappa(D)$ as follows

$$A'_\kappa(D) = \lim_{\Delta Q \to 0^-} \frac{A_\kappa(D + \Delta Q)}{\Delta Q} \tag{82}$$

Now we can compute $A'_0(D)$ by substituting $A_0(D + \Delta Q)$ into (82), as follows

$$A'_0(D) = \lim_{\Delta Q \to 0^-} \frac{\sum_{d=0}^D [d - D - \Delta Q]^+ f(d)}{\Delta Q} = 0. \tag{83}$$

Similarly, we are able to show that $A_\kappa(D) = 0$ for any $\kappa \in \{0, 1, 2, 3\}$.

E.3 Convexity

According to (77), $A_\kappa(Q)$ is the nonnegative weighted summation over $[d - Q^e_\kappa(\tau)]^+$. Specifically, $[d - Q^e_\kappa(\tau)]^+$ is given by

$$[Q - Q^e_\kappa(\tau)]^+ = \begin{cases} [d - Q]^+, & \text{if } \kappa = 0, \\ [d - Q - \tau]^+, & \text{if } \kappa \in \{1, 2\}, \\ [d - 2Q - \tau]^+, & \text{if } \kappa = 3, \end{cases} \tag{84}$$

which indicates that $[d - Q^e_\kappa(\tau)]^+$ is always convex in Q for any $\kappa \in \{0, 1, 2, 3\}$. Considering the nonnegative weighted summation is a convexity-preserving operation, thus $A_\kappa(Q)$ is convex in Q. $\blacksquare$

APPENDIX F
VARIANCE OF DEMAND DISTRIBUTION

We investigate the impact of the demand distribution variance by considering two different values. Specifically, we consider two log-normal distributions, which have the same mean value of 1GB but different variances (i.e., $\sigma = 0.8, 0.4$). In the following, we first show the distributions

(a) $\sigma = 0.8$ (b) $\sigma = 0.4$

Fig. 10: $A_\kappa(Q)$ vs Q and $f(d)$ vs d under different variances.

(a) $\sigma = 0.8$ (b) $\sigma = 0.4$

Fig. 11: Performance gain under different variances.

and the expected overage payments, then compare the performance gains under the two values of variances.

Fig. 10 shows the demand distributions and the corresponding expected overage payment. The horizontal axis in each sub-figure corresponds to the MNO's QoS. By comparing Fig. 10(a) and Fig. 10(b), we note that given the same data mechanism (e.g., $\kappa = 1$ with two red curves in the two sub-figures), a larger demand variance leads to higher expected overage payments. The impact of time-flexible data mechanism on reducing overage payment is stronger under a larger variance, since the differences among the three curves with markers in Fig. 10(a) is larger than that in Fig. 10(b).

Fig. 11 plots the performance gain of the time-flexible data mechanisms (compare with the traditional data mechanism) under the two values of variance. We note that a larger variance leads to a higher gain for both MNO's profit gain and users' payoff gain.

Based on the above discussions, we find that the time-flexible data mechanism plays a significant role, especially when the users' demand variance is relatively large.