SMART DATA PRICING IN WIRELESS DATA NETWORKS: AN ECONOMIC SOLUTION TO CONGESTION

ZHANG LIANG

Ph.D

The Hong Kong Polytechnic University

2016

THE HONG KONG POLYTECHNIC UNIVERSITY

DEPARTMENT OF COMPUTING

Smart Data Pricing in Wireless Data Networks: An Economic Solution to Congestion

ZHANG Liang

A Thesis Submitted in Partial Fulfillment of

the Requirements for the Degree of

Doctor of Philosophy

December 2015

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

_____ZHANG Liang_____(Name of student)

ABSTRACT

As the popularity of smart mobile devices, together with bandwidth-intensive applications, the data traffic for wireless data networks has grown tremendously in the past few years. This poses challenges for the network operators to improve or even maintain network quality of the system. People are concerned whether such demand increase can be fulfilled by simple infrastructure expansion. On the other hand, it also brings huge financial burden to the Internet service providers (ISPs) since supporting such demand-supply gap requires large investments. The pricing of data traffic and other services is central to the core challenges of network management, growth sustainability and monetization supports. The *smart data pricing* has great potential to tackle the issue of surging demand for network operators and may also bring great benefit to consumers, Internet service providers, and content providers.

In this thesis, we analyze the two main pricing proposals in smart data pricing, i.e., *time dependent pricing* (short for TDP) and *sponsored data plan* (short for SDP). We try to understand the rationale behind the two pricing models, as well as their impacts to the wireless data market, in particular, who will benefit and who will be hurt from these schemes. We also propose and analyze a new pricing proposal, *time dependent sponsoring* (short for TDS), that combines the advantages of TDP and SDP.

First, we focus on the *time dependent pricing*, which is a promising pricing method to relieve the congestion caused by the surging traffic demand. TDP captures the time-variation characteristic of demand by charging users dynamically over time and has the potential to even out time-of-the-day fluctuations in bandwidth consumption. We explore the design space of TDP. In particular, we focus on a number of schemes, e.g., the usage-based scheme, the flat-rate scheme, and a mixture of them which called a cap scheme.

Our main findings include: 1) the ISP obtains a higher profit with usage-based (or flat-rate) scheme if the capacity is insufficient (or sufficient); 2) the usage-based scheme usually achieves a higher consumer surplus and more efficient traffic utilization than the flat-rate scheme; and 3) the cap scheme is strongly preferred by the ISP to further increase its revenue.

Second, we analyze the *sponsored data plan*, a recent pricing proposal, i.e., when accessing contents from a particular content provider (short for CP), end users do not need to pay for that volume of traffic consumed, but the CP will sponsor for this data consumption. We build a two-class service model to analyze the consumers' traffic demand under the sponsored data plan with consideration of QoS. We use a two-stage Stackelberg game to characterize the interaction between CPs and the ISP and reveal a number of important findings. Our conclusions are: 1) When the ISP's capacity is sufficient, the sponsored data plan benefits consumers and CPs in the short run, but the ISP does not have incentives to further improve its service in the long run. 2) When ISP's capacity is insufficient, the ISP and end users may achieve a win-win trade, while the ISP and CPs always compete for the revenue. 3) The sponsored data plan may enlarge the unbalance in revenue distribution between different CPs; CPs with higher unit income and poorer technology support are more likely to prefer the sponsored data plan.

Third, we propose and study one new smart data pricing scheme, *time dependent sponsoring*, i.e., content providers can decide when and how much to sponsor their traffic. The main novelty of TDS is its potential to improve Internet resource utilization by migrating data traffic from peak times to valley times. We formulate a Stackelberg game model to study the interactions between the ISP, CPs and users, and derive the optimal sponsoring fractions over different times under TDS. In particular, we develop a dynamic programming algorithm to solve a non-convex optimization in sponsoring decisions. We find that: 1) TDS improves the CPs' bandwidth utilization, CPs' profit and consumers' welfare for slightly patient strategic users; but may result in controversial effects for

highly patient strategic users, and 2) when CPs provide different subsidizations to different groups, the bandwidth usage can be improved significantly and so are CPs' profit and consumers' welfare, and 3) well designed TDS reduces the waste of capacity and thus improves social welfare and ISP's profit, compared with TDP.

Keywords: smart data pricing, time dependent, sponsoring, Stackelberg game.

PUBLICATIONS

Conferences

1. **Liang Zhang**, Weijie Wu and Dan Wang, "TDS: Time-Dependent Sponsored Data Plan for Wireless Data Traffic Market", in *Proceeding of IEEE International Conference on Computer Communications (INFOCOM'16)*, April 10-15, 2016, San Francisco, CA, USA.

2. **Liang Zhang**, Weijie Wu and Dan Wang, "Sponsored Data Plan: A Two-Class Service Model in Wireless Data Networks", in *Proceeding of ACM SIGMETRICS*, June 15-19, 2015, Portland, Oregon.

3. **Liang Zhang**, Abraham Hang-yat Lam, and Dan Wang, "Strategy-proof Thermal Comfort Voting in Buildings", in *Proceeding of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings (BuildSys'14)*, November 5-6, 2014, Memphis, USA.

4. **Liang Zhang**, Weijie Wu and Dan Wang, "Time Dependent Pricing in Wireless Data Networks: Flat-rates vs. Usage-based Schemes", in *Proceeding of IEEE International Conference on Computer Communications (INFOCOM'14)*, April 27-May 2, 2014, Toronto, Canada.

Workshops and Posters

1. **Liang Zhang** and Dan Wang, "Sponsoring Content: Motivation and Pitfalls for Content Service Providers", in *IEEE INFOCOM Workshop on Smart Data Pricing (SDP'14)*, April 27-May 2, 2014, Toronto, Canada.

2. **Liang Zhang**, Weijie Wu and Dan Wang, "The Effectiveness of Time Dependent Pricing in Controlling Usage Incentives in Wireless Data Networks", in *Proceeding of ACM SIGCOMM*, August 12-15, Hong Kong.

ACKNOWLEDGEMENTS

Thanks to my supervisor, Prof. Dan Wang, for offering me opportunity to pursue my PhD study and supporting me continuously over these years. His expertise, vast knowledge and skill in many research areas really impressed me. The relaxed research environment provided me enough freedom to explore new research topics following my interest. Thanks to Dr. Weijie Wu, for providing early inspiration, a perfect introduction to paper writing and mathematical modeling. His strict requirement and professional guidance for doing research helped me to become a well-trained researcher. Thanks to my friends and group members of Prof. Wang's research group. The friendly working environment created by them and the time worked and discussed with them will become a nice memory in my life. Finally, thanks to my parents and sister for their long-term support and encouragement, which allow me to follow my dreams and finish this thesis.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Background

The data traffic for wireless data networks has grown tremendously in the past few years. It is reported via Cisco Visual Networking Index that the global mobile data traffic will increase nearly tenfold between 2014 and 2019. In particular, the average smartphone will generate 4.0 GB of traffic per month in 2019, a fivefold increase over the 2014 average of 819 MB per month [19]. The growth in per-device data consumption is fueled by the demand for bandwidth-intensive applications such as Dropbox, YouTube, etc. For example, the mobile video is predicted to grow at a compound annual growth rate of 66 percent between 2014 and 2019, the highest growth rate of any mobile application category [19]. Many Internet applications can be categorized as cloud applications, so as to overcome the memory capacity and precessing power limitation of mobile devices. The trend of intensive interactions of mobile devices and clouds improves the per-device data consumption further. The constantly updated mobile devices are also contributing to the increase in traffic volume, e.g., Apple's iPhone 6S has almost threefold screen resolution compared with iPhone 5S, allowing videos of higher quality to be streamed to the device [6].

This surging wireless traffic demand poses challenges for the Internet service providers (ISPs) to improve or even maintain network quality of the system. It forces ISPs to expand their current wired and wireless capacity via acquiring additional spectrum, deploying more base stations and Wi-Fi hopspots, and adopting new technologies like 4G and LTE/5G. Despite of this, people are still concerned whether such surging demand increase can be fulfilled by simple infrastructure expansion. On the other hand, the surging demand also brings huge financial burden to the ISPs since supporting such

demand-supply gap requires large investments. As such, ISPs have tried to transfer some of their network costs to consumers. For instance, AT&T and Verizon announced in July 2012 that they were offering shared data plans for all new consumers and discontinuing their old plans [16]. Many consumers suffered much higher bills due to such plans. The popularity of usage-tracking and data compression apps helped consumers to manage the traffic usage so as to avoid overage fees and save money. Beside consumers, new financial sources are also explored via new proposals. For example, AT&T finally announced its sponsored data program in January 2014, after a long time of planning, which allowed content providers (CPs) to sponsor the traffic from their users [8].

The pricing of data traffic and other services is central to the core challenges of network management, growth sustainability and monetization supports. The dominant pricing scheme in today's Internet is flat-rate pricing, i.e., ISPs charge a fixed service fee for unlimited data usage during a time period (e.g., one month), and within this period, users can consume the data traffic anytime they want. This pricing scheme is successful in broadband (i.e., wired) networks, where bandwidth resources are usually adequate, and indeed contributed greatly to the Internet growth in the past years. However, this type of pricing strategy usually encourages data usage from customers, which is not always suitable for wireless services where bandwidth is inadequate. Even in broadband networks, the growing congestion caused by increasing traffic demand, which has much slower growth rate than mobile data, has also motivated ISPs to abandon such flat-rate pricing. For example, Comcast have tried to cap their wired network users to 300 GB per month [9]. Even in early year 2008, Comcast also had made headlines with their decision to throttle Netflix as a way to curb network congestion [41].

Currently, ISPs propose pricing plans with a cap in wireless data networks. Users are either not allowed or highly charged for consuming traffic volume beyond this cap, so as to limit traffic demand and thus avoid serious traffic congestion. The cap is usually conservative; for example, Google revealed that almost 85% of the plans offer less than 10 GB data per month, and 36% offer less than 1 GB per month [28]. Such data caps would be easily reached in less than seven hours under the current 3G bandwidth [65]. On the

other hand, the traffic cap plans still cannot avoid the congestion caused by increasing demand [59]. Consumers usually suffer serious congestion, especially when they access the Internet via wireless data networks during peak time. Internet Service Providers (ISPs) are therefore turning to new pricing and penalty schemes so as to manage the demand effectively and also balance the cost of infrastructure expansion.

The *smart data pricing* thus attracts great attention to tackle the issue of surging demand for network operators. Instead of traditional methods of stunting the growth of demand directly via small caps, usage-based scheme, limited connection speed, etc., the *smart data pricing* alleviates the congestion caused by surging demand with more flexibility, e.g., traffic shifting, offloading, etc., which also brings great benefit to the main stockholders in wireless data networks, i.e., consumers, Internet service providers, and content providers. Generally, we classify the possible smart data pricing schemes into three types according to the methodologies of relieving congestion: 1) improving the efficiency of capacity utilization via traffic shifting, e.g., time dependent pricing, congestion aware pricing, etc., and 2) reducing traffic demand via mobile data offloading, e.g., Wi-Fi offloading, small cell offloading, opportunistic offloading, etc., and 3) increasing the traffic capacity via discovering new financial supports, e.g., sponsored data plan and priority pricing, etc. But any changes in pricing and accounting mechanisms should be carefully designed to avoid the potential negative effects for the entire network ecosystem.

The concept of demand regulation via pricing is not just limited to wired and wireless data networks. The pricing mechanisms have also been widely used in other fields like electricity and transportation networks to regulate demand and thus avoid serious congestion or overloading. For example, Borenstein [10] considered retail real-time pricing (RTP) in electricity industry, and Paschalidis and Tsitsiklis [63] proposed congest-dependent pricing in communication networks. However, there exist several features in data networks, which are different from that of other networks significantly, thus resulting in much different pricing paradigms. First, unlike other utilities such as water, electricity, gas or oil, bits are not physical goods, and therefore cannot be truly "con-

sumed" in any meaningful sense. The bandwidth in Internet is shared via multiple users through statistical multiplexing and therefore is reusable and only temporarily in use at an give time. This explains the facts that unlimited data plan is common in Internet while the metered scheme is dominated in electricity networks, water networks, etc. Second, the interaction between an ISP and its customers is much easier due to well-designed interfaces for most mobile devices. Moreover, the data-traces applications also help users to monitor and manage the data traffic. The pricing innovation required real-time user reaction and complex algorithms is possible in wireless data networks. In contrast, most household devices connected to the electricity network do not have the ability of usage monitoring. Third, many online applications such as movie downloads, software update are time elastic such that traffic shift can be carried out efficiently. Fourth, the issues for mobile users like consumer billing, privacy, and security associated with wireless data pricing are much more complex and sensitive than that in electricity and transportation networks.

1.2 Research Framework

In this thesis, we analyze the two typical pricing proposals in smart data pricing, i.e., *time dependent pricing* (short for TDP) and *sponsored data plan* (short for SDP). They alleviate congestion caused by surging demand from two different aspects, i.e., improving capacity efficiency and discovering new financial support. We try to understand the rationale behind the two pricing models, as well as their impacts to the wireless data market, in particular, who will benefit and who will be hurt from these schemes. We also propose and analyze a new pricing proposal, *time dependent sponsoring* (short for TDS), that combines the advantages of *time dependent pricing* and *sponsored data plan*. Figure 1.1 demonstrates the comparison between smart data pricing and traditional pricing schems. It also illustrates the relationship between these three pricing proposals (red and thicker lines) with other pricing proposals in wireless data networks. We analyze these three pricing schemes from different views and thus propose much different models to better explain and analyze them. In this section, we demonstrate the problems and these

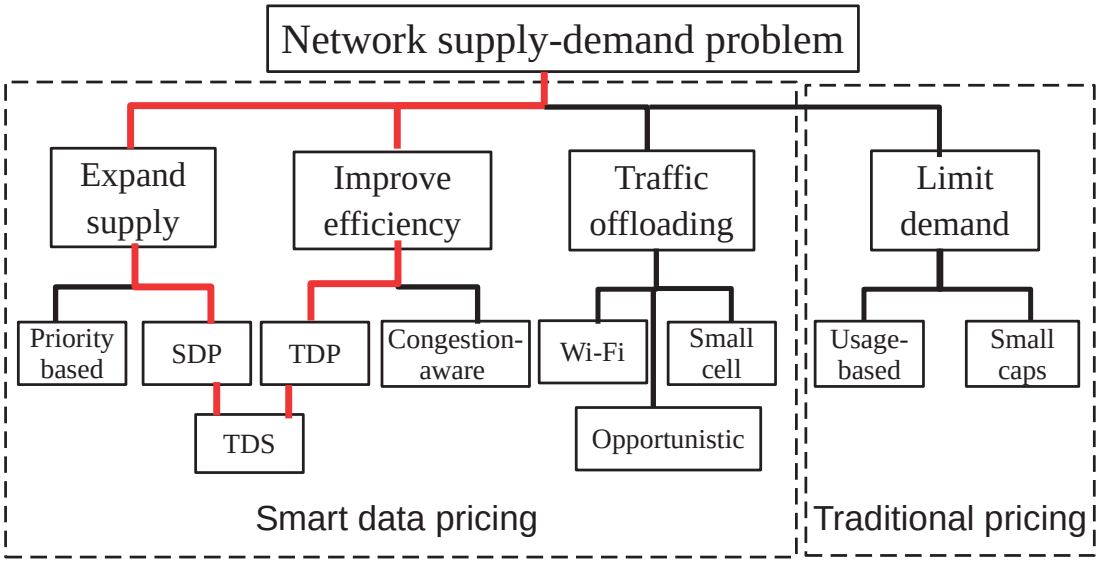


Figure 1.1: Smart data pricing vs. traditional pricing

models briefly and show our main contributions.

1.2.1 Time Dependent Pricing

In Chapter 3, we analyze the *time dependent pricing*, which is a promising pricing method to relieve the congestion caused by the surging traffic demand. TDP captures the time-variation characteristic of demand by charging users dynamically over time. It has great potential to even out time-of-the-day fluctuations in bandwidth consumption. Authors in [29, 76] declared that time dependent pricing can migrate demand from peak to off-peak times, and Ha et al. [29] designed a mechanism to do it via rewarding users. We argue that besides the migration effect, a high or low price can significantly change users usage pattern in peak and valley times. For instance, in WeChat, one may use video chat when the price is low, and switch to text chat when the price is high. On the other hand, much is unknown in its design space, in particular, what is the most effective and profitable time dependent pricing scheme, and how to incentivize the users to use the wireless bandwidth in an efficient manner. These problems are challenging because users' demands are highly dynamic and heterogenous, and there are complicated interactions between the users and ISPs.

To address these problems, we explore the design space of time dependent pricing in a monopoly ISP market and analyze three types of schemes: 1) flat-rate scheme, where a single price is proposed for unlimited usage (but this price can change from peak to valley times); 2) usage-based (or metering) scheme, where the total price equals to the unit price times the amount of usage (again, the unit price can also be time-varying); and 3) the cap then metered scheme (or cap scheme for short), i.e., setting a limit below which flat-rate scheme is applied and beyond which usage-based scheme is applied [40]. There have been extensive studies on comparison between the flat-rate scheme and the usage-based schem [40, 59], but they are restricted in broadband networks. Up till now, very few works have been focusing on time dependent pricing in wireless networks. Thus, we analyze the design principles of time dependent pricing under wireless environment.

Our Contributions:

- Built a Stackelberg game model to capture the interactions between a set of heterogeneous users and the monopoly ISP;
- Compared the usage-based scheme and flat-rate scheme in time dependent pricing in terms of ISP's profit, consumers' surplus and capacity utilization;
- Found that the ISP obtains a higher profit with usage-based (or flat-rate) scheme if the capacity is insufficient (or sufficient); the usage-based scheme usually achieves a higher consumer surplus and more efficient traffic utilization than the flat-rate scheme;
- Proposed the dynamic cap scheme that combines the advantages of both usage-based scheme and flat-rate scheme;
- Found that the dynamic cap scheme is strongly preferred by the ISP to further increase its revenue.

1.2.2 Sponsored Data Plan

In Chapter 4, we focus on the *sponsored data plan*, a recent pricing proposal, i.e., when accessing contents from a particular CP, end users do not need to pay for that volume of traffic consumed, but the CP will sponsor for this data consumption. It is hoped that this pricing strategy can create a positive cycle: End users are glad to access more contents which will not be counted into their data caps; Content providers can attract more users and views; ISPs can obtain higher revenue to support better quality of service (QoS) and carry out technology upgrade. Nevertheless, a key problem is whether such a plan may lead to unfair competition advantage to certain parties. Opponents, including network neutrality advocators, representatives from public interest groups, concern that such a plan will favor rich and big CPs over small ones [55]. This may impede Internet innovation and ultimately hurt consumers. Proponents, mostly ISPs and some CPs, argue that this plan can promote competition and improve efficiency [5]; ultimately, consumers will benefit from better services and cheaper traffic.

To address this problem, we build a two-class service model to analyze the consumers' traffic demand under the sponsored data plan. Under the plan, there will be a sponsored class services and an ordinary class services. The sponsored class brings higher revenue for CPs than the ordinary class. Yet to join this class, a CP needs to pay a non-trivial premium for each unit of content it delivers. Only a CP whose unit profit is high can afford such a premium. This discriminates CPs in the sense that they need to compete for capital, i.e., the capability they pay a premium for their content, rather than their services, i.e., the quality of the contents or the quality of services.

Our contributions:

- Built a two-class service model to analyze the consumers' traffic demand with consideration of QoS;
- Developed a two-stage Stackelberg game framework and analyzed the interactions between the ISP and CPs;

- Developed efficient polynomial time algorithms to search the equilibrium (or outcome) of Stackelberg game in an exponential solution space;

- Found that when the ISP's capacity is sufficient, the sponsored data plan benefits consumers and CPs in the short run, but the ISP does not have incentives to further improve its service in the long run;

- Found that when ISP's capacity is insufficient, the ISP and end users may achieve a win-win trade, while the ISP and CPs always compete for the revenue;

- Found that the sponsored data plan may enlarge the unbalance in revenue distribution between different CPs; those with higher unit income and poorer technology support are more likely to prefer the sponsored data plan.

1.2.3 Time Dependent Sponsoring

In Chapter 5, we propose a new pricing proposal, *time dependent sponsoring*, i.e., content providers can decide when and how much to sponsor their traffic, based on detailed analysis of the two main pricing proposals, time dependant pricing and sponsored data plan. The main novelty of TDS is its potential to improve Internet resource utilization by migrating data traffic from peak times to valley times. It is non-trivial to analyze TDS. The key challenges include: 1) how to model user behavior as different users may behave differently under TDS, 2) the demands under different times can be correlated due to traffic migration under different subsidizations, 3) the interaction among users, CPs and ISPs is complex, and there need appropriate models for the overall game, comprehensive discussions and interpretations.

Our contributions:

- Proposed TDS, i.e., time dependent sponsoring, and established a Stackelberg game model to capture the interactions among a monopoly ISP, a set of CPs and an arbitrary number of strategic users under this scheme;

- Proposed a dynamic programming based algorithm to solve the non-convex optimization introduced by formulating the ISP's and CPs' decisions under TDS in polynomial time;
- Found that TDS improves the CPs' bandwidth utilization, CPs' profit and consumers' welfare for slightly patient strategic users; but may result in controversial effects for highly patient strategic users;
- Found that when CPs provide different subsidizations to different groups, the CPs' bandwidth utilization can be improved significantly and so are CPs' profit and consumers' welfare;
- Found that well designed TDS reduces the waste of ISP's capacity and thus improves social welfare and ISP's profit, compared with the sponsored data plan.

1.3 Structure of this thesis

The rest of this thesis is organized as follows: In Chapter 2, we provide a detailed literature review. In Chapter 3, we analyze and compare the usage-based scheme and flat-rate scheme under time dependent pricing in wireless data networks. In Chapter 4, we analyze the sponsored data plan via building a two-class service model. In Chapter 5, we propose and analyze the time dependent sponsoring. In Chapter 6, we present the conclusion and discussion on future work.

CHAPTER 2
LITERATURE REVIEW

This chapter lists the literature review of three typical pricing proposals in smart data pricing, i.e., time dependent pricing, sponsored data plan, and mobile data offloading, from three different aspects of alleviating congestion, shown in Fig. 1.1. Time dependent pricing has the potential to even out the time-of-the-day fluctuation, and thus can improve the capacity efficiency. Sponsored data plan provides the platform for content providers to sponsor their users, which generates new financial sources so as to support the capacity extension. When these two methodologies still cannot catch up the growth of traffic demand, mobile data offloading becomes an efficient method to migrate the traffic from cellular networks to complementary networks such as WiFi, Femtocell and WiMax. All of the three proposals are widely discussed and studied in wireless data networks as the popularity of smart devices and the growth of mobile traffic demand [69].

2.1 Time Dependent Pricing

Time dependent pricing has been extensively studied to address congestion problems in various fields. Borenstein [10] studied retail real-time pricing (RTP) in electricity industry, and Paschalidis and Tsitsiklis [63] proposed congest-dependent pricing in communication networks. Until recently, researchers from academia and industry began to migrate the similar methodology into pricing the wired or wireless network access services [15, 37, 42, 47, 76, 84].

Jiang et al. [37] proposed a model with the time dependent pricing based on users' preference and congestion level. It analyzed the revenue and social welfare loss due to the insufficient information on users.

Lee et al. [42] proposed a token pricing scheme, which views the tokes as currency and allows user to accumulate daily tokens so as to exchange for services during peak hours. This mechanism encourages users to congest the network only when they have a high utility for the service. It provides an implicit time-varying incentives to shift data so as to avoid congestion. The work showed that users make a better use of resources and the social welfare increases under such proposal.

Loiseau et al. [47] proposed a raffle-based scheme for the decongestion of a shared resource. Raffle-based scheme provides a monetary reward in a raffle-like manner, i.e., the expected reward of a user is proportional to his percentage of contribution to the aggregate reduction in peak-time demand. The authors formulated a game-theoretic model for the decongestion problem and demonstrated that raffle-based schemes are useful in addressing congestion problem. The successive works by Loiseau et al. [46,48], compared the benefits of using the raffle-based scheme and time-of-the-day pricing for congestion management. They showed that both schemes can achieve an optimal level of decongestion at a unique Nash equilibrium. However, raffle-based scheme is more robust when the users' sensitivity to congestion is "sufficiently" convex.

Wong et al. [76] considered a model of rewarding users via delaying application sessions. Each session is characterized by a waiting function that reflects the willingness of the user to delay his entire session for given time. The authors showed how to minimize the cost problem by balancing the congestion cost and the reward amount. They designed efficient algorithms to determine the optimal time-dependent prices which basically belong to a time dependent usage-based scheme. Ha et al. [29] extended the work of [76] by presenting the architecture, implementation, and a user trial of the system, called TUBE. TUBE created a price-based feedback control loop between an ISP and its end users: the ISP computes TDP prices while end users are allowed to respond to the offered prices via a graphical user interface.

Ma et al. [49,50] provided an optimal design of time and location aware pricing scheme, which incentivizes users to even out traffic and reduce network congestion. They formu-

lated a two-stage decision problem, and treat it as a bilevel optimization problem. They proposed a derivative-free algorithm to solve the problem for any increasing concave user utility functions.

The above works have been making it a reality to charge the Internet access in a time dependent manner; however, there are very limited understandings on the theoretical rationales of the mechanism design. In particular, much is unknown on how to design a practical and effective time dependent pricing and how to compare various schemes. We find only one recent work from Hande et al. [32] which studies both usage-based and flat-rate schemes. That work considered time-varying consumers' utilities and capacity constraint and studied the strategy of dropping packets. It analyzed a combination of usage-based and flat-rate schemes where a fixed access fee is charged, irrespective of the data rate and a linear flat rate is charged for extra usage. However, the authors considered homogeneous utility function of customers which does not capture the real market. They modeled the problem from an ISP's point of view, but they did not consider the user surplus or social welfare. Our work in Chapter 3 differs from [32] in that 1) we borrow the idea of bundling from Nabipay et al. [57], and consider the user heterogeneity; 2) we rigorously show what factors/conditions make flat-rate or usage-based scheme more profitable; 3) we show how traffic cap strategy combines the advantages of flat-rate and usage-based schemes; and 4) we compare the schemes from a comprehensive viewpoint, including the profit of ISPs and the surplus of customers.

2.2 Sponsored Data Plan

Sponsored data plan has become an attractive research topic since it was proposed by AT&T. Under this plan, content providers can transfer part of their revenue to consumers so as to pursue higher traffic usage and remedy their consumers' low willingness to pay. Similar revenue transfer and sharing also happen between other interest groups with complementary requirements in the Internet [31, 52, 58, 77]. Based on the seminal work of two-sided markets by Armstrong [7], prior works [31, 58] studied the two-sided mar-

kets in the Internet, i.e., CPs and end users are the two sides that interact in a market enabled by the platform of ISPs. Njoroge et al. [58] found that through CP-side pricing, ISPs could extract higher surplus and maintain higher investment levels. Hande et al. [31] set up a two-sided model in a rate allocation market and concluded that subsidizing end users' cost of connectivity by pricing content providers may benefit both end users and CPs. Xu et al. [79] proposed a cooperative profit-distribution model for eyeball ISPs and peer-assisted content providers based on Nash Bargaining Solution. Authors of [52, 77] studied the profit-sharing mechanism of multi-lateral ISP settlements. Ma et al. [52] proposed a Shapley profit-sharing mechanism. At the Nash equilibrium, the routing and connecting/peering strategies maximize aggregate network profits. Wu et al. [77] proposed a Nash bargaining process and found that all ISPs are simultaneously better than the noncooperative equilibrium.

The sponsored data plan faces new challenges due to the complexity of users' behaviors and attracts lots of works [3,4,38,51,85]. From the economic point of view, Andrews et al. [4] studied the contractual relationship between CPs and ISPs with random demand. They concluded that a coordinating contract can maximize the total system profit, and that the additional profit caused by sponsored data plan can be split between CPs and ISPs in an arbitrary manner. Our previous work [83] formulated a competition problem between one large CP and one small CP to show whether sponsored content indeed favors large and rich CPs. Ma [51] captured the regulated subsidization competition among CPs under a neutral network and concluded that certain CPs might be harmed with a main reason being the high access prices instead of the existence of subsidization. Joe-Wong et al. [38] formulated the interaction among ISPs, CPs and heterogeneous users and derived their optimal behaviors. They found that sponsorship favors less cost-constrained CPs and more cost-constrained users, exacerbating CP inequalities but making user demand even more. From technical point of view, Raj et al. [65] developed a new computing abstraction, called SIMlet, based on the idea of split billing. Andrews et al. [3] developed a detailed methodology for extracting the parameters required by models in their previous work [4] and discussed how to select the proper sites to join in

the sponsored data plan.

The above works provide initial analysis on the sponsored data plan; however, the understanding of sponsored data plan is still limited. In particular, much is unknown on whether this strategy brings unfair competition and hurts consumers, as well as how end users behave and QoS changes upon the adoption of this plan. In Chapter 4, we model the sponsored data plan as a two-class service model, which has been widely adopted by [33, 45, 54, 70, 82]. Xu [45] provided the technical support for two-class services with different quality of services by proposing a multicast protocol. Shetty et al. [70] investigated the effects of transition from a single-service class to two-service classes in the Internet by considering the interaction between end users and multiple ISPs. Yuksel et al. [82] focused on transit ISPs and quantified the extra capacity requirement for an over-provisioned classless network compared with the class network. Hermalin and Katz [33] examined the welfare effects of product-line restrictions and analyzed the case of two technologically restricted quality levels. However, all the above works focused on the ISP's side but not the CPs' choices. We have only found that Ma and Misra [54] considered the problem from similar directions with us. We adopt a similar methodology with Ma and Misra [53,54] to obtain the equilibrium of the Stackelberg game. However, our work does not consider QoS differentiation, but focuses on the sponsored data plan.

On the other hand, all these studies, in particular [51], have demonstrated that the sponsored data plan may vitalize the Internet growth. Intrinsically, the sponsored data plan can establish a more balanced finance model. Nevertheless, interpretations from studies [38, 83] also indicate that the sponsored data plan may lead to greater data traffic; since it is the CPs' incentive to deliver more content to end users and it is the users' incentive to consume more when others are paying. The work in Chapter 5, for the first time, focuses on how to sponsor data so as to avoid worse congestion with the current insufficiency of traffic bandwidth. We proposed and studied time dependent sponsoring. In time dependent sponsoring, CPs can decide the fraction of traffic to sponsor to end users for a given time, and that fraction may vary over time. Thus, it has the potential to improve resource utilization

2.3 Mobile Data Offloading

Mobile data offloading, allowing end users to transfer data originally targeted to flow through the mobile/cellular networks via complementary wireless technologies, like Wi-Fi, Femtocell, etc., has become a hot topic recently. Obviously, it has the potential to alleviate congestion caused by surging demand in cellular networks and make better usage of available bandwidth resources in the complementary networks. Besides these, mobile data offloading also has the potential to increase the overall throughput, reduce the content delivery time, and improve the energy efficiency. The challenges of mobile data offloading, ranging from infrastructure coordination, mobility of users [44], service continuity [81] to pricing and business models [39], have attracted great interest in the research community [2, 67].

From the technical view, Lee et al. [44] presented a quantitative study on the benefit of offloading data through Wi-Fi APs bringing to network providers and end-users. The collected traces of availability for Wi-Fi during two and a half weeks in Seoul, Korea, reveal that non-delayed offloading could relieve a large portion of cellular traffic. Hu et al. [34] focused on the QoS improvement as a function of the number of APs due to non-delayed offloading. Based on an accurate radio propagation model in an urban scenario, the author carried out simulations and disclosed a linear increase in the average throughput per user, as the density of APs increases. The work [13], studied AP deployment problem aiming at maximizing the fraction of offloaded traffic. For delayed offloading, the work [71] proposed prediction based offloading, which combines the prediction of node mobility with the knowledge of the geo-localization of fixed APs to enhance the offloading process. Dimatteo et al. [22] proposed a network-centered architecture called MADNet, which integrates multiple comunication methods like cellular, Wi-Fi APs, and mobile-to-mobile communications, and showed through simulations that almost half of the cellular traffic could be offloaded by only a few hundreds APs deployed citywide. Ristanovic et al. in [68] proposed energy-efficient offloading algorithms for delay tolerant applications and showed that they can offload a significant amount of traffic and also extend the battery lifetime.

From the economic view, Joe-Wong et al. [39] developed a model of user adoption for a base technology (e.g., 3G) and a bundle of the base plus a supplementary technology (e.g., 3G + WiFi). The author showed that user adoption converges to a unique, stable equilibrium point, and derived analytical conditions under which nonintuitive adoption behaviors occur. Gao et al. [25] analyzed the cooperation and agreement of mobile cellular network operators (MNOs) and AP owners (APOs) by using Nash bargaining theory. Iosifidis et al. [36] studied a market where MNOs lease third party-owned Wi-Fi or femtocell access points to offload their mobile data traffic on demand. Based on particular characteristics and challenges of wireless networks, the authors proposed an iterative double-auction mechanism, which satisfies the desirable economic properties and maximizes the welfare of the market. Lee et al. [43] proposed a two-stage sequential game model to study the economic profits gained by mobile service providers and by users due to delayed Wi-Fi offloading. Zhou et al. [86] investigated the tradeoff between the amount of offloaded traffic and the users' satisfaction. They proposed a novel incentive framework based on reverse auctions to motivate users to leverage their delay tolerance for cellular traffic offloading. Cheung et al. [18] considered the QoS requirement of the Wi-Fi offloading and aimed to achieve a good tradeoff between the user's payment and its QoS characterized by the file transfer deadline. Dong et al. [23], on the other hand, proposed iDEAL, a novel auction-based incentive framework, which allows a cellular service provider to leverage resources from third-party resource owners on demand by buying capacity through reverse auctions.

Although there already exist lots of studies on data offloading in both of economic and technical views, few of them consider the combination among mobile data offloading, time dependent pricing, and sponsored data plan. Only Cheung et al. [17] studied the congestion-aware network selection and data offloading problem in an integrated cellular Wi-Fi system. They formulated the interactions of the users' congestion-aware network selection decisions across multiple time slots as a non-cooperative network selection game (NSG) and proved that the NSG is equivalent to a congestion game. The combination between time dependent pricing and mobile data offloading or sponsored data plan

and mobile data offloading, may generate more benefit and also bring new problems. For example, the current sponsored data plan allows each CP to fully or partially sponsor its consumers' traffic usage directly. An alternative way is to sponsor its users via providing free WiFi offloading, which is much cheaper and more efficient. We will focus on this problem as our future work.

CHAPTER 3

TIME DEPENDENT PRICING IN WIRELESS DATA NETWORKS: FLAT-RATE VS. USAGE-BASED SCHEMES

3.1 Overview

With the advances of bandwidth-intensive mobile device such as smart phones, tablet computers, etc., the data traffic for wireless data networks has grown tremendously in the past few years. It is reported a further increase by more than ten times of the current volume is expected in the next five years [20]. This poses challenges for the network operators to consistently provide good quality services. There are studies addressing this problem from technical points of view, including data measurement [35], caching designs [64], smart spectrum utilization [78], and architectural redevelopment [80]. Nevertheless, researchers also debate that whether such demand increases can be fulfilled by technical solutions only [14].

In this chapter, we consider this problem from a pricing point of view. To see our motivation, on one hand, the traffic demand is highly volatile over time, e.g., the demand in peak hours can be more than ten times than that in valley hours [29]. It is neither physically easy nor economically profitable to purely rely on technical solutions to meet the extreme peak demand. On the other hand, users' behaviors lead to volatile traffic demands; and pricing has been proven as an effective way to shape users' behaviors [73, 74]. For example, by charging a higher price, users may choose to use low-bandwidth applications or reduce unnecessary consumption during the peak times.

The dominant pricing scheme in today's Internet is time-independent flat-rate pricing, i.e., Internet service providers (ISPs) charge a fixed service fee for unlimited data usage during a time period (e.g., one month), and within this period, users can consume the data traffic anytime they want. This is successful in broadband (i.e., wired) networks

as these networks own adequate bandwidth resources. However, this type of pricing strategy usually encourages data usage from customers, which is not always suitable for wireless services where bandwidth is inadequate. For example, WeChat, a very popular mobile social application in China, consumes data traffic to send text, voice and photos. Under the time-independent flat-rate pricing model, people may relentlessly upload photos and "short talk" of trivial errands whenever they want. This causes increasing congestion problems since people consume traffic during peak hours, and important data transmissions may be delayed or even rejected.

To handle this problem, *time dependent* pricing [29, 76] have been recently introduced for wireless data networks. It considers the time variance feature of users' demands, and charges the users *dynamically over time*. Such pricing has been emerging recently in practice. For instance, BSNL in India offers unlimited night time (2-8 am) downloads on a monthly data plan of RS 500 (or USD $10); in US, some ISPs have begun experimenting time dependent pricing plans. Authors in [29, 76] declared that time dependent pricing can migrate demand from peak to off-peak times, and Ha et al. [29] designed a mechanism to do it via rewarding users. We argue that besides the *migration* effect, a high or low price can significantly *change* users' usage pattern in peak and valley times. For instance, in WeChat, one may use video chat when the price is low, and switch to text chat when the price is high.

Although time dependent pricing has been proposed, much is unknown in its design space, in particular, what is the most effective and profitable time dependent pricing scheme, and how to incentivize the users to use the wireless bandwidth in an efficient manner. These problems are challenging because users' demands are highly dynamic and heterogenous, and there are complicated interactions between the users and ISPs. In this chapter, we explore the design space of time dependent pricing in a monopoly ISP market, and provide important insights on how to design practical and effective pricing schemes. In particular, we consider three types of schemes: 1) flat-rate scheme, where a single price is proposed for unlimited usage (but this price can change from peak to valley times); 2) usage-based (or metering) scheme, where the total price equals to the

unit price times the amount of usage (again, the unit price can also be time-varying); and 3) the "cap then metered" scheme (or cap scheme for short), i.e., setting a limit below which flat-rate scheme is applied and beyond which usage-based scheme is applied [40]. There have been extensive studies on comparison between the flat-rate scheme and the usage-based scheme [40, 59], but they are restricted in broadband networks. Up till now, very few works have been focusing on time dependent pricing in wireless networks. In this chapter, we analyze the design principles of time dependent pricing under wireless environment. We use a Stackelberg game model to capture the interactions between a set of heterogeneous users and the monopoly ISP, and explore the optimal pricing schemes for the ISP. We evaluate the schemes in terms of the ISP's profit, users' surplus, bandwidth utilization and the effectiveness of bandwidth usage. Our main findings are:

- The ISP obtains a higher profit with usage-based (or flat-rate) scheme if its capacity is insufficient (or sufficient);
- Comparing with the flat-rate scheme, the usage-based scheme usually achieves a higher consumer surplus and a more efficient utilization of the traffic.
- The cap scheme is preferred by the ISP to further increase its revenue, but consumers may not benefit from it.

The rest of the chapter is organized as follows: We discuss the users' service valuation model in Section 3.2. In Section 3.3, we compare the flat-rate and usage-based schemes. We discuss the cap scheme in Section 3.4 and provide numerical results in Section 3.5. Finally, we summarize our work in Section 3.6.

3.2 Users' Service Valuation Model

In this section, we formulate a model on how users evaluate the valuation of any particular service, and based on that, we capture how users decide the amount of traffic to use for any given price. This sets up the basis for analyzing various pricing schemes in later sections.

We consider a time period T (e.g., one day or one month), and divide it into time slots $[t-1,t], t = 1, \cdots, T$. We assume that each user has a valuation on a particular wireless service. In each time slot, a user decides whether and how much to use a service based on his valuation and the service price. Only when his valuation of the service is larger than or equal to the service price, the user will subscribe to such service. For example, if a user thinks that the traffic usage of watching a video brings him a huge cost which is larger than his valuation of the video, he may not watch this video, but he may opt to consume other forms of services (e.g., reading emails).

We assume there are totally I independent services $i = 1, 2, \ldots, I$, and we consider how users decide their valuation on any service i. Let θ_i^t be the maximal possible demand for service i during the time slot $[t-1,t]$[1]. A user can decide to consume any amount of traffic $x_i^t \leq \theta_i^t$. If $x_i^t < \theta_i^t$, it means the user does not consume the maximal demand. This represents that the user consumes partial service (e.g., he discusses with his friend on the most important issues via WeChat but he avoids telling jokes, or he watches the video with screen freezing from time to time). We define $\omega_i^t = x_i^t/\theta_i^t$, which is the ratio between the actual usage and the maximal possible demand. Let c_i denote the users' *per unit valuation* of service i. If $x_i^t = \theta_i^t$, then the user's valuation on this service is $c_i\theta_i^t$ during $[t-1,t]$. If $x_i^t < \theta_i^t$, then his valuation decreases by a certain factor, and we use a satisfaction function: $f_i : [0,1] \to [0,1]$ to represent it. This satisfaction function satisfies $f_i(0) = 0$ and $f_i(1) = 1$. We assume that $f_i(\cdot)$ is a non-decreasing, twice differentiable function and it is concave or convex in the interval $[0,1]$. Thus, the users' valuation for service i during time slot $[t-1,t]$, denoted as Y_i^t, is

$$Y_i^t = c_i\theta_i^t f_i(\omega_i^t). \tag{3.1}$$

We assume c_i is independent of time, but users can have different maximal demands during different slots. For instance, the per unit valuation for WeChat is the same at any time during a day, while the usage demand may be volatile over time. In addition, the

[1]The amount of the traffic demand may change over time. For instance, users' demand on the video may be higher at 11 pm than at 6 am [1].

per unit valuations for different services can also be much different. For example, the per unit value for SMS can be much greater than that of voice [61]. We also assume θ_i^t are non-negative random variables that reflect the heterogeneity of consumers' maximal traffic demand. Given the time slot $[t-1,t]$, the maximal traffic usage for different service i, i.e., θ_i^t, is assumed to be independent of each other. Thus, Y_i^t is a non-negative independent random variable.

We also assume that the valuations of different services are additive. Therefore, given the values of θ_i^t, the value of using all services in $[t-1,t]$ is

$$Y^t = \sum_{i=1}^{I} c_i \theta_i^t f_i(\omega_i^t). \tag{3.2}$$

We denote that θ_i^t has the cumulative distribution function $\Theta_{\theta_i^t}(s_i^t) = \mathbf{Pr}\{\theta_i^t \leq s_i^t\}$ with finite mean u_i^t and finite standard variance σ_i^t. In particular, u_i^t represents the average traffic usage for service i in slot $[t-1,t]$, which is important in our later analysis. We define the joint cumulative distribution function of $(\theta_1^t, \theta_2^t, ..., \theta_I^t)$ as $\Theta_t(\mathbf{s}^t) = \mathbf{Pr}\{\theta_1^t \leq s_1^t, \theta_2^t \leq s_2^t, ..., \theta_I^t \leq s_I^t\}$, where $\mathbf{s}^t = (s_1^t, s_2^t, ..., s_I^t)$.

3.2.1 Discussion on the satisfaction function

The satisfaction functions have different features for various services. For instance, in an online video service like Netflix, users' satisfaction drops rapidly when ω_i^t decreases (i.e., a large gap between the maximal demand and the actual traffic consumption). This is because receiving the data less than the required playback rate leads to frequent screen freeze, which significantly reduces the quality of experience. In contrast, in an online chat service like WeChat, users' satisfaction may still be high even if ω_i^t is low. This is because people can usually use only a few sentences to express the core message, and they can use text chat instead of video chat. In this chapter, we define a user's satisfaction function as follows:

$$f_i(\omega_i) = \omega_i^{\beta_i}, \tag{3.3}$$

where β_i is called the *traffic sensitivity* of service i. Large $\beta_i (> 1)$ represents services with high requirement on integrity, e.g., video service like Netflix; while small $\beta_i (< 1)$

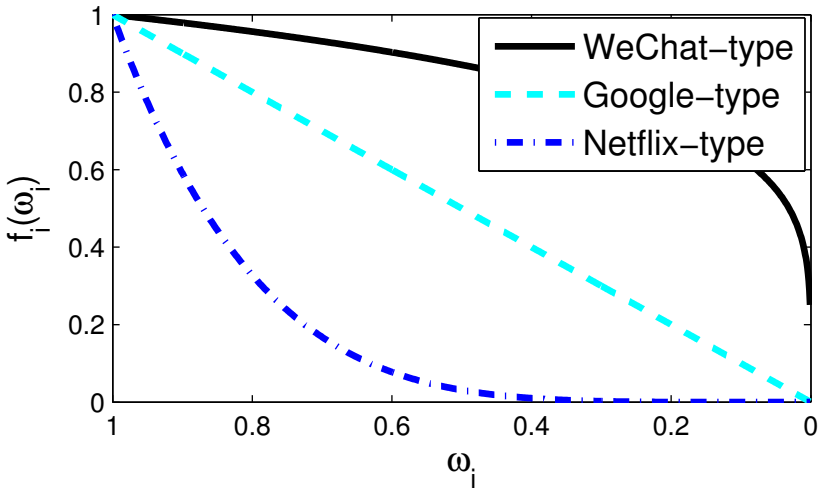


Figure 3.1: Satisfaction function

represents low sensitivity services like WeChat. There are also medium-sensitivity services, e.g., web service like Google. Fig. 3.1 illustrates the satisfaction functions of these three types, with parameters $(c_1, \theta_1, \beta_1) = [0.2, 50, 5]$, $(c_2, \theta_2, \beta_2) = [5, 1, 1]$ and $(c_3, \theta_3, \beta_3) = [1, 10, 0.2]$, respectively. They represent Netflix-type (high maximal demand, high traffic sensitivity), Google-type (low maximal demand, medium traffic sensitivity) and WeChat-type (medium maximal demand, low traffic sensitivity) services. Fig. 3.1 shows that to achieve half of the maximal valuation, Netflix-type services need at least 75% of maximal demand; while Google-type and WeChat-type services only need 50% or 5%. In our later analysis, many results are based on the form of satisfaction function defined in this subsection, and we are interested to observe the impact of traffic sensitivity on the pricing schemes.

3.3 Flat-Rate vs. Usage-Based Schemes

In this section, we formulate a two-stage Stackelberg game model [62] to capture the interactions between the monopoly ISP and the heterogeneous users. The first stage of this game is that the ISP determines the pricing scheme, and the second stage is that the consumers decide whether to join in the network and how much traffic to consume. It is natural to assume that the ISP is the first mover and the consumers are followers that make their decision according the prices. To obtain the Stackelberg equilibrium of the

game, we can use the backward induction [62]. In particular, we first consider the traffic consumption determined by users for any given pricing scheme by the ISP. By knowing the consumers' best responses, the ISP decides its optimal pricing scheme, based on which the traffic consumption of users can be also determined.

Based on this game framework, we will analyze the Stackelberg equilibrium under both flat-rate and usage-based schemes, and we will compare them via a number of performance measures. Since the major cost of an ISP is on infrastructure constructions, we ignore its marginal cost for delivering the data.[2] Therefore, the ISP's profit (or utility) equals the total service fee charged from all users. Let μ be the capacity constraint of the ISP during any time slot, i.e., the maximal amount of traffic that can be provided by the ISP.

3.3.1 Usage-based Scheme

Due to network neutrality rules, we assume that the ISP charges the same price h^t per unit traffic for any kind of services during $[t-1, t]$. We normalize the the total number of users to be one.

In order to analylize the Stackelberg equilibrium, we use backward induction and first consider the second stage of the game, i.e., given h^t, users maximize their utility function by choosing the traffic consumption x_i^t for any service i:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_u(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - h^t \sum_{i=1}^{I} x_i^t \\ \text{s.t.} \quad & 0 \le x_i^t \le \theta_i^t, 1 \le i \le I. \end{aligned} \tag{3.4}$$

The optimal solution always exists and is:

$$x_i^{t*} = \begin{cases} 0 \quad \text{or} \quad \theta_i^t & \text{if } f_i''(\cdot) \ge 0, \\ \theta_i^t \min\left\{1, f_i'^{-1}(h^t/c_i)\right\} & \text{if } f_i''(\cdot) < 0, \end{cases} \tag{3.5}$$

where $f_i'^{-1}(\cdot)$ is the inverse function of first order derivative of the satisfaction function $f_i(\cdot)$, $f_i''(\cdot) \ge 0$ means $f(\cdot)$ is a convex function and $f_i''(\cdot) < 0$ means concavity. When

[2] Even though some arguments state that the marginal cost may be not ignorable, the consideration of such cost does not affect the final result of this paper.

$f_i''(\cdot) \geq 0$, the users' utility for service i is $\max\{0, (c_i - h^t)\theta_i^t\}$. Thus, $x_i^{t*} = 0$ if $h^t < c_i$ or $x_i^{t*} = \theta_i^t$ otherwise. When $f_i''(\cdot) < 0$, $f_i'^{-1}(\cdot)$ is a decreasing function and x_i^{t*} is non-increasing in h^t. The total data consumption for service i from all users is:

$$\begin{aligned} D_i^t(h^t) &= \int x_i^{t*} d\Theta_t \\ &= \begin{cases} 0 \quad \text{or} \quad u_i^t & \text{if } f_i''(\cdot) \geq 0, \\ u_i^t \min\left\{1, f_i'^{-1}(h^t/c_i)\right\} & \text{if } f_i''(\cdot) < 0, \end{cases} \end{aligned} \tag{3.6}$$

where u_i^t means the average data consumption for service i in $[t-1, t]$. The total data consumption cannot exceed the traffic capacity of the ISP,[3] i.e.,

$$\sum_{i=1}^{I} D_i^t(h^t) \leq \mu. \tag{3.7}$$

We next analyze the first stage of the Stackelberg game. Knowing the best responses of users, the ISP maximizes its profit by charging prices that solve the following optimization:

$$\begin{aligned} \max_{\{h^t\}_t} \quad & \Pi_u = \sum_{t=1}^{T} \sum_{i=1}^{I} h^t D_i^t(h^t) \\ \text{s.t.} \quad & \sum_{i=1}^{I} D_i^t(h^t) \leq \mu \quad \forall t. \end{aligned} \tag{3.8}$$

Define $l_u^t = \min\{l \geq 0 : \sum_{i=1}^{I} D_i^t(l) \leq \mu\}$. Since $D_i^t(\cdot)$ is a non-increasing and continuous function, l_u^t means the lowest price such that the total consumption does not exceed the ISP's capacity. Denote the ISP's utility in $[t-1, t]$ as $\pi_u^t(\cdot)$, we have $\pi_u^t(0) = 0$ and $\pi_u^t(\infty) = 0$. Since $\pi_u^t(\cdot)$ is a continuous function, the optimal solution of above optimization exists, which we denote as h^{t*}. The optimal solution $(\mathbf{x}^{t*}, h^{t*})$, obtained by backward induction, is a *Stackelberg equilibrium* of the game, where $\mathbf{x}^{t*} = (x_1^{t*}, ..., x_I^{t*})$. We denote the optimal profit during time slot $[t-1, t]$ as π_u^{t*}, so $\Pi_u^* = \sum_t \pi_u^{t*}$. In particular, when $h^{t*} = l_u^t$, it means the optimal price is to make the traffic consumption equal to the ISP's capacity. We can imagine that if there is no capacity constraint, the Stackelberg equilibrium will induce a larger amount of traffic consumption. So in this

[3]Here, we assume that the traffic capacity is hard capacity and cannot be changed in short time, like previous works [32, 37].

sense, we say the capacity is *insufficient* for usage-based scheme because with a larger μ the ISP can achieve a higher utility. When $h^{t*} > l_u^t$, the capacity is not fully utilized in the Stackelberg equilibrium. In other words, the capacity is *sufficient* for the usage-based scheme.

3.3.2 Flat-rate Scheme

In the previous subsection, we have analyzed the interplay between the monopoly ISP and users under usage-based scheme for time dependent pricing. Now we analyze the flat-rate scheme for time dependent pricing. We still use the two-stage Stackelberg game model and the analysis is quite similar to the previous case. If flat-rate pricing scheme is applied, then the ISP charges a uniform price g^t for unlimited data consumption during $[t-1, t]$, but the price may vary depending on t. We first analyze the second stage game. Given the flat-rate price h^t in slot $[t-1, t]$, each user maximizes its utility function by choosing the traffic consumption x_i^t for any service i:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_f(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - g^t, \\ \text{s.t.} \quad & 0 \le x_i^t \le \theta_i^t, \quad 1 \le i \le I. \end{aligned} \tag{3.9}$$

Since $f_i(\cdot)$ is a non-decreasing function and $f_i(1) = 1$, the optimal solution is $x_i^{t*} = \theta_i^t$ and $U_f(\mathbf{x}^{t*}) = \sum_{i=1}^{I} c_i \theta_i^t - g^t$. This means users always use as much as possible by flat-rate scheme. A user decides to access the network if and only if $U_f(\mathbf{x}^{t*}) \ge 0$. When g^t is high, only those with high valuation of all services will access the network. Thus, the fraction of users accessing the network during $[t-1, t]$ is:

$$\mathbf{Pr}\left\{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t\right\} = \int_{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t} d\Theta_t. \tag{3.10}$$

Now we focus on the first stage game. Knowing the best responses of users' traffic consumption, the ISP maximizes its utility by solving

$$\begin{aligned} \max_{\{g^t\}_t} \quad & \Pi_f = \sum_{t=1}^{T} g^t \mathbf{Pr}\left\{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t\right\} \\ \text{s.t.} \quad & \int_{\sum_{i=1}^{I} c_i \theta_i^t \ge g^t} \sum_{i=1}^{I} \theta_i^t d\Theta_t \le \mu \quad \forall t. \end{aligned} \tag{3.11}$$

Define $H^t(g^t) = \int_{\sum_{i=1}^I c_i\theta_i^t \geq g^t} \sum_{i=1}^I \theta_i^t d\Theta_t$, i.e., users' traffic consumption given price g^t. Since $H^t(\cdot)$ is non-increasing and continuous, the capacity constraint is equivalent to $g^t \geq l_f^t$, where $l_f^t = \min\{l \geq 0 : H^t(l) \leq \mu\}$. Denote the profit function during time slot $[t-1, t]$ for flat-rate scheme as $\pi_f^t(\cdot)$, we have $\pi_f^t(0) = 0$ and $\pi_f^t(\infty) = 0$. Since $\pi_f^t(\cdot)$ is a continuous function, the optimal solution of above optimization exists, denoted as g^{t*}. Therefore, $(\mathbf{x}^{t*}, h^{t*})$ is a *Stackelberg equilibrium* of the game, where $\mathbf{x}^{t*} = (\theta_1^t, ..., \theta_I^t)$. We denote the maximal profit during time slot $[t-1, t]$ as π_f^{t*}. Note that the above optimization problem is also a probility problem. Thus, we can bound this maximal profit by the following lemma.

Lemma 3.3.1. *Denote* $\epsilon^t = I^{-1/3}\left(\frac{\max_i\{c_i\sigma_i^t\}}{\min_i\{c_i u_i^t\}}\right)^{2/3}$. *The optimal profit during slot* $[t-1, t]$ *satisfies:*

$$\pi_f^{t*}\begin{cases} \geq (1-2\epsilon^t)\sum_{i=1}^I c_i u_i^t & \text{if} \quad g^{t*} > l_f^t, \\ \leq \max_i\{c_i\}\mu & \text{if} \quad g^{t*} = l_f^t. \end{cases} \tag{3.12}$$

Proof. We first consider the case of sufficient capacity, i.e., $g^{t*} > l_f^t$. The optimization problem can be simplified as $\pi_f^{t*} = \max_{g^t} g^t \mathbf{Pr}\{\sum_{i=1}^I c_i\theta_i^t \geq g^t\}$. Denote $u = \sum_{i=1}^I c_i u_i^t$ and $\sigma^2 = \sum_{i=1}^I c_i^2(\sigma_i^t)^2$. By letting $g^t = (1-\epsilon)u$, we have

$$\begin{aligned} \pi_f^{t*} &\geq (1-\epsilon)u\mathbf{Pr}\{\textstyle\sum_{i=1}^I c_i\theta_i^t \geq (1-\epsilon)u\} \\ &\geq (1-\epsilon)u(1-\mathbf{Pr}\{\mid \textstyle\sum_{i=1}^I c_i\theta_i^t - u \mid \geq \epsilon u\}). \end{aligned} \tag{3.13}$$

From Chebyshov's inequality, we can know that $\mathbf{Pr}\{\mid \sum_{i=1}^I c_i\theta_i^t - u \mid \geq \epsilon u\} \leq \frac{\sigma^2}{(\epsilon u)^2}$. Combined with Eq. 3.13, we have $\pi_f^{t*} \geq (1-\epsilon)u[1-\frac{\sigma^2}{(\epsilon u)^2}] \geq u[1-\epsilon-\frac{\sigma^2}{(\epsilon u)^2}]$. If we take $\epsilon = (\frac{\sigma}{u})^{2/3}$, it follows that $\pi_f^{t*} \geq u(1-2\epsilon)$. Since $\frac{\sigma}{u} \leq I^{-1/2}\frac{\max_i\{c_i\sigma_i^t\}}{\min_i\{c_i u_i^t\}}$, we have $\epsilon \leq \epsilon^t$. Thus, we prove that $\pi_f^{t*} \geq u(1-2\epsilon_1^t)$.

We next consider the case of insufficient capacity, i.e., $g^{t*} = l_f^t$. It also means that $\int_{\sum_{i=1}^I c_i\theta_i^t \geq g^{t*}} \sum_{i=1}^I \theta_i^t d\Theta_t = \mu$. Denote $c^* = \max_i\{c_i\}$. For any g^t, we have:

$$\begin{aligned} \int_{\sum_{i=1}^I c_i\theta_i^t \geq g^t} \sum_{i=1}^I \theta_i^t d\Theta_t &= 1/c^* \int_{\sum_{i=1}^I c_i\theta_i^t \geq g^t} \sum_{i=1}^I c^*\theta_i^t d\Theta_t \\ &\geq 1/c^* \int_{\sum_{i=1}^I c_i\theta_i^t \geq g^t} g^t dG_t \\ &= 1/c^* g^t \mathbf{Pr}\{\textstyle\sum_{i=1}^I c_i\theta_i^t \geq g^t\}. \end{aligned} \tag{3.14}$$

Then, we have $\pi_f^{t*} = g^{t*}\mathbf{Pr}\{\sum_{i=1}^I c_i\theta_i^t \geq g^{t*}\} \leq c^*\mu$. Therefore, we prove the lemma.

□

Lemma 3.3.1 shows that when I, the number of services, is large enough, the ISP can almost achieve the maximal possible profit (which is $\sum_{i=1}^{I} c_i u_i^t$) by flat-rate scheme when the capacity is *sufficient*, i.e., $g^{t*} > l_f^t$. The intuition is that the flat-rate scheme can *reduce the variance* of users' valuations on different services, so that the ISP can easily set up a single price to attract many users. We can also see that when the ISP's capacity is *insufficient*, i.e., $g^{t*} = l_f^t$, the ISP's maximal profit is constraint by this capacity.

3.3.3 Comparison of Usage-based and Flat-rate Schemes

Now we compare usage-based and flat-rate schemes from various viewpoints. We let the satisfaction function be $f_i(\omega) = \omega^{\beta_i}$, $\beta_i = \beta (0 < \beta < 1)$ and $c_i = c$. Denote $u^t = \sum_{i=1}^{I} u_i^t$. We start our analysis by comparing the ISP's profit, and we have the following theorem.

Theorem 3.3.1. *If $u^t \leq \mu$, then there exists I_0 such that for any $I \geq I_0$, $\pi_f^{t*} > \pi_u^{t*}$; if $u^t \geq \beta^{1/(\beta-1)}\mu$, then $\pi_f^{t*} \leq \pi_u^{t*}$.*

Proof. When $u^t \leq \mu$, the capacity is sufficient for both usage-based and flat-rate schemes. For usage-based scheme, the optimal solution and the maximal profit during $[t-1, t]$ are $h^{t*} = \beta c$ and $\pi_u^{t*} = \beta c u^t$. Note that when the capacity is sufficient, according to the lemma 3.3.1, we have that $\pi_f^{t*} \geq (1 - 2\epsilon^t) c u^t$. By letting $I \geq (\frac{2}{1-\beta})^3 (\frac{\max_i \sigma_i^t}{\min_i u_i^t})^2$, we have $\pi_f^{t*} \geq \beta c u^t = \pi_u^{t*}$.

When $u^t \geq \beta^{1/(\beta-1)}\mu$, the maximal profit of flat-rate scheme is upper bounded by $c\mu$ according to Lemma 3.3.1. The capacity of usage-based scheme is also insufficient. The optimal solution and the maximal profit are $h^{t*} = \beta c (\frac{u^t}{\mu})^{1-\beta}$ and $\pi_u^{t*} = \beta c u (\frac{u^t}{\mu})^{1-\beta} \geq c\mu$. Then, we have $\pi_u^{t*} \geq \pi_f^{t*}$ and this completes the proof. □

The first part of Theorem 3.3.1 shows that when the ISP's capacity is larger than the maximal possible demand from users, and I is large enough, then the ISP will have a higher profit when adopting flat-rate scheme. In fact, flat-rate scheme can almost achieve a profit of cu^t but usage-based scheme can achieve at most $\beta c u^t$. The second part shows

that when the capacity is small, the usage-based scheme can achieve more profit than usage-based; when the traffic sensitivity β is larger, the usage-based scheme achieves higher profit.

We also compare the two pricing schemes from users' viewpoint, and we have the following definition.

Definition 3.3.1. *Consumers' surplus is the difference between consumers' average valuation of services and the service fee charged by the ISP.*

Denote the consumers' surplus of usage-based scheme and flat-rate scheme during time slot $[t-1, t]$ as ψ_u^t and ψ_f^t respectively. We have the following theorem.

Theorem 3.3.2. *If $u^t \leq \mu$, then there exists an I_0 such that for any $I \geq I_0$, $\psi_u^t > \psi_f^t$; if $u^t \geq (1-\beta)^{1/(\beta-1)}\mu$, then $\psi_u^t > \psi_f^t$.*

Proof. We first consider the case $u^t \leq \mu$. It means that capacity is sufficient for both the usage-based scheme and the flat-rate scheme. For the consumers' surplus of flat-rate scheme, we have:

$$\begin{aligned} \psi_f^t &= \textstyle\int_{c\sum_{i=1}^I \theta_i^t \geq g^t} (c\sum_{i=1}^I \theta_i^t - g^t) d\Theta_t \\ &\leq cu^t - (1-2\epsilon^t)cu^t = 2\epsilon^t cu^t. \end{aligned} \tag{3.15}$$

Note that the consumers' surplus of usage-based scheme is $\psi_u^t = (1-\beta)cu^t$. By letting $I > (\frac{2}{1-\beta})^3(\frac{\max_i \sigma_i^t}{\min_i u_i^t})^2$, we have $\psi_u^t > \psi_f^t$. When $u^t \geq (1-\beta)^{1/(\beta-1)}\mu$, we have $\psi_u^t = (1-\beta)cu^\beta(\mu^t)^{1-\beta} \geq c\mu > \psi_f^t$ as desired in the theorem. □

Theorem 3.3.2 shows that the consumers' surplus of usage-based scheme is higher than that of flat-rate scheme when the capacity is large enough or small enough. The underlying reason is that the flat-rate scheme reduces the heterogeneity of users' valuation, so the ISP can charge the price closer to the consumers' valuation and this reduces the consumers' surplus.

We also consider the following two other metrics to compare usage-based and flat-rate schemes.

Definition 3.3.2. *Capacity utilization is the ratio of average data consumption during the whole period T over the largest data consumption during period T*

Definition 3.3.3. *The traffic efficiency (or per-unit traffic valuation) is the consumers' average valuation of services divided by the average traffic consumption.*

The usage-based and flat-rate scheme have different performances for capacity utilization and traffic efficiency. For flat-rate scheme, it can even out the varying valuation for different services, so as to reduce the heterogeneity of users' valuation. This characteristic makes flat-rate scheme attract most of the demand when the capacity is sufficient. When the capacity is insufficient, the capacity is fully utilized. The flat-rate price attracts the consumers with high total valuation, but not high per-unit traffic valuation. This is against improving traffic efficiency even when the capacity is insufficient. For usage-based scheme, it always filters out the traffic with valuation lower than the optimal price per unit even when the capacity is sufficient. When the capacity is insufficient, the monopoly ISP makes higher price per unit to obtain higher profit. This also means higher per-unit traffic valuation. The traffic efficiency is greatly improved. Thus, the flat-rate scheme is more likely to have higher capacity utilization while the usage-based scheme is more likely to have higher traffic efficiency. Theoretically, it is hard to give religious results, but we will validate our analysis via numerical results in later sections.

Remarks: From the ISP's point of view, it achieves a higher profit under the usage-based scheme when the capacity is insufficient, or under the flat-rate scheme when the capacity is sufficient. The proper adoption of flat-rate and usage-based schemes for time-dependent pricing strategy provides an effective method for the monopoly ISP to improve its profit. From the consumers' point of view, the usage-based scheme usually brings a higher consumers' surplus than flat-rate scheme. In addition, the usage-based scheme usually brings a higher traffic efficiency while the flat-rate scheme usually leads to a higher capacity utilization.

3.4 Traffic Cap Scheme

In Section 3.3 we have compared usage-based and flat-rate pricing schemes under time dependent pricing. In fact, these two schemes both have limitations. The usage-based scheme does not attract most users to access the wireless service, while the flat-rate scheme does not limit the usage of each user and this is why "bandwidth hogs" exist. In reality, many companies apply a "cap then metered" scheme, or "cap scheme" for short, which is a mixture of the above two schemes. To illustrate, AT&T charges $20 for 300MB and $30 for 3GB per month in "AT&T individual plan". Users enjoy a flat-rate pricing as long as their traffic consumption is no larger than this threshold, and a usage-based pricing is applied when the usage is beyond the threshold[4]. In this section, we explore the rationale of using the cap scheme under time dependent pricing, where the prices and the threshold can change over time.

The interplay between the ISP and consumers is still a Stackelberg game. Similar to the previous analysis, we start by analyzing the second stage game. Given the price g^t and traffic cap C^t during time slot $[t-1, t]$, users decide the amount of traffic to use by maximizing their utility function:

$$\begin{aligned} \max_{\mathbf{x}^t} \quad & U_c(\mathbf{x}^t) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - g^t \\ \text{s.t.} \quad & \sum_{i=1}^{I} x_i^t \le C^t, 0 \le x_i^t \le \theta_i^t, 1 \le i \le I. \end{aligned} \tag{3.16}$$

We have the following proposition to quantify its solution:

Proposition 3.4.1. *Given traffic cap C^t, there exists a λ^{t*} such that the optimal solution of the following optimization problem*

$$\begin{aligned} \max_{\boldsymbol{x}^t} \quad & \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^t/\theta_i^t) - \lambda^{t*} \sum_{i=1}^{I} x_i^t \\ \textit{s.t.} \quad & 0 \le x_i^t \le \theta_i^t, 1 \le i \le I, \end{aligned} \tag{3.17}$$

is a global optimum for Problem (3.16) if $f_i''(\cdot) < 0$ for any i.

[4]Usually, the users suppress their traffic consumption under the threshold due to the high per unit price when the usage is beyond the threshold.

Proof. Denote the optimal solution of Problem (3.16) as $\mathbf{x}^{t*} = (x_1^{t*}, x_2^{t*}, ..., x_I^{t*})$. The Lagrangian is:

$$\begin{aligned} &L(\mathbf{x}^{t*}, \lambda, \mathbf{v}, \mathbf{w}) = -U_{tc}(\mathbf{x}^{t*}) + \nu(\textstyle\sum_{i=1}^{I} x_i^{t*} - C^t) \\ &- \textstyle\sum_{i=1}^{I} v_i x_i^{t*} + \sum_{i=1}^{I} w_i(x_i^{t*} - \theta_i^t). \end{aligned} \tag{3.18}$$

The optimal solution to Problem (3.17) satisfies the KKT conditions if we assign $\nu = \lambda^{t*}$. We consider the Hessian matrix of the Lagrangian:

$$\triangledown^2 L(\mathbf{x}^{t*}) = -\mathbf{diag}\left(\frac{c_1}{\theta_1^t} f''\left(\frac{x_1^{t*}}{\theta_1^t}\right), \cdots, \frac{c_I}{\theta_I^t} f''\left(\frac{x_I^{t*}}{\theta_I^t}\right)\right). \tag{3.19}$$

When $f_i''(\cdot) < 0$ holds on for any i, we have that $\mathbf{y}^T L(\mathbf{x}^{t*})\mathbf{y} \geq 0$ for any $\mathbf{y} \neq 0$. Thus, the optimal solution to Problem (3.17) is the global optimum of Problem (3.16). □

According to Proposition 3.4.1, the optimal traffic consumption for Problem (3.16) can be obtained by solving Problem (3.17). To some extent, the cap scheme can be treated as the usage-based scheme with unit price λ^{t*}. Denote the optimal traffic consumption as $\mathbf{x}^{t*}$. The users' utility can be expressed as $U_c(\mathbf{x}^{t*}) = \sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^{t*}/\theta_i^t) - g^t$. The users access the network charged by traffic cap scheme if and only if $U_c(\mathbf{x}^{t*}) \geq 0$, and the fraction of these users is:

$$\mathbf{Pr}\{U_c(\mathbf{x}^{t*}) \geq 0\} = \int_{\sum_{i=1}^{I} c_i \theta_i^t f_i(x_i^{t*}/\theta_i^t) \geq g^t} d\Theta_t. \tag{3.20}$$

Now we analyze the first stage game. Knowing the best responses from consumers, the ISP maximizes its profit by charging a price g^t and setting a traffic cap C^t that solve:

$$\begin{aligned} \max_{\{g^t, C^t\}_t} \quad & \Pi_c = \sum_{t=1}^{T} g^t \mathbf{Pr}\{U_c(\mathbf{x}^{t*}) \geq 0\} \\ \text{s.t.} \quad & \int_{U_c(\mathbf{x}^{t*}) \geq 0} \sum_{i=1}^{I} x_i^{t*} d\Theta_t \leq \mu \quad \forall t. \end{aligned} \tag{3.21}$$

Given any C^t, due to similar reason with flat-rate scheme, there exists an optimal solution to the above problem and we denote it as $g^{t*}(C^t)$. So there exists an optimal solution for Problem (3.21), which we denote as $(g^{t*}(C^{t*}), C^{t*})$. Therefore, by the backward induction, we know that there exists a *Stackelberg equilibrium* using cap scheme and it is $(\mathbf{x}^{t*}, g^{t*}(C^{t*}), C^{t*})$.

In general, it is hard to quantify the properties of the Stackelberg equilibrium using the cap scheme. In order to show some interesting insights, we consider a special case where the traffic sensitivity $\beta_i = \beta(\beta \in [0,1])$ and the per unit valuation $c_i = c$. Define $\Phi^t = \max_s s\mathbf{Pr}\{\sum_{i=1}^I \theta_i^t \geq s\}$. In fact, $c\Phi^t$ is the maximal possible profit the ISP can obtain if $\mu = \infty$. We define the *cap benefit* of the ISP as the ratio of the ISP's optimal profit with traffic cap scheme over that with flat-rate scheme. Denote CB_p^t as the cap benefit of the ISP during time interval $[t-1, t]$. We have the following theorem.

Theorem 3.4.1. *If $\Phi^t > \mu$, then CB_p^t satisfies: 1) it is increasing in Φ^t and decreasing in β; and 2) $CB_p^t \geq (\frac{\Phi^t}{\mu})^{1-\beta}$.*

Proof. We substitute the variable g^t by $c_1 = \frac{g^t}{C^t}$. We can divide the original problem into two optimization problems by considering new conditions $c_1 \geq c$ and $c_1 < c$. We first consider the case $c_1 \geq c$. For any consumer with $\sum_{i=1}^I \theta_i^t < C^t$, $\lambda^{t*} = 0$ and $U_f(\mathbf{x}^{t*}) = c\sum_{i=1}^I \theta_i^t - g^t \leq cC^t - c_1C^t \leq 0$. It means that these consumers will not access the network. Thus, we only need to consider the users with $\sum_{i=1}^I \theta_i^t \geq C^t$. Under this case, $\lambda^{t*} = \beta c\left(\frac{C^t}{\sum_{i=1}^I \theta_i^t}\right)^{\beta-1}$ and $U_c(\mathbf{x}^{t*}) = c(C^t)^\beta(\sum_{i=1}^I \theta_i^t)^{1-\beta} - g^t$. The optimization problem becomes:

$$\begin{aligned} \max_{\{c_1, C^t\}} \quad & \pi_c^t = c_1 C^t \mathbf{Pr}\left\{\sum_{i=1}^I \theta_i^t \geq \left(\frac{c_1}{c}\right)^{\frac{1}{1-\beta}} C^t\right\} \\ \text{s.t.} \quad & C^t \mathbf{Pr}\left\{\sum_{i=1}^I \theta_i^t \geq (\frac{c_1}{c})^{\frac{1}{1-\beta}} C^t\right\} \leq \mu. \end{aligned} \tag{3.22}$$

Define $\Phi^t = \max_s s\mathbf{Pr}\{\sum_{i=1}^I \theta_i^t \geq s\} = \max_s s\int_{\sum \theta_i^t \geq s} d\Theta_t$ and denote s^* as one optimal solution. The maximum profit of the above optimization problem is:

$$\begin{aligned} \pi_c^{t*} &= c_1(\tfrac{c_1}{c})^{\frac{1}{\beta-1}}(\tfrac{c_1}{c})^{\frac{1}{1-\beta}} C^t \mathbf{Pr}\{\textstyle\sum_{i=1}^I \theta_i^t \geq (\frac{c_1}{c})^{\frac{1}{1-\beta}} C^t\} \\ &\leq c_1(\tfrac{c_1}{c})^{\frac{1}{\beta-1}} \Phi^t. \end{aligned} \tag{3.23}$$

Let $c_1 = c(\frac{\Phi^t}{\mu})^{1-\beta}$ and $C^t = \frac{\mu}{\Phi^t} s^*$. The above upper bound will be achievable and the constraint can also be satisfied. The maximal profit for the ISP will be $c(\Phi^t)^{1-\beta}\mu^\beta$. Then, we need to prove that $c_1 = c(\frac{\Phi^t}{\mu})^{1-\beta}$ and $C^t = \frac{\mu}{\Phi^t} s^*$ are the optimal solutions under both cases, i.e., $c_1 \geq c$ and $c_1 < c$. If not, the optimal profit under the case $c_1 < c$

will be higher than that under the case $c_1 \geq c$. Denote the optimal solution as (g^{t*}, C^{t*}). It means that $g^{t*} < cC^{t*}$. Under the case $c_1 < c$, we have the optimization problem

$$\max_{\{c_1, C^t\}} \pi_c^t = c_1 C^t \mathbf{Pr}\left\{\sum_{i=1}^{I} \theta_i^t \geq \frac{c_1}{c} C^t\right\} \tag{3.24}$$

with the capacity constraint

$$\int_{\frac{c_1}{c}C^t \leq \sum_{i=1}^{I} \theta_i^t \leq C^t} \sum_{i=1}^{I} \theta_i^t d\Theta_t + C^t \int_{\sum_{i=1}^{I} \theta_i^t \geq C^t} d\Theta_t \leq \mu. \tag{3.25}$$

Note that given $c_1^* = \frac{g^{t*}}{C^{t*}}$, for any $C^t < C^{t*}$, as C^t decreases, the total traffic will be non-increasing. We let $C^t = g^{t*}/c < C^{t*}$ and have

$$\begin{array}{l} \int_{\frac{c_1^*}{c}C^{t*} \leq \sum_{i=1}^{I} \theta_i^t \leq C^{t*}} \sum_{i=1}^{I} \theta_i^t d\Theta_t + C^{t*} \int_{\sum_{i=1}^{I} \theta_i^t \geq C^{t*}} d\Theta_t \\ \geq \frac{c_1^*}{c} C^{t*} \int_{\sum_{i=1}^{I} \theta_i^t \geq \frac{c_1^*}{c} C^{t*}} d\Theta_t = 1/c\pi_c^{t*}. \end{array} \tag{3.26}$$

Since $\pi_c^{t*} \geq c(\Phi^t)^{1-\beta}\mu^\beta$, we have $\mu \geq (\Phi^t)^{1-\beta}\mu^\beta$. Then, we have $\mu \geq \Phi^t$ that contradicts to the condition that $\mu < \Phi^t$. Thus, the $(g^{t*}, C^{t*}) = \left(c(\frac{\mu}{\Phi^t})^\beta s^*, \frac{\mu}{\Phi^t} s^*\right)$ is the optimal solution.

For flat-rate scheme, the maximal profit of the ISP will be no more than $c\mu$. Then, we have $CB_p^t \geq \frac{c(\Phi^t)^{1-\beta}\mu^\beta}{\pi_b} \geq \frac{c(\Phi^t)^{1-\beta}\mu^\beta}{c\mu} = (\frac{\Phi^t}{\mu})^{1-\beta}$. It is clear that π_f^{t*} is independent with β and Φ^t. This completes the proof. □

Theorem 3.4.1 indicates that the ISP's cap benefit is always larger than one when $\Phi^t > \mu$, and it increases with respect to Φ^t and decreases with respect to μ. This means when the capacity is insufficient, the cap benefit becomes more dominant. This is because the cap scheme reduces high volume of traffic consumption. We also note that small β means high cap benefit. This is because low β indicates that consumers conserve high unit valuation of customers under small cap threshold, and these customers accept high price charged by the ISP, increasing the ISP's profit.

We also analyze the traffic cap scheme from the consumers' point of view. Similarly, we can define the cap benefit of consumers' surplus, and we denote its value in $[t-1, t]$ as CB_s^t. We have the following theorem.

Theorem 3.4.2. *If $\Phi^t > \mu$, then CB_s^t decreases when β increases, and $CB_s^t \to 0$ when $\beta \to 1$.*

Proof. The consumers' surplus for traffic cap scheme during time interval $[t-1, t]$, denoted as ψ_c^t, is $\psi_c^t = c\left(\frac{\mu}{\Phi^t}\right)^{\beta} s^* \int_{\sum_{i=1}^I \theta_i^t \geq s^*} \left[\left(\frac{\sum_{i=1}^I \theta_i^t}{s^*}\right)^{1-\beta} - 1\right] d\Theta_t$. Since $\frac{\mu}{\Phi^t} < 1$ and $\frac{\sum_{i=1}^I \theta_i^t}{s^*} \geq 1$, we know that ψ_c^t decreases when β increases. When $\beta \to 1$, we get $\psi_c^t \to 0$. It is clear the consumer's surplus for flat-rate scheme is independent of β. This completes the proof. □

Theorem 3.4.2 shows that the traffic cap strategy cannot always improve the consumers' surplus. When β is small, the consumers' surplus is high using cap scheme, while under the flat-rate scheme it is independent of β. When β increases, consumers' surplus reduces; and when $\beta \to 1$, the consumers' surplus approaches zero under the traffic cap scheme.

We can similarly define cap benefit of traffic efficiency and denote its value in $[t-1, t]$ as CB_e^t. We have:

Theorem 3.4.3. *If $\Phi^t > \mu$, then CB_e^t satisfies: 1) it is decreasing in β and μ; 2) $CB_e^t \geq (\frac{\Phi^t}{\mu})^{1-\beta}$; and 3) $CB_e^t \to 1$ as $\beta \to 1$.*

Proof. The traffic efficiency of the traffic cap scheme during $[t-1, t]$, denoted as ϕ_c^t, is

$$\phi_c^t = c(\Phi^t)^{-\beta} \mu^{\beta-1} s^* \int_{\sum_{i=1}^I \theta_i^t \geq s^*} \left[\left(\frac{\sum_{i=1}^I \theta_i^t}{s^*}\right)^{1-\beta}\right] d\Theta_t. \tag{3.27}$$

The traffic efficiency of the flat-rate scheme is c. Then, we have

$$\begin{aligned} CB_e^t &= (\Phi^t)^{-\beta} \mu^{\beta-1} s^* \int_{\sum_{i=1}^I \theta_i^t \geq s^*} \left[\left(\frac{\sum_{i=1}^I \theta_i^t}{s^*}\right)^{1-\beta}\right] d\Theta_t \\ &\geq (\Phi^t)^{-\beta} \mu^{\beta-1} s^* \int_{\sum_{i=1}^I \theta_i^t \geq s^*} d\Theta_t = (\tfrac{\Phi^t}{\mu})^{1-\beta}. \end{aligned} \tag{3.28}$$

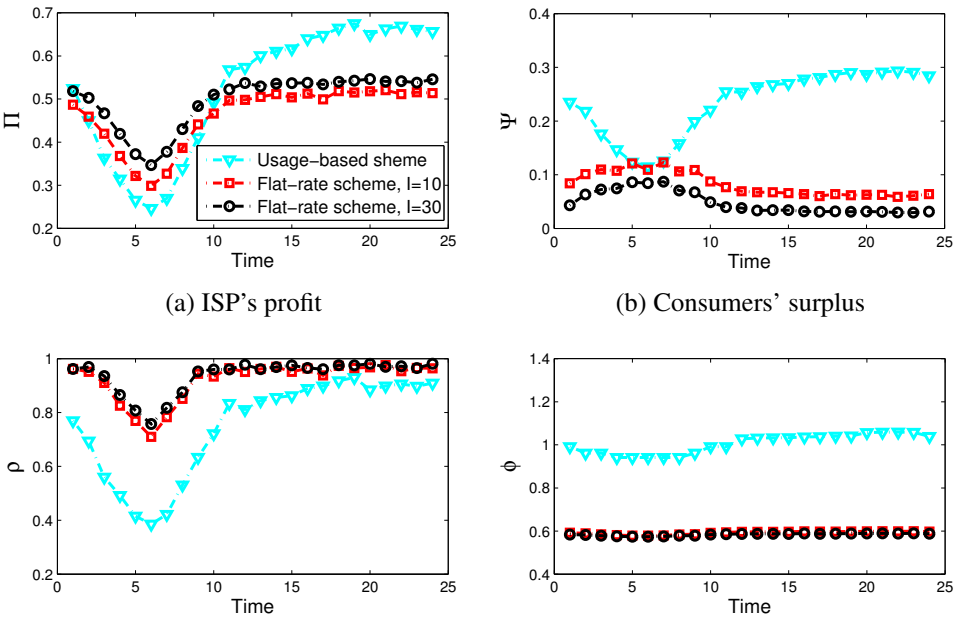
Since $\frac{\mu}{\Phi^t} < 1$ and $\frac{\sum_{i=1}^I \theta_i^t}{s^*} \geq 1$, we have CB_e^t is a decreasing function with respect to β and μ. When $\beta \to 1$, we have $CB_e^t \to 1$. □

Theorem 3.4.3 shows that the efficiency can increase by adopting traffic cap strategy when $\Phi^t > \mu$. When β is small, the benefit is large because the consumers consume the data in a more efficient way. The traffic efficiency is high when the capacity is less than Φ^t (or insufficient). The traffic cap strategy improves the traffic efficiency by replacing low-valuation traffic with high-valuation traffic. For example, when the capacity is insufficient, a user may use it to read emails but not watching video because the per-unit valuation of reading email is much higher. This also means that the traffic cap strategy can improve traffic efficiency while keeping high capacity utilization. When the capacity is sufficient, the traffic cap strategy will just work like a flat-rate scheme.

Remarks: The cap strategy combines the advantages of usage-based and flat-rate schemes. When the capacity is sufficient, the cap strategy improves the capacity utilization, which is similar to the effect of flat-rate scheme. When the capacity is insufficient, the cap strategy improves traffic efficiency, which is similar to the effect of usage-based scheme. Therefore, the ISP has a strong incentive to introduce this cap into its pricing strategy. However, consumer's surplus may not always be as large as that under the flat-rate scheme.

3.5 Numerical results

In this section, we provide numerical examples for quantitative study on the key features of the three schemes discussed above. We set the satisfaction function in the form of Eq. (3.3). The default number of services is set as 10. The per unit valuation for service i is randomly chosen from $[0, 1]$. The distributions of the maximal demand during peak time are assumed to be uniform distributions $U([0, \alpha_i])$, where α_i is randomly selected from $[0, 10]$. The traffic sensitivity β_i is randomly selected from $[0, 1]$ if not specified otherwise. We divide a day into 24 time slots as [29]. The maximal demands during different slots are obtained by multiplying a discount function in terms of time from a 24-hours traffic usage data [1] normalized in $[0, 1]$. To satisfy the maximal demand for all time slots, the capacity per service needs to be around 2.5. In practice, the capacity



(a) ISP's profit

(b) Consumers' surplus

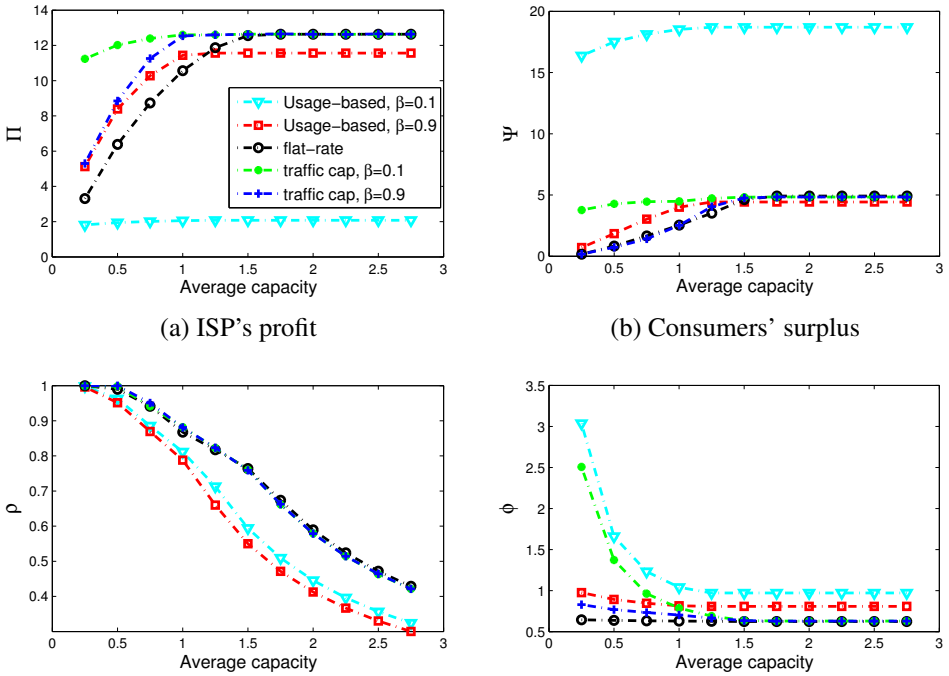
(c) Capacity utilization

(d) Traffic efficiency

Figure 3.2: Usage-based scheme vs. flat-rate scheme

is always insufficient during peak time and sufficient during valley time in wireless data networks, so we set the capacity per service as 1 by default. We consider three schemes for time dependent pricing: usage-based scheme, flat-rate scheme and traffic cap scheme. The performance measures include the ISP's average profit per service Π, consumers' average surplus per service Ψ, capacity utilization ρ and traffic efficiency ϕ.

We first compare the usage-based and flat-rate schemes. Fig. 3.2(a) shows the ISP's average profit per service during different time slots. In valley time, e.g., 5 am, the flat-rate scheme leads to a higher profit than usage-based scheme. The main reason is that the flat-rate scheme attracts more traffic usage (which is verified in Fig. 3.2(c)). In peak time, e.g., 10 pm, the ISP benefits more from usage-based scheme. This is because the usage-based scheme improves the traffic efficiency during peak time (which is verified in Fig. 3.2(d)). The traffic efficiency for usage-based scheme is almost twice as that of flat-rate scheme. We also compare the flat-rate scheme when the numbers of services changes. As the number increases, the ISP obtains a higher profit. The reason is a large number of services means low heterogeneity of the valuation in all services. More users can be attracted by a single price so that the capacity utilization is high (which is



(a) ISP's profit

(b) Consumers' surplus

(c) Capacity utilization

(d) Traffic efficiency

Figure 3.3: Comparison of three schemes under different capacities

verified by Fig. 3.2(b)). Yet, the consumers' surplus reduces when the heterogeneity of the valuation decreases, as is shown in Fig. 3.2(b).

We then compare the cap and flat-rate schemes with various capacities. Fig. 3.3(a) demonstrates that the ISP always benefits more from traffic cap scheme. A smaller capacity means a larger profit of the ISP using the traffic cap scheme. In addition, a lower traffic sensitivity indicates a higher profit of the ISP. For instance, when the traffic sensitivity is large, e.g., $\beta = 0.9$, and the average capacity per service is small, e.g., $\mu = 0.25$, the profit of the ISP for traffic cap scheme is around 1.5 times of that in flat-rate scheme. When the traffic sensitivity is small, e.g., $\beta = 0.1$, the benefit is more than 3 times than that in flat-rate scheme. Fig. 3.3(c) and Fig. 3.3(d) show that the capacity utilizations for cap and flat-rate schemes are almost the same; while the traffic efficiency for traffic cap scheme is much higher, especially when the capacity is small. It shows the traffic cap scheme does not increase capacity utilization but does improve traffic efficiency. Fig. 3.3(b) shows that consumers benefit from traffic cap scheme when the traffic sensitivity is small. When the traffic sensitivity is large, the consumers' surplus

may reduce.

We also compare the cap and usage-based schemes under different capacities. Fig. 3.3(a) and Fig. 3.3(b) show that the ISP strongly prefers traffic cap strategy while consumers' surplus is usually much higher when using the usage-based scheme. The main reason is that cap scheme always has the advantage of reducing the heterogeneity of the consumers' valuation. This enables the ISP to earn profit from consumers and reduce consumers' surplus. Fig. 3.3(c) shows that the usage-based scheme always has a low capacity utilization, and a smaller traffic sensitivity means lower capacity utilization. Fig. 3.3(d) shows that both traffic cap and usage-based schemes have high traffic efficiency.

3.6 Summary

In this chapter, we explore the design space of practical and effective schemes for time dependent pricing in a monopoly ISP market. We model the users valuation for different services in a wireless data network. We use game theoretic analysis to capture the interplay between consumers and the ISP. Based on this, we compare three schemes, i.e., usagebased scheme, flat-rate scheme and cap scheme, in terms of the ISPs profit, users surplus, capacity utilization and traffic efficiency, respectively. Our important findings includes: 1) the monopoly ISP obtains a higher profit using usage-based (or flat-rate) scheme if the capacity is insufficient (or sufficient); 2) the usage-based scheme usually achieves a higher consumer surplus and better traffic efficiency than flat-rate scheme; and 3) the ISP prefers using the cap scheme to further increase its revenue, but consumers may not benefit under the cap scheme. We believe our findings provide important insights for ISPs to design effective pricing schemes.

CHAPTER 4

SPONSORED DATA PLAN: A TWO-CLASS SERVICE MODEL IN WIRELESS DATA NETWORKS

4.1 Overview

With the popularity of bandwidth-intensive mobile devices like smart phones and tablet computers, data traffic is increasing fast recently. The trend of intensive interactions of mobile devices and public clouds suggests that the amount of future wireless data traffic can be even daunting. This poses huge burden to the Internet service providers (ISPs) since supporting such demand-supply gap requires large investments. To share such costs with users, flat-rate pricing plans used in broadband networks are phasing out. ISPs now propose pricing plans with a cap. Users are either not allowed or highly charged for consuming traffic volume beyond this cap. The cap is usually conservative; for example, Google revealed that almost 85% of the plans offer less than 10 GB data per month, and 36% offer less than 1 GB per month [28]. Such data caps would be easily reached in less than seven hours under the current 3G bandwidth [65]. There is thus a great demand on better pricing models. One research direction is time dependent pricing [30, 84]. The key observation is that the user traffic demands are not uniform at different times. Therefore, higher (lower) prices can be applied to peak (off-peak) hours.

Another recent proposal is *sponsored data plan*, originated from 1-800 services of phone calls [8]. In particular, ISPs provide platforms for the content providers (CPs) to sponsor their end users, but with some payments or indirectly sharing advertising revenue, such that when end users access the content from one CP joining the sponsored data plan, their traffic from this CP is partially or fully exempted from their data caps. For example, Google has joined with India's Bharti Airtel to offer free access to certain Google-based services such as Gmail, Google+ and first page of web sites via Google search without

ringing up data charges [27]. It is hoped that this pricing strategy can create a positive cycle: End users are glad to access more contents which will not be counted into their data caps; Content providers can attract more users and views; ISPs can obtain higher revenue to support better quality of service (QoS) and carry out technology upgrade. Early works [4, 83] have confirmed its benefits to CPs and end users.

Nevertheless, a key problem is whether such a plan may lead to unfair competition advantage to certain parties. Similar debates, e.g., network neutrality, appeared in the past. Opponents, including network neutrality advocators, representatives from public interest groups, concern that such a plan will favor rich and big CPs over small ones [55]. This may impede Internet innovation and ultimately hurt consumers. Proponents, mostly ISPs and some CPs, argue that this plan can promote competition and improve efficiency [5]; ultimately, consumers will benefit from better services and cheaper traffic. The pioneer ISP, AT&T, expressed its confidence that the sponsored data plan complies with Federal Communications Commission (FCC) network neutrality rules. The core of network neutrality is that packet flows at the ISP level should not bear priorities[1]. This is because when congestion occurs, low priority flows can be dropped before high priority flows. Sponsored content, however, does not trigger differentiated services at the ISP flow level. When congestion occurs, packet flows of sponsored content have the same probability to get dropped as non-sponsored content. The FCC is inclined to side with proponents though its chairman claimed that the commission will carefully watch the sponsored data program and intervene if it finds the sponsored content practice violates the Open Internet Order [24]. After a long time of planning, AT&T finally announced its sponsored data program in January 2014. Its sponsored data partner, Syntonic Wireless, launched "toll-free" content store six months later [26].

In this chapter, we show that an unfair competition advantage may still exist. In particular, there may be a competition disadvantage for certain CPs, which ultimately hurts users, such that a tiered service will be created. Under the plan, there will be a sponsored

[1]Priorities for traffic engineering might be acceptable; yet priorities targeting on certain application types (e.g., P2P) or some particular CPs (e.g., Google), should be prohibited.

class services and an ordinary class services. The sponsored class brings higher revenue for CPs than the ordinary class. Yet to join this class, a CP needs to pay a non-trivial premium for each unit of content it delivers. Only a CP whose unit profit is high can afford such a premium. This discriminates CPs in the sense that they need to compete for capital, i.e., the capability they pay a premium for their content, rather than their services, i.e., the quality of the contents or the quality of services. In addition, we find that in the long run, the ISP has no incentives to enlarge its traffic cap when its traffic capacity is sufficient; this greatly hurts consumers and CPs. Although these results seem discouraging for the sponsored data plan, we also show that if the traffic cap is regulated properly so as to guarantee the majority part of capacity being allocated to the ordinary class, the ISP's optimal strategy is aligned with the consumers' surplus.

In this chapter, we study a set of CPs, a monopolistic ISP and a set of users[2]. We model the users' traffic demand dynamics under the sponsoring plan with QoS consideration (Sec. 4.2.1 and 4.2.2). Based on this, we develop a Stackelberg game framework (Sec. 4.2.3 and 4.2.4) and analyze the interactions between the ISP and CPs (Sec. 4.3 and 4.4). In particular, we find that the equilibrium may not always exist and discuss the outcome of the interactions. We also develop efficient polynomial time algorithms to search the equilibrium in exponential solution space. Finally, we find the best strategy of the ISP and CPs, i.e., the solution to the Stackelberg game. Our major findings are:

- If the ISP's capacity is sufficient, then the sponsored data plan benefits consumers and CPs in the short run, but in the long run, the ISP has no incentives to enlarge its traffic cap, which hurt consumers and CPs.
- If the ISP's capacity is insufficient, then the ISP has strong incentives to enlarge its capacity, which benefits both CPs and consumers. However, the ISP's optimal strategy to enlarge its profit is always contrary to CPs' surplus. If the traffic cap is regulated properly, this strategy is aligned with the consumers' surplus.

[2]In this thesis, we use the terms "users" and "consumers" interchangeably.

- The sponsored data plan may enlarge the unbalance in revenue distribution between different CPs; those with higher unit income and poorer technology support are more likely to prefer the sponsored data plan.

4.2 General Model

In this section, we model the market with three parties: a set of CPs $\mathcal{N}$ ($N = |\mathcal{N}|$), a monopolistic ISP and a set of end users with a total number M. The CPs provide services to end users. We assume that one CP supplies only one service. If a CP provides multiple services, then we treat it as multiple virtual CPs, each serving one particular service. The ISP provides Internet access services to CPs and end users. Usually, there is a transmission bottleneck for the connection services between CPs and end users. We define the traffic capacity (or capacity for short) of the ISP, denoted by μ, as the maximal possible amount of traffic volume that can be transmitted through the bottleneck during a fixed period[3]. Based on the above model, we can use a triple $(\mathcal{N}, \mu, M)$ to represent the whole system.

4.2.1 Consumers' Traffic Demand

Users have different preferences towards various contents and services of CPs. We use *valuation* to present this preference. Facing different choices in the service set, a user prefers accessing a service with a high per unit valuation when he has a usage limitation (or traffic cap). We assume a user's per-unit traffic valuation has a decreasing trend. For example, a user may have a high valuation on a VoIP service since he needs to have an important discussion with his friend, but when he finishes this discussion, the extra traffic he consumes, e.g., for telling jokes, is with a low valuation. In other words, the marginal valuation decreases with the traffic volume consumed. We define a strictly decreasing *valuation density function* $g_i(\cdot)$ to capture this feature, and the total valuation of consuming x_i amount of traffic is given by $\int_0^{x_i} g_i(s)ds$. Further, we assume

[3]Please be noted that the definition of "*capacity*" is different from the ISP's *cap* announced to the users; the later concept refers to the maximal traffic that a user is allowed consume during a fixed period.

$\int_0^\infty g_i(s)ds < \infty$.

Each service requires a certain bandwidth to achieve good QoS. For instance, a bandwidth of 500 Kbps is required for YouTube videos. We denote this maximal requirement for the service of CP i (or service i for short) as $\hat{b}_i$. However, in reality it may not be totally satisfied due to ISP's insufficient capacity. We denote the achievable bandwidth as b_i. Obviously, we have $b_i \leq \hat{b}_i$. When $\hat{b}_i$ cannot be satisfied, QoS decreases, resulting phenomenons like frequent screen freeze in video display. We define $q_i = b_i/\hat{b}_i$ as the ratio of achievable bandwidth over the maximal bandwidth requirement. It reflects the extent of QoS degradation; when $q_i < 1$, it may lead to a reduction of users' valuations. In later parts of this chapter, we call q_i "*QoS index*". We capture this effect by a *QoS satisfaction function*: $h_i(\cdot) : [0,1] \to [0,1]$, where $h_i(0) = 0$ and $h_i(1) = 1$. We assume that it is a non-decreasing and continuous function in q_i. When $q_i < 1$, the marginal valuation decreases to $g_i(\cdot)h_i(q_i)$, and thus the total valuation for consuming x_i amount of traffic becomes $\int_0^{x_i} g_i(s)h_i(q_i)ds$.

We assume that an end user accesses a service if and only if his per unit traffic valuation of this service is higher than a pre-set threshold denoted by t_i. This threshold may come from the cost of bearing irritating pop-ups. Due to the decreasing marginal valuation of a service, there exists a *usage threshold* for a user where his marginal valuation is equal to t_i. We define the *usage threshold* for service i as

$$\theta_i = \max\{s : g_i(s)h_i(q_i) \geq t_i\},$$

which reflects the maximal possible traffic usage for service i. When the inverse function of $g_i(\cdot)$ exists, denoted as $g_i^{-1}(\cdot)$, we have $\theta_i = g_i^{-1}\left(\frac{t_i}{h_i(q_i)}\right)$. For any traffic usage $x_i \in [0, \theta_i]$, let us define the users' *utility* as[4]

$$\psi_i(x_i) = \int_0^{x_i} [g_i(s)h_i(q_i) - t_i]\, ds. \tag{4.1}$$

[4]We do not include the Internet access fee charged by the ISP into the formula, since the access fee is a constant and does not impact any result.

We assume that the utility of consuming different services are additive. Therefore, the utility of accessing all services with traffic usage $\mathbf{x} = (x_1, \cdots, x_N)$ is

$$\psi(\mathbf{x}) = \sum_{i \in \mathcal{N}} \int_0^{x_i} [g_i(s) h_i(q_i) - t_i] \, ds. \tag{4.2}$$

In later analysis, we also call it the *surplus* of a consumer.

Now let us observe the ISP's role in users' consumption decisions. An ISP usually applies a "flat-rate-like" pricing scheme but with a cap, i.e., end users can use the traffic below the cap like by paying a flat rate, but they are not allowed to consume traffic volume beyond this cap, or are charged by a much higher price for usage beyond[5]. Denote the cap as C. In this chapter, we assume that each user's usage is below this cap. This assumption is for mathematical tractability; in reality, users do usually limit their usage below this cap due to the high fee charged for beyond. Under the sponsored data plan, traffic consumption of a particular service can be partially or totally exempted from this cap. We denote $\mathcal{O}$ (or $\mathcal{S}$) as the set of ordinary (or sponsoring) content providers (or services). Content providers in $\mathcal{S}$ sponsor the total traffic volume consumed on their content, while those in $\mathcal{O}$ do not participate in the sponsored plan. Each CP in $\mathcal{N}$ is in either $\mathcal{S}$ or $\mathcal{O}$. Thus, given the QoS index vector $\mathbf{q} = (q_1, \cdots, q_N)$, an end user can decide his optimal traffic usage by maximizing his utility:

$$\begin{aligned} \max_{\mathbf{x}} \quad & \psi(\mathbf{x}) = \sum_{i \in \mathcal{N}} \int_0^{x_i} [g_i(s) h_i(q_i) - t_i] \, ds, \\ \text{s.t.} \quad & \sum_{i \in \mathcal{O}} x_i \le C, \ 0 \le x_i \le \theta_i. \end{aligned} \tag{4.3}$$

Note that for end users, the total traffic usage for non-sponsored services should not exceed the cap. Given the choices of CPs, i.e., $(\mathcal{O}, \mathcal{S})$, the above optimization can be solved by KKT conditions [12]. Therefore, we have

Lemma 4.2.1. *A user's optimal usage of the content provided by CP i, denoted as x_i, is:*

$$x_i = \begin{cases} \max\left\{0, g_i^{-1}\left(\frac{t_i + \nu}{h_i(q_i)}\right)\right\} & i \in \mathcal{O}, \\ \theta_i & i \in \mathcal{S}, \end{cases} \tag{4.4}$$

[5]For instance, in the "AT&T individual plan" for 4G smart phones, AT&T charges \$10 for a traffic cap 1 GB per month, but another \$5 for any additional 50 MB.

where ν is the Lagrange multiplier associated with the cap constraint. In addition, ν is non-decreasing with respect to q_i, and non-increasing with respect to C.

Proof. We introduce the Lagrange multiplier ν for the cap constraint, w_i for constraint $x_i \geq 0$ and v_i for constraint $x_i \leq \theta_i$. Then, we have the KKT conditions:

$$\nu \geq 0, w_i \geq 0, v_i \geq 0, 0 \leq x_i^* \leq \theta_i, \forall i \in \mathcal{N} \tag{4.5}$$

$$\nu(\sum_{i\in\mathcal{O}} x_i^* - C) = 0, w_i x_i^* = 0, v_i(x_i^* - \theta_i) = 0, \forall i \in \mathcal{N} \tag{4.6}$$

$$-[g_i(x_i^*)h_i(q_i) - t_i] + \nu - w_i + v_i = 0, \forall i \in \mathcal{O} \tag{4.7}$$

$$-[g_i(x_i^*)h_i(q_i) - t_i] - w_i + v_i = 0, \forall i \in \mathcal{S} \tag{4.8}$$

For any $i \in \mathcal{S}$, if $v_i > 0$, then $x_i^* = \theta_i$. Then, $w_i = v_i > 0$ and thus $x_i^* = 0$. This is impossible. Thus, for any $i \in \mathcal{S}$, $v_i = 0$. Note that $w_i \geq 0$ and $g_i(x_i^*)h_i(q_i) - t_i \geq 0$ when $0 \leq x_i^* \leq \theta_i$. Then, we have $w_i = 0$ and $x_i^* = \theta_i$. Then, we consider any $i \in \mathcal{O}$. Note that $\theta_i = g_i^{-1}\left(\frac{t_i}{h_i(q_i)}\right)$. If $v_i > 0$, then $x_i^* = \theta_i$, and thus $w_i = 0$ and $g_i(x_i^*)h_i(q_i) - t_i = 0$. That means $\nu + v_i = 0$. Since $\nu \geq 0$ and $v_i > 0$, this is impossible. Thus, we have $v_i = 0$. If $x_i^* > 0$, then $w_i = 0$ and $x_i^* = g_i^{-1}\left(\frac{t_i+\nu}{h_i(q_i)}\right)$. Note that $x_i^* \geq 0$. Thus, $x_i^* = \max\left\{0, g_i^{-1}\left(\frac{t_i+\nu}{h_i(q_i)}\right)\right\}$. Note that $x_i^*(\nu)$ is non-increasing with respect to ν and $\sum_{i\in\mathcal{O}} x_i^*(\nu) = C$. Thus, ν is non-increasing with respect to C. Also note that x_i^* is non-decreasing with respect to q_i. Thus, ν is non-decreasing with respect to q_i. Hence, we complete the proof. □

Lemma 4.2.1 derives the optimal traffic usage of customers. The traffic usage for services in $\mathcal{S}$ always approaches the usage threshold, but that in $\mathcal{O}$ is constrained by the traffic cap and the usage threshold.

We define ν as the *level of competition* for the traffic cap, because a high ν indicates a low traffic cap, so CPs in $\mathcal{O}$ face intense competition to attract users' consumption within this limited cap. When C is large, the traffic usage of services in $\mathcal{O}$ also approaches the usage threshold. Services in $\mathcal{S}$ has no impact on the demand of traffic in $\mathcal{O}$. Lemma 4.2.1

also states that the level of competition is also affected by the QoS index. A higher QoS index means a higher level of competition since it increases the traffic demand.

Discussion on QoS Satisfaction Functions: Users may have different requirements on QoS for different services. For example, for real-time applications like Netflix, the value of the QoS satisfaction function decreases dramatically with respect to q_i. This is because inadequate bandwidth for the realtime applications greatly hurts users' experience. In contrast, for delay-tolerant services like email, reduction in QoS does not hurt users' experience too much. Therefore, in this chapter, we define the QoS satisfaction function in the following form:

$$h_i(q_i) = {q_i}^{\gamma_i}, \tag{4.9}$$

where γ_i is called the *quality sensitivity* for service i. A large γ_i represents a service with a high sensitivity on the quality, while a small γ_i represents one with a low sensitivity.

Discussion on Valuation Density Functions: Define $\alpha_i = \lim_{x_i \to 0} g_i(x_i)$. The value of the valuation density function approaches the maximum α_i when the user's consumption on service i approaches zero. As the traffic amount increases to infinity, the marginal valuation decreases to zero, i.e., $\lim_{x_i \to \infty} g_i(x_i) = 0$. The above requirement is needed to guarantee $\int_0^\infty g_i(s)ds < \infty$. In particular, we consider the following canonical form of valuation density function:

$$g_i(x_i) = \alpha_i e^{-\beta_i x_i}, \tag{4.10}$$

where β_i captures the *traffic sensitivity* on the valuation of service i. A higher traffic sensitivity indicates that the valuation of per unit traffic decreases more rapidly when more traffic is consumed by end users.

Illustration: We consider an example where two CPs are with parameters $(\alpha_1, \beta_1, \gamma_1) = (3, 2, 0.5)$ and $(\alpha_2, \beta_2, \gamma_2) = (1.5, 0.5, 1.5)$; CP 1 provides an Email type service with a high per unit valuation, a high traffic sensitivity and a low quality sensitivity, while CP 2 provides a video type service with a low per unit valuation, a low traffic sensitivity and a high quality sensitivity. Let the pre-set threshold be $t_i = 1$.

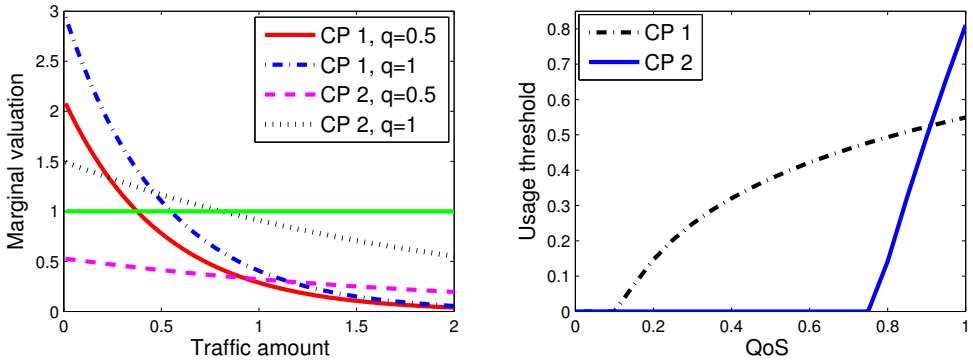


Figure 4.1: An example of consumers' valuation model

Figure 4.1 shows the marginal valuation with respect to the traffic consumed by both CPs (the left subfigure) and the usage threshold of the two CPs with respect to the QoS index (the right subfigure). The marginal valuation decreases with respect to the traffic amount consumed, and the decreasing trend becomes more rapid when congestion happens. For example, the marginal valuation of CP 1 decreases 63% when the traffic amount increases from 0.5 to 1. When the QoS index decreases, e.g., from 1 to 0.5, the marginal valuation of CP 1 decreases 28% further. A user consumes the traffic only when the marginal valuation is higher than the pre-set threshold. The critical point, i.e., the usage threshold, shifts to the left when congestion happens. When the marginal valuation line is all below the pre-set threshold line, the end user consumes no traffic and does not receive such service. The relationship between the usage threshold and the QoS index are shown in the right sub-figure. Each CP needs some QoS guarantee to attract its end users, while different CPs have different requirements. For example, CP 1 requires that the QoS index is larger than 0.1, while CP 2 requires it larger than 0.75. Comparing to CP 1, quality degradation has a more serious effect on CP 2. Decrease of the QoS index from 1 to 0.9 results in 39% reduction of the usage threshold for CP 2 but only 4% for CP 1.

4.2.2 Capacity Sufficiency and Rate Allocation Mechanism

In this subsection we analyze the aggregated traffic demand of all services, based on which we define the sufficiency and insufficiency of the ISP's capacity. Later, we capture

the interactions between the traffic demand, sufficiency of capacity, and the quality of service.

Let us first consider the aggregated traffic demand of M end users receiving services in the sponsored class $\mathcal{S}$, i.e., $\sum_{i\in\mathcal{S}} Mx_i(\boldsymbol{q})$. According to lemma 4.2.1, the optimal usage for each service approaches the usage threshold, i.e., $x_i = \theta_i(q_i)$ for any $i \in \mathcal{S}$. The aggregated traffic demand for services in $\mathcal{S}$ is $\sum_{i\in\mathcal{S}} M\theta_i(q_i)$. Then let us consider the aggregated traffic demand in $\mathcal{O}$, i.e., $\sum_{i\in\mathcal{O}} Mx_i(\boldsymbol{q})$. When the traffic cap is sufficiently large, each end user's optimal traffic usage is the usage threshold, so the aggregated traffic demand in $\mathcal{O}$ is $\sum_{i\in\mathcal{O}} M\theta_i(q_i)$. Otherwise, if the cap for the users is smaller than $\sum_{i\in\mathcal{O}} M\theta_i(q_i)$, then the aggregated traffic demand is MC where C is the cap set by the ISP. Therefore, the aggregated demand for services in $\mathcal{O}$ is $M\min\left\{\sum_{i\in\mathcal{O}}\theta_i(q_i), C\right\}$. The aggregated traffic demand from all users is

$$D(\mathbf{q}) = \sum_{i\in\mathcal{S}} M\theta_i(q_i) + M\min\left\{\sum_{i\in\mathcal{O}}\theta_i(q_i), C\right\}. \tag{4.11}$$

Now let us formally define the sufficiency (insufficiency) of the ISP's capacity.

Definition 4.2.1. *We say that the ISP's capacity μ is sufficient if $\mu \geq D(\boldsymbol{1})$, or insufficient otherwise.*

When μ is insufficient, we can by no means guarantee that each customer receives each service under the best QoS; in other words, congestion happens. Lots of bandwidth allocating mechanisms have been used to address the rate allocation problem under congestion [54, 56]. One well adopted method is the proportional share mechanism [54], where each flow reduces the same percentage of rates under congestion. In other words, the ratios of the achievable bandwidth over the maximal bandwidth requirement for any two services i and j are the same, i.e., $b_i : \hat{b}_i = b_j : \hat{b}_j$, so the QoS indices for each service are the same, i.e., $q_i = q$ for any $i \in \mathcal{N}$.[6] Thus, the traffic demand function $D(\mathbf{q})$

[6]In practice, the QoS index q may be dynamic over time due to volatile traffic demand (e.g., high QoS index during valley period and low QoS index during peak period). However, the QoS index during peak period is relatively stable [21]. In this chapter, we focus on the traffic capacity during peak period and treat q as the average QoS index over the period.

can be simplified by $D(q)$. When $q < 1$, users' demand for each service reduces. When reaching a steady state, the traffic demand is equal to the capacity of the ISP:

$$D(q) = \mu. \tag{4.12}$$

We call the QoS index q that satisfies the above equation an *equilibrium QoS*. Let us define $\lambda = \mu/M$ as the average capacity (or per user capacity). Given the sets $(\mathcal{O}, \mathcal{S})$, the equilibrium QoS is captured by the following lemma.

Lemma 4.2.2. *Given the sets $(\mathcal{O}, \mathcal{S})$, there is a unique equilibrium QoS $q \in [0, 1]$. Further, it is a non-decreasing function with respect to λ, and a non-increasing function with respect to C.*

Proof. Since $g_i(\cdot)$ is a decreasing function, its inverse function $g_i^{-1}(\cdot)$ is also a decreasing function. Since $h_i(q_i)$ increases with q_i, $\theta_i(q) = g_i^{-1}(\frac{t_i}{h_i(q)})$ increases with q. Denote left side of the equation as $Q_C(q) = \min\left\{\sum_{j\in\mathcal{O}}\theta_j(q), C\right\} + \sum_{j\in\mathcal{S}}\theta_j(q)$. $Q_C(q)$ increases with q during interval $[0, 1]$. When $Q_C(1) \leq \mu$, $q^* = 1$. Otherwise, these exists one unique $q^* \in [0, 1]$ such that $Q_C(q^*) = \mu/M = \lambda$. Consider traffic cap C_1 and traffic cap C_2 satisfying $C_1 \leq C_2$. Then, $Q_{C_1}(q) \leq Q_{C_2}(q)$. Denote the equilibrium QoS under two traffic caps as q_1^* and q_2^*. Then, we have $Q_{C_1}(q_1^*) = Q_{C_2}(q_2^*) \geq Q_{C_1}(q_2^*)$. Since $Q_C(q)$ increases with q, we have $q_1^* \geq q_2^*$. Thus, the equilibrium QoS is a non-increasing function with respect to C. Considering two per user capacity λ_1, λ_2 and $\lambda_1 \geq \lambda_2$, then $Q_C(q_1') \geq Q_C(q_2')$, where q_1' and q_2' are the equilibrium QoS under per user capacity λ_1 and λ_2 respectively. Since $Q_C(q)$ increases with q, we have $q_1' \geq q_2'$. Thus, the equilibrium QoS is a non-decreasing function with respect to λ. Hence, we complete the proof. □

Lemma 4.2.2 shows that the ISP can improve QoS by enlarging its capacity, in particular, when $q < 1$. It is also interesting to note that if we merely increase the ISP's traffic cap, i.e., users are allowed to consume more traffic, then QoS becomes worse since the traffic demand from users increases but the ISP's capacity remains the same.

4.2.3 Utility of CPs and the ISP

Now let us formally define the utility functions of content providers and the Internet service provider. This serves as the foundation for our further game analysis.

Utility of CPs. We use v_i to denote the per unit revenue of CP i. Content providers may have quite different per unit revenue [61]. For example, Google search has a much higher per unit revenue than YouTube. The revenue can be generated by advertisements (e.g., YouTube), or value-added services (e.g., Tencent), or other e-commerce (e.g., Amazon). The cost of CP i consists of two parts: 1) the cost c_i for the connection service of per unit traffic, and 2) the additional cost p for per unit sponsored traffic. Thus, the utility of CP i, denoted by ϕ_i, is:

$$\phi_i(c_i, p) = \begin{cases} (v_i - c_i)x_i(q) & i \in \mathcal{O}, \\ (v_i - c_i - p)\theta_i(q) & i \in \mathcal{S}. \end{cases} \tag{4.13}$$

Content providers' surplus, defined as the summation of utilities of all CPs, can be expressed as

$$\phi = \sum_{i \in \mathcal{N}} \phi_i. \tag{4.14}$$

Utility of the ISP. We use the ISP's revenue to represent its utility[7] , mainly from two sources: 1) the unit price charged to CPs for the connection service, i.e., c_i, and 2) the unit price charged to CPs for the sponsored traffic, i.e., p. We omit the price charged to end users because it is only a constant under the cap scheme. Thus, the utility (or payoff) of the ISP, denoted by π, is:

$$\pi(c_i, p) = \sum_{i \in \mathcal{S}} (c_i + p)\theta_i(q) + \sum_{i \in \mathcal{O}} c_i x_i(q). \tag{4.15}$$

Note that we treat the unit sponsoring price for different CPs as equal so as to cope with the network neutrality rules.

[7]We ignore the cost for delivering per unit traffic since the fixed cost is majority while the marginal cost is negligible.

4.2.4 A Two-stage Stackelberg Game

We model the interactions of the ISP and CPs as a two-stage Stackelberg game in the system $(M, \mu, \mathcal{N})$. In particular, we have the following settings:

- *Players*: The ISP and the set of CPs.
- *Strategies*: The ISP decides the unit price charged to CPs for the sponsored traffic, and the traffic cap for end users, i.e., the ISP's strategy profile is $s_I \in \{(p, C) : p \geq 0, C \geq 0\}$. Each CP decides to join either the ordinary class or the sponsored class. We use $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{S})$ to denote CPs' strategy profile, with $\mathcal{O} \cap \mathcal{S} = \emptyset$ and $\mathcal{O} \cup \mathcal{S} = \mathcal{N}$.
- *Rules*: The ISP is the first mover who decides its price and traffic cap and announces them to CPs and end users. CPs are second movers and decide which class to join. Each CP makes its own decision independently.
- *Outcome*: The outcome is determined by backward induction. In particular, given any ISP's decision, each CP chooses which class to join to maximize its utility. Based on this knowledge, the ISP decides its optimal price and traffic cap that maximize its utility.

Note that we do not include the decision of c_i, i.e., the unit price for connection services, into the ISP's strategy profile. This is because we want to focus on the sponsored data scheme, which influences customers' decisions, but has limited impacts on c_i. Therefore we assume c_i is predetermined and known. We apply the Stackelberg game where the ISP is the first mover and CPs are second movers. This reflects the reality where ISPs usually have the monopolistic power and are active to promote the sponsored data plan. Once the ISP fixes its charging scheme, it cannot frequently change it as its contract with CPs and end users are normally of long term. After the ISP's decision, CPs decide whether they sponsor the content. Since CPs make their decisions simultaneously, we call their decision process a *simultaneous game* denoted by $(M, \mu, \mathcal{N}, s_I)$. Following

the backward induction, we analyze the CPs' decisions, i.e., the simultaneous game, in Section 4.3, and later, the ISP's decision, in Section 4.4.

4.3 Content Providers' Decisions

In this section, we analyze content providers' decisions, i.e., the outcome of the simultaneous game $(M, \mu, \mathcal{N}, s_I)$. In the decision phase, a CP joins a particular class ($\mathcal{O}$ or $\mathcal{S}$) where he can obtain a higher utility. Note that upon joining a particular class, this CP may impact the QoS index and the traffic consumption of other services. However, we consider when the number of CPs is large, this effect is ignorable, and thus define *competitive equilibrium* as follows:

Definition 4.3.1. *A strategy profile $s_{\mathcal{N}} = (\mathcal{O}, \mathcal{S})$ is a competitive equilibrium of the game $(M, \mu, \mathcal{N}, s_I)$ if for any CP i, its utility satisfies:*

$$\frac{v_i - c_i - p}{v_i - c_i} \begin{cases} \leq \frac{x_i(\mathcal{O},\mathcal{S})}{\widetilde{\theta}_i(\mathcal{O},\mathcal{S})} & \text{if } i \in \mathcal{O}, \\ > \frac{\widetilde{x}_i(\mathcal{O},\mathcal{S})}{\theta_i(\mathcal{O},\mathcal{S})} & \text{if } i \in \mathcal{S}, \end{cases} \tag{4.16}$$

where $\widetilde{x}_i$ and $\widetilde{\theta}_i$ are the estimation of the ex-post traffic usage $x_i(\mathcal{O} \cup \{i\}, \mathcal{S}/\{i\})$ and $\theta_i(\mathcal{O}/\{i\}, \mathcal{S} \cup \{i\})$ accordingly.

Definition 4.3.1 states that under the competitive equilibrium, CPs in each class cannot obtain a higher profit by joining the other class. The competitive equilibrium depends on the estimation of the ex-post traffic usage for $\widetilde{x}_i$ and $\widetilde{\theta}_i$ which are obtained by $\widetilde{x}_i = g_i^{-1}\left(\frac{t_i+\widetilde{\nu}}{h_i(\widetilde{q})}\right)$ and $\widetilde{\theta}_i = g_i^{-1}\left(\frac{t_i}{h_i(\widetilde{q})}\right)$, i.e., we can calculate the ex-post traffic usage by estimating the ex-post QoS $\widetilde{q}$ and ex-post level of competition $\widetilde{\nu}$. This estimation for the two parameters $(\widetilde{\nu}, \widetilde{q})$, called *congestion metric*, alleviates each CP from estimating all other CPs' characteristics and greatly simplifies our analysis.

4.3.1 Outcome of the Simultaneous Game

Intuitively, we can use the competitive equilibrium to capture the steady state of the system, or the outcome of the simultaneous game. However, we will show later that

there does not always exist a competitive equilibrium for the game. In this section, we derive the conditions for the existence of the competitive equilibrium, and explore how to determine the outcome of the game when there is no equilibrium. We also design an algorithm to quickly find out the unique outcome of the game over the feasible space of exponential size.

Sufficient Capacity

According to Definition 4.2.1, when the ISP has a sufficient capacity, it can support the traffic demand from all users when all services are with the best QoS, i.e., $q = 1$. This is the reality of the broadband (or wired) network, in particular, with fiber access. As for the wireless network, with the development of 4G LTE, the total demand may be fully satisfied in the future so as to fulfill a sufficient capacity condition. In this case, an end user's traffic consumption of service i is:

$$x_i = \begin{cases} \max\left\{0, g_i^{-1}(t_i + \widetilde{\nu})\right\} & i \in \mathcal{O}, \\ g_i^{-1}(t_i) & i \in \mathcal{S}. \end{cases} \tag{4.17}$$

To see each CP's decision, we first define the *relative priority* of CP i as:

$$\rho_i = g_i\left((v_i - c_i - p)g_i^{-1}(t_i)/(v_i - c_i)\right) - t_i. \tag{4.18}$$

This relative priority is the highest critical level of competition that CP i can tolerate in the ordinary class. A smaller relative priority of a CP means a higher incentive or priority to join the sponsored class $\mathcal{S}$. Then, CP i's choice is:

$$i \in \begin{cases} \mathcal{O} & \text{if } \rho_i \geq \widetilde{\nu}, \\ \mathcal{S} & \text{if } \rho_i < \widetilde{\nu}. \end{cases} \tag{4.19}$$

The key point to find an equilibrium is to decide the corresponding level of competition $\widetilde{\nu}$. We first relabel the CPs according to a non-increasing order of ρ_i such that $\rho_i \geq \rho_j$ if $i < j$. Then, we define set $\mathcal{H}_l$ as the set of first l CPs. Given the choices of CPs as $(\mathcal{H}_l, \mathcal{N}/\mathcal{H}_l)$, we can obtain the level of competition $\nu(\mathcal{H}_l)$ according to the optimization problem (4.3), and this value increases with respect to l. When the sequences $\{\nu(\mathcal{H}_l)\}$ and $\{\rho_l\}$ are smooth enough, i.e., the differences between any two neighboring

elements are small enough, we can approximately view them as continuous sequences. The intersection points of these two sequences, if exist, are the levels of competition $\widetilde{\nu}$. Otherwise, the level of competition is zero or $\nu(\mathcal{H}_N)$. We use this rough description to illustrate the idea on finding the level of competition; in what follows we describe the detailed conditions for the existence of competitive equilibria based on this idea.

Theorem 4.3.1. *If there does not exist a positive number l such that $\nu(\mathcal{H}_{l-1}) < \rho_l < \nu(\mathcal{H}_l)$, then there exists at least one competitive equilibrium.*

Proof. If $\rho_1 \leq \nu(\mathcal{H}_1)$ or $\rho_N \geq \nu(\mathcal{H}_N)$, the equilibrium is $(\emptyset, \mathcal{N})$ and $(\mathcal{N}, \emptyset)$ respectively. When $\rho_1 \geq \nu(\mathcal{H}_1)$ and $\rho_N \leq \nu(\mathcal{H}_N)$, there exist one l that $\rho_l \geq \nu(\mathcal{H}_l)$ and $\rho_{l+1} \leq \nu(\mathcal{H}_{l+1})$. If $\rho_{l+1} = \nu(\mathcal{H}_{l+1})$, then $(\mathcal{H}_{l+1}, \mathcal{N}/\mathcal{H}_{l+1})$ is one equilibrium. Otherwise, $\rho_{l+1} \leq \nu(\mathcal{H}_l)$ according to the conditions. It means that $(\mathcal{H}_l, \mathcal{N}/\mathcal{H}_l)$ is one equilibrium. □

When there exists a positive number l such that $\nu(\mathcal{H}_{l-1}) < \rho_l < \nu(\mathcal{H}_l)$, neither $\mathcal{S}$ nor $\mathcal{O}$ will be chosen by CP l. When CP l chooses $\mathcal{O}$, then $\mathcal{S}$ becomes a better choice since $\rho_l < \nu(\mathcal{H}_l)$; when CP l chooses $\mathcal{S}$, then $\mathcal{O}$ becomes a better choice since $\nu(\mathcal{H}_{l-1}) < \rho_l$. The intuition is that CP l's decision to join either class results in a jump on the level of competition, i.e., $\nu(\mathcal{H}_l) - \nu(\mathcal{H}_{l-1})$. This jump can in turn change the original decision of CP l.

When there is no competitive equilibrium, we still need to analyze the outcome of this game. Due to the fact that the sponsored data plan contracts between the ISP and CPs are usually of long term, it is impossible for CPs to always change their decisions, but in fact, their decisions will last for a relatively long time. To determine such decisions as the outcome of the game, we assume that any particular CP, say, CP l, makes its decision according to the following nearest point rule:

$$l \in \begin{cases} \mathcal{O} & \text{if } \nu(\mathcal{H}_l) - \rho_l \leq \rho_l - \nu(\mathcal{H}_{l-1}), \\ \mathcal{S} & \text{if } \nu(\mathcal{H}_l) - \rho_l > \rho_l - \nu(\mathcal{H}_{l-1}). \end{cases} \tag{4.20}$$

The key point for this rule is that the unstable CP joins the set with the level of competition nearer to its relative priority. A binary-search algorithm can be designed to find

the outcome of the game when the ISP's capacity is sufficient. We denote the binary-search algorithm as $FindEqInt()$ with input $\{\rho_l\}$, $\{\mathcal{H}_l\}$ and $\{\nu(\mathcal{H}_l)\}$. This algorithm adopts half-interval search to find the intersection point between sorted sequence $\{\rho_l\}$ and $\{\nu(\mathcal{H}_l)\}$ according to the nearest point rule.

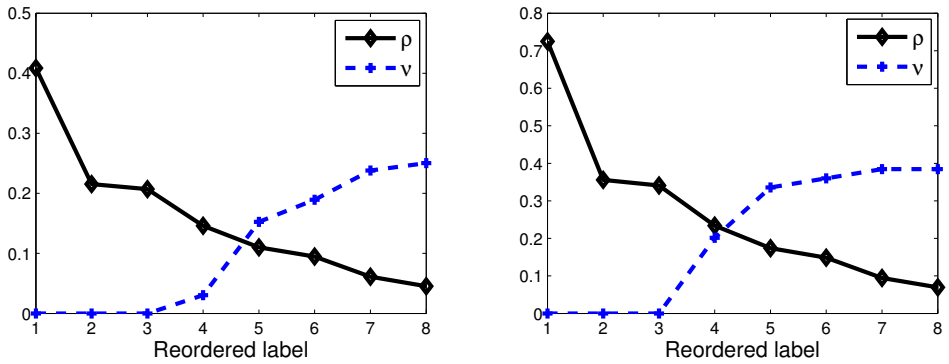
Numerical Example: To intuitively understand the outcome of the game, we give a numerical example of eight CPs with parameters $\alpha_i \in \{1, 3\}$, $\beta_i \in \{1, 2\}$ and $\gamma_i \in \{0.5, 1.5\}$.[8] We set $c_i = 1$ and $t_i = 0.5$. The per unit revenues for the eight CPs are $\{3, \cdots, 10\}$ accordingly.

Figure 4.2 shows the level of competition sequences $\{\nu(\mathcal{H}_l)\}$ and relative priority sequences $\{\rho_l\}$ under two cases of the ISP's strategies, i.e., $s_1 = (1, 5)$ and $s_2 = (1.5, 4)$. The set $\mathcal{H}_N = \{2, 4, 1, 6, 8, 3, 5, 7\}$ and the first four CPs join $\mathcal{O}$ under both strategies s_1 and s_2 since $\rho_l > \nu(\mathcal{H}_l)$ for $l \in \{1, 2, 3, 4\}$. Then let us consider CP 8 under s_1. When it joins $\mathcal{O}$, it finds $\rho_5 < \nu(\mathcal{H}_5)$ and thus joining $\mathcal{S}$ is a better choice. However, when it joins $\mathcal{S}$, it finds $\rho_5 > \nu(\mathcal{H}_4)$ and thus joining $\mathcal{O}$ is better. Therefore, there is no equilibrium for CP 8. This oscillation does not happen under s_2 since $\rho_5 < \nu(\mathcal{H}_4)$. Thus CP 8 joins $\mathcal{S}$, and the equilibrium of the simultaneous game is $\mathcal{O} = \{2, 4, 1, 6\}$ and $\mathcal{S} = \{8, 3, 5, 7\}$. When we adopt the nearest point rule in Eq. 4.20, CP 8 joins the$\mathcal{O}$ under s_1, so the outcome of the game is $\mathcal{O} = \{2, 4, 1, 6, 8\}$ and $\mathcal{S} = \{3, 5, 7\}$.

Insufficient Capacity

Currently wireless data networks often lack capacity. ISPs often cannot support all traffic demand under the best possible QoS. Traffic cap is usually set to limit end users' traffic usage so as to alleviate the congestion problem. Finding an equilibrium of the game $(M, \mu, \mathcal{N}, s_I)$ under the insufficient capacity is complex since we need to consider a pair of parameters $(\widetilde{\nu}, \widetilde{q})$ rather than one single parameter $\widetilde{\nu}$. The interactions between $\widetilde{\nu}$ and $\widetilde{q}$ make the problem more complicated. To find an equilibrium, we first fix the QoS

[8] In this example, the sequence of parameters of CP i is given by $(\gamma\beta\alpha)_2 = (i-1)_{10}$. $\alpha = 0$ indicates CP i chooses the first value of α_i, i.e., $\alpha_i = 1$; otherwise $\alpha_i = 3$ is chosen. Similarly, β_i and γ_i are determined by sequence β and γ.



(a) $s_1 = (1, 5)$ (b) $s_2 = (1.5, 4)$

Figure 4.2: Examples of equilibria under sufficient capacity

index q. Then we can obtain the relative priority sequence $\{\rho_l(q)\}$ under this QoS index. Similar to the sufficient capacity case, the competitive equilibrium $\big(\mathcal{O}(q), \mathcal{S}(q)\big)$ can be calculated according to the binary-search algorithm. This equilibrium results in a new QoS index, i.e., $q'\big(\mathcal{O}(q), \mathcal{S}(q)\big)$. When this new QoS index is equal to the original QoS index, i.e., $q'\big(\mathcal{O}(q), \mathcal{S}(q)\big) = q$, then we obtain a competitive equilibrium $\big(\mathcal{O}(q), \mathcal{S}(q)\big)$. The following theorem quantifies the condition for the existence of competitive equilibria under the insufficient capacity case.

Theorem 4.3.2. *If there does not exist a QoS index q^* such that $q\big(\mathcal{O}(q^+), \mathcal{S}(q^+)\big) < q^* < q\big(\mathcal{O}(q^-), \mathcal{S}(q^-)\big)$ where $q^- = q^* - \epsilon$, $q^+ = q^* + \epsilon$ (ϵ is any sufficiently small positive number), then there exists at least one competitive equilibrium.*

Proof. Define the function $Q(q)$ as the QoS induced by CPs' choices $\big(\mathcal{O}(q), \mathcal{S}(q)\big)$. We first consider the two steps that determine the CPs' choices. The first one is to rerank the sequences of CPs according to their relative priorities and the second step is to find the across point of priority sequence and level of competition sequence. These two steps have finite states and each state respects to one induced QoS. That means the $Q(q)$ has finite valuations during $[0, 1]$. We then prove the theorem by contradiction. Note that $Q(0) = 1$ and $Q(1) \leq 1$. If there does not exist any competitive equilibrium, then $Q(1) < 1$. Then, there must exist some discontinuous point q that $Q(q^+) < q < Q(q^-)$ that contradicts to the conditions. Thus, we finish the proof. □

If there exists a QoS index q^* such that $q\big(\mathcal{O}(q^+),\mathcal{S}(q^+)\big) < q^* < q\big(\mathcal{O}(q^-),\mathcal{S}(q^-)\big)$, then the choices of CPs oscillate. Under different choices of CPs, the QoS index changes around q^* but never converges to a stable point. Similarly, in order to capture the outcome of the simultaneous game, we also adopt the nearest point rule, i.e.,

$$(\mathcal{O},\mathcal{S}) \leftarrow \begin{cases} \big(\mathcal{O}(q^-),\mathcal{S}(q^-)\big) & \text{if } q^* \geq \overline{q}, \\ \big(\mathcal{O}(q^+),\mathcal{S}(q^+)\big) & \text{if } q^* < \overline{q}, \end{cases} \tag{4.21}$$

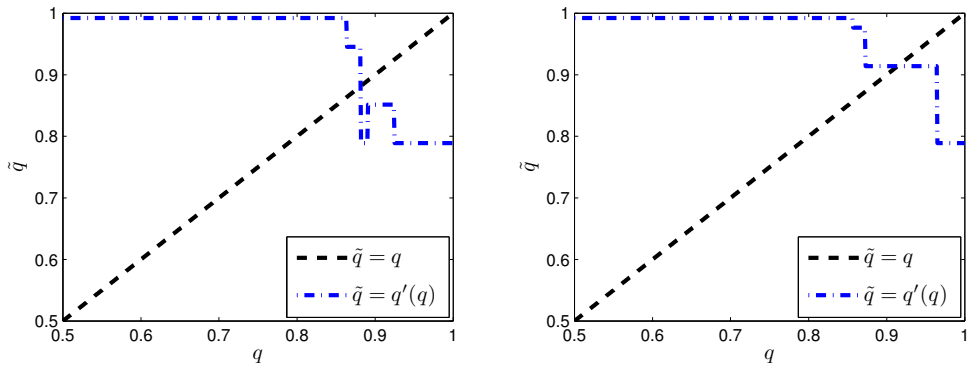
where $\overline{q} = \frac{1}{2}\left(q\left(\mathcal{O}(q^-),\mathcal{S}(q^-)\right) + q\left(\mathcal{O}(q^+),\mathcal{S}(q^+)\right)\right)$.

Algorithm 1 FindEq()

Input: $(N,\mu,\mathcal{M},s_I)$
Output: $(\mathcal{O},\mathcal{S})$
1: Initialize $(\nu[0],q[0])$;
2: Calculate $(\mathcal{O}_{[0]},\mathcal{S}_{[0]})$ induced by $(\nu[0],q[0])$;
3: $t \leftarrow 0$;
4: **do**
5: Calculate $\{\rho_l(q[t])\}$ and sort them according to a non-increasing order;
6: Calculate $\{\mathcal{H}_l[t]\}$ and $\{\nu(\mathcal{H}_l[t])\}$;
7: $(\mathcal{O}_{[t+1]},\mathcal{S}_{[t+1]}) \leftarrow FindEqInt(\{\rho_l(q[t])\},\{\mathcal{H}_l[t]\},\{\nu(\mathcal{H}_l[t])\})$;
8: Calculate $(\nu'[t],q'[t])$ induced by $(\mathcal{O}_{[t+1]},\mathcal{S}_{[t+1]})$;
9: $q[t+1] \leftarrow q[t] + g[t](q'[t] - q[t])$;
10: $t \leftarrow t+1$;
11: **until** $t < T$ or $(\mathcal{O}_{[t]},\mathcal{S}_{[t]}) == (\mathcal{O}_{[t-1]},\mathcal{S}_{[t-1]})$
12: **return** $(\mathcal{O}_{[t]},\mathcal{S}_{[t]})$.

We design Algorithm 1, called $FindEq()$, to search the outcome of the game under the insufficient capacity. It starts with initializing congestion metric $(\nu[0],q[0])$ and calculating CPs' choices $(\mathcal{O}_{[0]},\mathcal{S}_{[0]})$ (line 1 to 3). In each step t, after obtaining the relative priority of CPs under the QoS index $q[t]$, the algorithm calculates the outcome according to the binary-search algorithm $FindEqInt()$ (line 5 to 7). Then it updates the QoS index $q[t+1]$ based on $q[t]$ and the induced QoS index $q'[t]$ (line 8 to 10). The step size parameter $g[t]$ can be time-static or decreasing in t. The algorithm terminates when the round time approaches the maximal number, i.e., T, or the outcome is stable (line 11).

Numerical Example: We provide a numerical example with the same CPs under the sufficient capacity case (in Section 4.3.1). We set the capacity as $\mu = 6$. Figure 4.3 shows the QoS index under the outcome for two ISP's strategies, i.e., $s_1 = (0.5,5)$ and



(a) $s_1 = (0.5, 5)$ (b) $s_2 = (1, 4)$

Figure 4.3: Examples of equilibria under insufficient capacity

$s_2 = (1, 4)$. This QoS index is determined by the cross point of $\tilde{q} = q'(\mathcal{O}(q), \mathcal{S}(q))$ and $\tilde{q} = q$ as discussed previously. Figure 4.3(a) shows non-existence of equilibra. We have $q^* = 0.8815$ such that $q\big(\mathcal{O}(q^+), \mathcal{S}(q^+)\big) < q^* < q\big(\mathcal{O}(q^-), \mathcal{S}(q^-)\big)$. The CPs' choices oscillate between $(\mathcal{O}, \mathcal{S}) = (\{2, 6, 8, 4, 5, 7\}, \{1, 3\})$ with QoS index $\tilde{q} = 0.9453$ and $(\mathcal{O}, \mathcal{S}) = (\{2, 6, 8, 4, 5\}, \{1, 3, 7\})$ with QoS index $\tilde{q} = 0.7891$. When we adopt the nearest point rule in Eq. 4.21, $\bar{q} = 0.8672 < q^*$, the outcome is $(\mathcal{O}, \mathcal{S}) = (\{2, 6, 8, 4, 5, 7\}, \{1, 3\})$. Figure 4.3(b) shows the case with no oscillation and thus there is a unique equilibrium $(\mathcal{O}, \mathcal{S}) = (\{2, 4, 1\}, \{3, 5, 6, 7, 8\})$.

4.3.2 Characteristics of the Outcome

Given the ISP's decision $s_I = (p, C)$, content providers play the simultaneous game $(M, \mu, \mathcal{N}, s_I)$. We denote the outcome of the game $(M, \mu, \mathcal{N}, s_I)$ as $(\mathcal{O}, \mathcal{S})$ and the corresponding congestion metric as (ν, q). Given different decisions of the ISP, CPs play different simultaneous games and may lead to different outcomes. We have the following theorem to quantify the condition that leads to the same outcome.

Theorem 4.3.3. *Consider a new decision of the ISP $s_I' \succeq s_I$ and denote the outcome of the new simultaneous game $(M, \mu, \mathcal{N}, s_I')$ as $(\mathcal{O}', \mathcal{S}')$. If $(\mathcal{O}', \mathcal{S}') = (\mathcal{O}, \mathcal{S})$, then for any $\tilde{s}_I$ satisfying $s_I \preceq \tilde{s}_I \preceq s_I'$, $(\mathcal{O}, \mathcal{S})$ is also the outcome of the new game $(M, \mu, \mathcal{N}, \tilde{s}_I)$.*

Proof. We consider the relative priority $\rho_i(q) = h_i(q) g_i((v_i - c_i - p) g_i^{-1}(\frac{t_i}{h_i(q)}))/(v_i -$

$c_i)) - t_i$. Before we prove the theorem above, we first prove that $\rho_i(q)$ increases with q. Denote $x = g_i^{-1}(\frac{t_i}{h_i(q)})$ and thus $h_i(q) = \frac{t_i}{g_i(x)}$. Then, the relative priority function becomes $\rho_i(x(q)) = t_i g_i((v_i - c_i - p)x/(v_i - c_i))/g_i(x) - t_i$. Define $H_i(x) = g_i(\epsilon x)/g_i(x)(\epsilon > 1)$. Since $g_i^{-1}(\cdot)$ decreases with q and $h_i(\cdot)$ increases with q, we only need to prove that $H(x)$ decreases with x. Since $\ln(H_i(x))$ has the same monotonicity with $H_i(x)$, we just consider the function $\ln H_i(x) = k_i(\epsilon x) - k_i(x)$, where $k_i(x) = \ln(g_i(x))$. Note that $g_i(x) = \alpha_i \exp(-\beta_i x)$. Then, we have $\ln H_i(x) = \beta_i(1-\epsilon)x$ decreasing with x. Thus, we prove that $\rho_i(q)$ decreases with q. Note that ρ_i increases with p. Then, we consider the CPs' choice $(\mathcal{O}, \mathcal{S})$ under ISP's strategy $\tilde{s}$. The QoS under choice $(\mathcal{O}, \mathcal{S})$ is not affected by p, but non-increasing with C according to Lemma 2. Then, we have $q(s) \geq q(\tilde{s}) \geq q(s')$. Since ρ_i decreases with q and increases with p, we have $\rho_i(s) \leq \rho_i(\tilde{s}) \leq \rho_i(s')$. When we obtain new CPs' choice under QoS $q(\tilde{s})$ according to CPs' choice in Eq 4.19, it is the same as $(\mathcal{O}, \mathcal{S})$. That means $(\mathcal{O}, \mathcal{S})$ is also the equilibrium under ISP's strategy $\tilde{s}$. □

Theorem 4.3.3 states that when two decisions of the ISP lead to the same outcome of the simultaneous game, then any decision in between has the same outcome. This stability of CPs' outcome provides some flexibility for the ISP's strategy. For example, the ISP may increase the price charged to the sponsored data and reduce the traffic cap slightly so as to increase its revenue and improve QoS, where the simultaneous game has the same outcome. As $N \to \infty$, this stability shrinks and finally disappears.

Corollary 4.3.1. *Denote the relative priority of CP i under the outcome of the game $(M, \mu, \mathcal{N}, s_I)$ and the new game $(M, \mu, \mathcal{N}, s_I')$ as ρ_i and ρ_i', respectively. If $s_I' \succeq s_I$ and $(\mathcal{O}, \mathcal{S})$ is the outcome of both games, then $\rho_i' - \nu' \geq \rho_i - \nu$.*

Corollary 4.3.1 states that the gap between the relative priority and the level of competition for each CP increases when the ISP increases its price charged to the sponsored data or the traffic cap. According to CPs' choices in Eq. 4.19, this gives each CP higher incentives to join the ordinary class.

Theorem 4.3.4. *The congestion metric* (ν', q') *under the outcome of the new game* $(M, \epsilon\mu, \mathcal{N}, s_I)(\epsilon \geq 1)$ *satisfies at least one of the following properties: 1)* $\nu' \leq \nu$*; 2)* $q' \geq q$*.*

Proof. To prove this theorem, we consider the *sponsored benefit function*, defined as $f_i(\widetilde{\nu}, \widetilde{q}) = g_i^{-1}(\frac{t_i+\widetilde{\nu}}{h_i(\widetilde{q})})/g_i^{-1}(\frac{t_i}{h_i(\widetilde{q})})$. It reflects the relative benefit of joining sponsored class for CP i. We prove that $f_i(\widetilde{\nu}, \widetilde{q})$ is decreasing with $\widetilde{\nu}$ and increasing with $\widetilde{q}$ at first. This property will also be used in the proofs of later theorems.

Since $g(\cdot)$ is a decreasing function, then its inverse function is also decreasing. Thus, $f_i(\widetilde{\nu}, \widetilde{q})$ is decreasing with $\widetilde{\nu}$. Then, we consider $\widetilde{q}$. Substituting $x_i = \frac{t_i}{h_i(\widetilde{q})}$ and $\epsilon_i = (t_i+\widetilde{\nu})/t_i$, The monotonicity problem for $\widetilde{q}$ can be obtained by the monotonicity problem of function $f_i(x_i) = g_i^{-1}(\epsilon_i x_i)/g_i^{-1}(x_i)$ respecting to x_i. Since $\ln(f_i(x_i))$ has the same monotonicity with $f_i(x_i)$, we just consider the function $\ln f(x_i) = k_i(\epsilon_i x_i) - k_i(x_i)$, where $k_i(x_i) = \ln(g_i^{-1}(x_i))$. Since $g_i(x_i)$ is a decreasing and convex function, $g_i^{'}(x_i) < 0$ and $g_i^{''}(x_i) \geq 0$. Considering the first and second order derivative for function $k_i(x_i)$, we have

$$k_i^{'}(x_i) = 1/(g_i^{-1}(x_i)g_i^{'}(x_i)) < 0, \tag{4.22}$$

and

$$k_i^{''}(x_i) = -(g_i^{''}(x_i)g_i^{-1}(x_i) + 1)(g_i^{-1}(x_i))^2(g_i^{'}(x_i))^2 < 0. \tag{4.23}$$

Then, considering the first order derivative for function $\ln(f_i(x_i))$ and we have:

$$\frac{\partial \ln f_i(x_i)}{\partial x_i} = \epsilon_i k_i^{'}(\epsilon_i x_i) - k_i^{'}(x_i) \leq k_i^{'}(\epsilon_i x_i) - k_i^{'}(x_i) < 0. \tag{4.24}$$

It means that $f_i(x_i)$ is decreasing with x_i. Since $h_i(\widetilde{q})$ is increasing with $\widetilde{q}$, $f_i(x_i(\widetilde{q}))$ is increasing with $\widetilde{q}$.

Then, we can prove the theorem by contradiction. We assume that there exist one new equilibrium $\mathcal{O}^{'}$ and $\mathcal{S}^{'}$ such that $\nu'_{s_I} > \nu_{s_I}$ and $q'_{s_I} < q_{s_I}$. For any $j \in \mathcal{S}$, we know that $f_i(\nu_{s_I}, q_{s_I}) > f_i(\nu'_{s_I}, q'_{s_I})$. It means that $j \in \mathcal{S}'$ and thus $\mathcal{S} \subseteq \mathcal{S}'$. Since $\mathcal{O} \cup \mathcal{S} = \mathcal{N}$ and $\mathcal{O}' \cup \mathcal{S}' = \mathcal{N}$, $\mathcal{O}' \subseteq \mathcal{O}$. According to Lemma 1, for any $j \in \mathcal{O}'$, $x'_j < x_j$. That means

$\sum_{j\in\mathcal{O}'} x'_j < \sum_{j\in\mathcal{O}'} x'_j$ and thus $\sum_{j\in\mathcal{O}'} x'_j < C$. This is impossible since $\nu'_{s_I} > 0$. Thus, we finish the proof. □

Theorem 4.3.4 states that if the ISP's capacity increases, then QoS improves, or the level of competition reduces. High QoS increases the traffic demand in $\mathcal{S}$, while low level of competition increases the traffic demand in $\mathcal{O}$. Therefore, when the ISP's capacity increases, it may benefit CPs in $\mathcal{S}$, or those in $\mathcal{O}$, or both.

Each CP's decision depends on their own features (v_i, c_i), the quality sensitivity γ_i, the pre-set threshold t_i of end users, and the traffic sensitivity β_i. To investigate the incentives of various CPs on joining the sponsored plan, we study a set $\mathcal{T}$ of CPs with the same pre-set threshold t_i and traffic sensitivity β_i but different features (v_i, c_i) and quality sensitivity γ_i. This setting represents those with similar services but differing in size or technology. We say that CPs in $\mathcal{T}$ are of the *same type*. We have the following theorem.

Theorem 4.3.5. *If CP $j \in \mathcal{T}$ joins $\mathcal{S}$ under the outcome of the game $(M, \mu, \mathcal{N}, s_I)$, then any other CP $i \in \mathcal{T}$ that satisfies $v_i - c_i \geq v_j - c_j$ and $\gamma_i \geq \gamma_j$ also joins $\mathcal{S}$.*

Proof. Since $i, j \in \mathcal{T}$, we have $g_i^{-1}(\cdot) = g_j^{-1}(\cdot)$. Since $h_i(q) \leq h_j(q)$, we have $f_i(\nu, q) \leq f_j(\nu, q)$. Since $v_i - c_i \geq v_j - c_j$, we have $\frac{v_i - c_i - p}{v_i - c_i} \geq \frac{v_j - c_j - p}{v_j - c_j}$. Thus, $\frac{v_i - c_i - p}{v_i - c_i} - f_i(\nu, q) \geq \frac{v_j - c_j - p}{v_j - c_j} - f_j(\nu, q)$. if CP $j \in \mathcal{T}$ joins $\mathcal{S}$, we have $\frac{v_j - c_j - p}{v_j - c_j} - f_j(\nu, q) > 0$. Then, we have $\frac{v_i - c_i - p}{v_i - c_i} > f_i(\nu, q)$ and thus CP i joins $\mathcal{S}$. □

Theorem 4.3.5 indicates that for the same type of CPs, those with high per unit revenue or quality sensitivity usually have high incentives to join the sponsored class, and in turn they have potential to achieve higher revenue. This may result in unfair competition and encourage CPs to pursue capital instead of improving their quality of service.

4.4 Monopolistic ISP's Strategy

In the previous section, we have analyzed the outcome of the simultaneous game, i.e., the second stage of the Stackelberg game. In this section, we discuss the first stage of

the Stackelberg game, i.e., the monopolistic ISP's best choice, so that we can understand the outcome of the Stackelberg game and its impacts to CPs and end users.

4.4.1 Sufficient Capacity

When the ISP has a sufficient capacity, it can support all demands with the best QoS. However, this does not imply that the ISP has an incentive to release the cap to users.

Theorem 4.4.1. *Given any strategy $s_I = (p, C)$ of the ISP and a sufficiently small $\epsilon > 0$, the strategy s_I is dominated by $s_I^+ = (p, C + \epsilon)$ if and only if CPs' decisions remain unchanged under the new game $(M, \mu, \mathcal{N}, s_I^+)$.*

Proof. Sufficiency: Denote the traffic usage under the game $(M, \mu, \mathcal{N}, s_I^+)$ as x_i^+ for service i. The traffic usage for the services in ordinary class non-decreases with traffic cap according to Lemma 1. If the CPs' decisions remain unchanged under the new game $(M, \mu, \mathcal{N}, s_I^+)$, i.e., still $(\mathcal{O}, \mathcal{S})$, then we have $x_i \leq x_i^+$ for any $i \in \mathcal{O}$. In addition, the traffic usage for sponsored class keep the same. The ISP can obtain no less profit from the new strategy s_I^+. Thus, the strategy of s_I is dominated by s_I^+.

Necessity: Define the set $R^+(\rho) = \{j : \rho_j \geq \rho\}$. We first show that if $i \in \mathcal{O}$, than $R^+(\rho_i) \subseteq \mathcal{O}$. Denote the level of competition under the equilibrium as $\widetilde{\nu}$. When $i \in \mathcal{O}$, it means $\rho_i \geq \widetilde{\nu}$. Then, for any $j \in R_i^+$, $\rho_j \geq \widetilde{\nu}$. Thus, CP j joins $\mathcal{O}$ and thus $R^+(\rho_i) \subseteq \mathcal{O}$. Define $\bar{\rho}(\mathcal{O}) = \min\{\rho_i : i \in \mathcal{O}\}$. Then, $R^+(\bar{\rho}(\mathcal{O})) = \mathcal{O}$ and $\bar{\rho}(\mathcal{O}) \geq \tilde{\nu}$. We then prove the necessary condition by contradiction. We assume that the equilibrium under the new game is changed. Denote the new equilibrium as $(\mathcal{O}^+, \mathcal{S}^+)$ and the corresponding level of competition as $\widetilde{\nu}^+$. Obviously, $\widetilde{\nu}^+ < \tilde{\nu}$. If not, $\mathcal{O}^+ = R^+(\widetilde{\nu}^+) \subseteq R^+(\widetilde{\nu}) = \mathcal{O}$. This makes $\widetilde{\nu}^+ < \widetilde{\nu}$ and results in contradiction. Thus, $\mathcal{O} \subset \mathcal{O}^+$ and $\mathcal{S}^+ \subset \mathcal{S}$. This results in a profit loss $\sum_{i \in \mathcal{O}^+/\mathcal{O}} [(p + c_i)\theta_i - c_i x_i]$. On the other hand, the profit obtained from the additional capacity ϵ is upper bounded by $\epsilon \max\{c_i\}$, that is ignorable compared with the profit loss when ϵ is small enough. That means the ISP prefers s_I to s_I^+ and this contradicts the assumption. Thus, we complete the proof. □

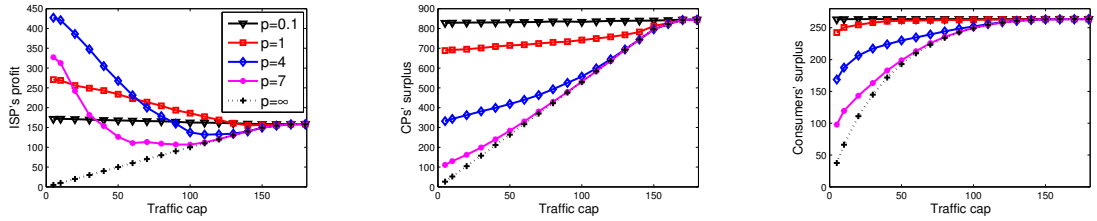
Theorem 4.4.1 states that the ISP has an incentive to enlarge its traffic cap until the CPs' decisions change. The intuition is that when CP's decisions remain unchanged, a large cap increases the revenue from $\mathcal{O}$. Yet, this may lead CPs in $\mathcal{S}$ switching to $\mathcal{O}$. When this happens, the profit from $\mathcal{S}$ has a jump of reduction. The total profit of the ISP increases if the profit increase from $\mathcal{O}$ dominates the loss from $\mathcal{S}$, or decreases otherwise. If the sponsored data plan is prohibited, i.e., $p \to \infty$, then the ISP will set the cap to infinity. This is because when the capacity is not a constraint, the ISP wants to attract users' demand as much as possible, so that it delivers as much traffic as possible for CPs, and this generates a large income to the ISP charged from CPs.

We conduct simulations with 100 CPs and one ISP to explore the key features of the ISP's strategy. The pre-set threshold t_i is randomly selected from $[0.1, 1]$[9]. The per unit traffic cost of connection services for each CP is normalized as $c_i = 1$. We set the CPs' per unit revenue v_i randomly distributed over $[1, 10]$ and excludes the CPs unable to afford connection services. The quality sensitivity γ_i is uniformly distributed over $[0, 2]$. The parameter pair (α_i, β_i) is chosen randomly from $[1, 10] \times [1, 2]$[10]. We set the ISP's per user capacity as 500, larger than the maximal capacity needed for one user, representing a sufficient capacity. Note that our simulations do not depend on particular settings, and our purpose is to show qualitative trends in general.

We first consider the ISP's optimal traffic cap under different prices, as shown in Figure 4.4(a). When the traffic cap is small, e.g., $C = 20$, the ISP's profit π decreases with the traffic cap. This means the profit loss from $\mathcal{S}$ dominates the increase from $\mathcal{O}$. Charging higher prices to CPs leads to a larger reduction of the ISP's profit when the traffic cap increases. When the traffic cap is large, e.g., $C = 100$, the ISP's profit π increases with respect to the traffic cap. This means the profit increase from $\mathcal{O}$ dominates the loss from $\mathcal{S}$. This happens when most CPs join $\mathcal{O}$. Enlarging the traffic cap increases the ISP's

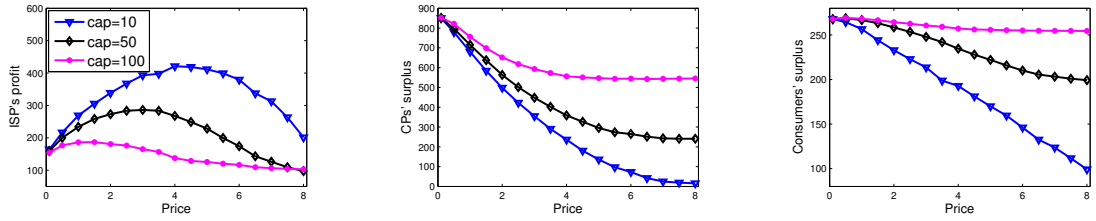
[9]We exclude the interval $[0, 0.1]$ since consumers always have non-ignorable t_i; otherwise they will consume infinite traffic.

[10]We exclude $[0, 1]$ for α_i to ensure non-zero traffic usage and narrow the range of β_i to avoid some CPs' traffic dominating the capacity.



(a) π vs. traffic cap (b) ϕ vs. traffic cap (c) ψ vs. traffic cap

Figure 4.4: π, ϕ, ψ under various prices versus traffic cap



(a) π vs. traffic price (b) ϕ vs. traffic price (c) ψ vs. traffic price

Figure 4.5: π, ϕ, ψ under various traffic caps versus price

profit until the traffic usage in $\mathcal{O}$ approaches the maximal demand. Figure 4.4(b)(c) show that CPs' and consumers' utilities both increase with respect to the traffic cap.

We then show the ISP's optimal price under various traffic caps in Figure 4.5(a). When the price is low, e.g., $p = 0.1$, the ISP's profit increases with the price due to high revenue obtained from $\mathcal{S}$. When the price is high, e.g., $p = 7$, the ISP's profit reduces with the price since more CPs choose to join $\mathcal{O}$, resulting in the reduction of the sponsored traffic. The ISP decides the optimal price that balances the per unit income from the sponsored traffic, and the amount of this traffic. Figure 4.5(b)(c) show that CPs' and consumers' utilities decrease with respect to the traffic cap.

Remark: When the ISP's capacity is sufficient, the ISP, CPs and consumers all benefit from the sponsored data plan in the short run. However, the ISP may not have incentives to enlarge its traffic cap. Keeping a small traffic cap (e.g., $C = 10$) and charging a high price (e.g., $p = 4$) to the sponsored traffic can bring in more revenue for the ISP. This selfish strategy greatly hurts the benefits of both CPs and consumers in the long run. To remedy this problem, the authority may need to put some regulations to the ISP so as to protect the consumers' surplus in the long run. In general, there are two regulation

methods. The first method is to allow the sponsored data plan but regulate the traffic cap, i.e., encouraging the ISP to enlarge the cap. The price for the sponsored data will also decrease accordingly. The other method is to forbid the sponsored data plan. In this case, the ISP has incentives to extend its traffic cap or even provide limitless usage service.

4.4.2 Insufficient Capacity

When the ISP's capacity is insufficient, the traffic cap set by the ISP is a good choice to limit the traffic of consumers. We analyze the ISP's strategy with insufficient capacity under the sponsored data plan. We also assume that $c_i = c$ in this subsection. We have the following theorem:

Theorem 4.4.2. *Given the ISP's strategy $s_I = (p, C)$ and a traffic cap $C' \leq C$, s_I is always dominated by (p, C') if the outcome of the simultaneous game $(M, \mu, \mathcal{N}, s_I)$ satisfies $\mathcal{S} \neq \emptyset$ and $q < 1$.*

Proof. Since the sponsored class is non-empty and $q < 1$, the total capacity is fully utilized. If $\sum_{i\in\mathcal{O}} x_i < C$, the QoS and CPs' choices keep the same as C decreases until $\sum_{i\in\mathcal{O}} x_i = C$. That means the ISP's profit keeps the same. If $\sum_{i\in\mathcal{O}} x_i = C$, then $\sum_{i\in\mathcal{S}} x_i = \lambda - C$. Denote the ISP's profit as $\pi(C)$ with the variable C and then it becomes $\pi(C) = cC + (p + c)(\lambda - C) = (P + c)\lambda - pC$. This is a decreasing function with C. Thus, we complete the proof. □

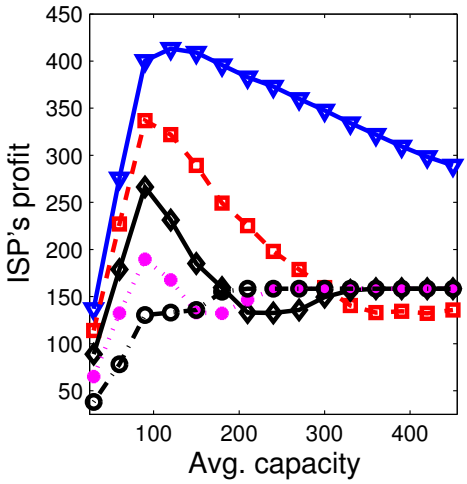
Theorem 4.4.2 says that the ISP is willing to set a small traffic cap so as to increase its profit. The reasons are: 1) the capacity is fully utilized, so it is good to limit users' consumption; and 2) the sponsored traffic brings in more profit to the ISP than the ordinary traffic. When a smaller traffic cap is given, more CPs will join $\mathcal{S}$, indicating a larger profit.

We also evaluate the effects of the sponsored strategy under an insufficient capacity via simulations. The basic settings are the same as the previous subsection. The price

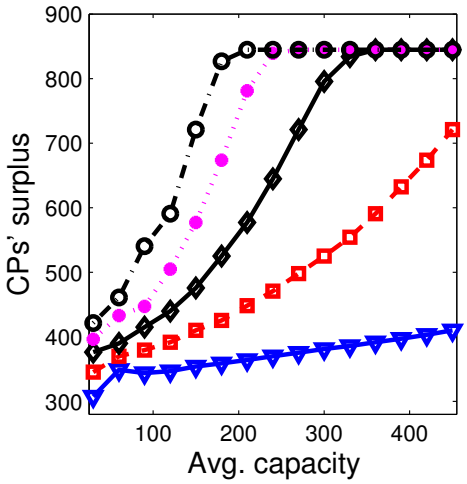
charged for the sponsored traffic is $p = 4$, and the average capacity is $\lambda = 100$ unless otherwise specified. We define the traffic cap ratio as $\kappa = C/\lambda$.

We first consider the effect of the ISP's capacity under various traffic caps, as shown in Figure 4.6. Figure 4.6(a) shows the ISP's profit π under different capacities. When the capacity is small, the ISP's profit increases linearly with respect to the capacity, i.e., $(c + (1 - \kappa)p)\lambda$. This happens when the capacity allocated to $\mathcal{S}$ (or sponsored capacity for short) is fully utilized. When this capacity is under-utilized, the ISP's profit reduces when the capacity increases since more CPs join $\mathcal{O}$. In addition, the ISP can obtain a higher profit by reducing the traffic cap. Figure 4.6(b) shows that the CPs' surplus increases with respect to the traffic capacity until all CPs join $\mathcal{O}$. Figure 4.6(c) shows that consumers' surplus increases with respect to the ISP's capacity. When the sponsored capacity is under-utilized, consumers' surplus increases at a much lower rate. In addition, CPs and consumers benefit more from the increasing cap. Figure 4.6(d) states that QoS improves with respect to the capacity as long as the sponsored capacity is fully utilized. Given a fixed capacity, a larger traffic cap may not indicate a higher QoS.

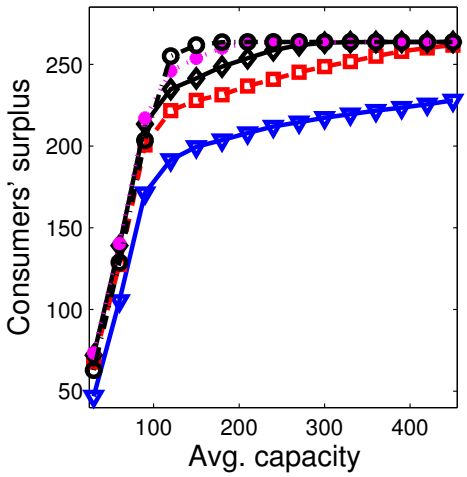
We then focus on the effect of the price charged to the sponsored traffic under various traffic caps, as shown in Figure 4.7. Figure 4.7(a) shows the ISP's profit π under different prices. When the price p is small, the ISP's profit increases linearly with respect to the price, i.e., $c\lambda + (1 - \kappa)\lambda p$. This happens when most CPs can afford the cost of the sponsored traffic so that the sponsored capacity is fully utilized. When p is large, the ISP's profit reduces as the price increases. This happens when the sponsored capacity is under-utilized. The optimal price of the ISP is around $p = 4$. Figure 4.7(b) shows that CPs always prefer higher traffic cap and lower price charged to the sponsored traffic. Figure 4.7(c) shows that consumers' surplus is almost aligned with the ISP's profit. This happens when the traffic cap is not too small. The optimal price for end users is also around $p = 4$. The intuition is that a low price results in serious QoS degradation and reduces the valuation of per unit traffic, while a large price results in the sponsored capacity under-utilized and reduces the traffic amount. Figure 4.7(d) shows that the QoS



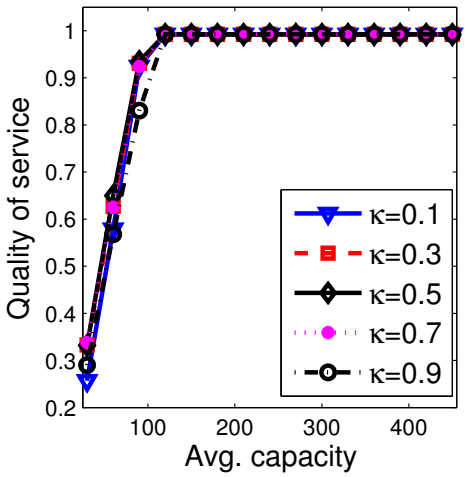
(a) π vs. capacity



(b) ϕ vs. capacity

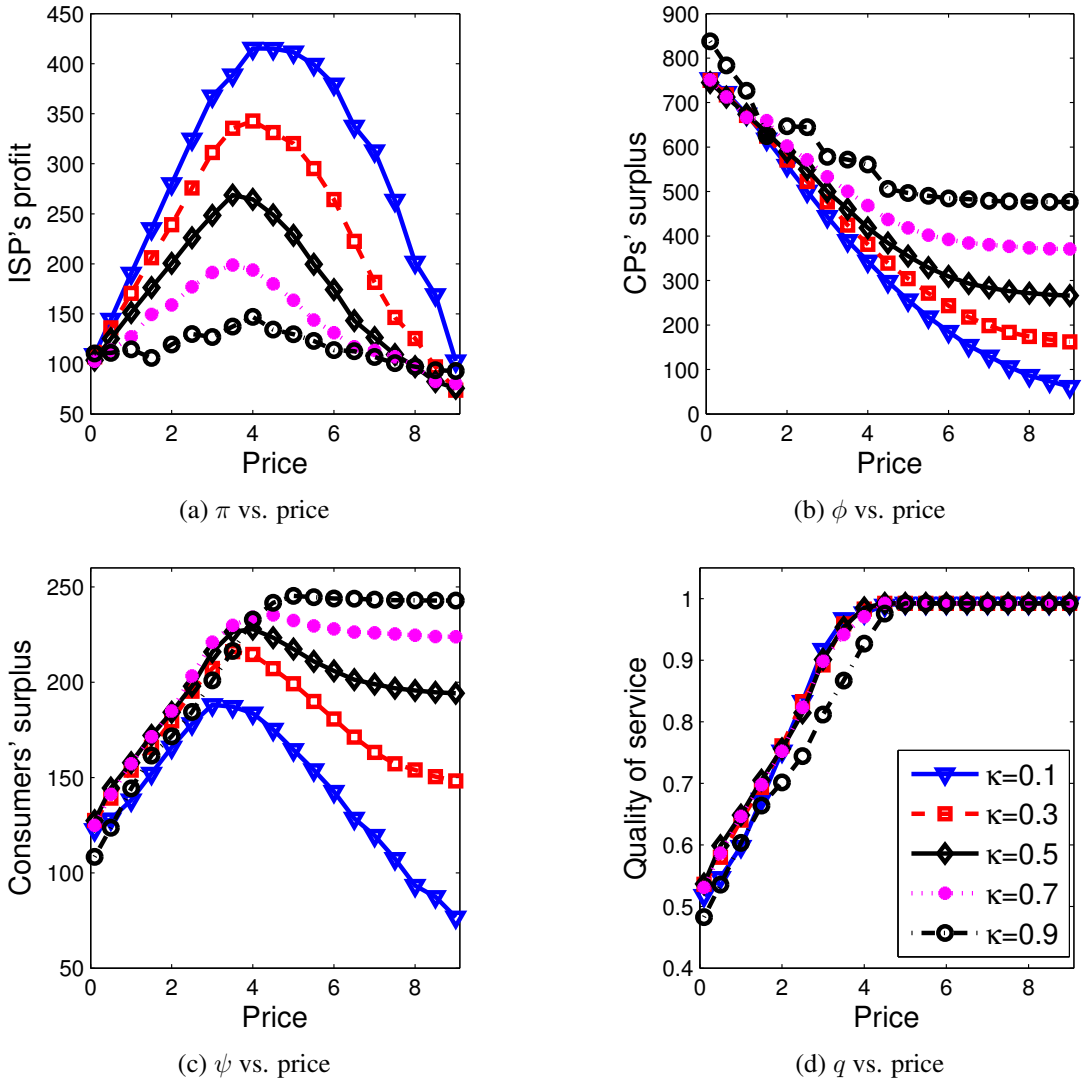


(c) ψ vs. capacity



(d) q vs. capacity

Figure 4.6: π, ϕ, ψ, q versus capacity



(a) π vs. price

(b) ϕ vs. price

(c) ψ vs. price

(d) q vs. price

Figure 4.7: π, ϕ, ψ, q versus price

index increases with respect to the price until it reaches its maximal value, i.e., $q = 1$, since more CPs join $\mathcal{O}$ and less traffic is consumed.

Remark: When the capacity is insufficient, the ISP still has no incentives to enlarge the traffic cap under the sponsored strategy. The traffic cap has a negligible effect on the QoS. Other factors, i.e., the price and the capacity, can be utilized to impact QoS. This may relieve the concerns of QoS degradation by the sponsored data plan. The ISP has a strong incentive to enlarge its traffic capacity until the best QoS achieves. This benefits both CPs and users. The ISP also prefers to set a high price due to a linear increase of its profit, i.e., $(c + (1 - \kappa)p)\lambda$. Yet, when the price is too high, more CPs will join the

ordinary class and the sponsored capacity may be under-utilized. A suitable traffic cap and price charged for the sponsored traffic can increase the ISP's profit so as to support it on enlarging the traffic capacity in the future. The CPs prefer the sponsored strategy with a large traffic cap and a low sponsored price, which is always on the contrary to the ISP's optimal strategy. The ISP's and consumers' surplus can increase simultaneously if the majority of capacity is allocated to the ordinary class.

4.5 Discussion and Limitation

Sponsored data plan was originated from the 1-800 services, but the market for wireless data networks is quite different. Currently, the main 3G and 4G LTE data plans set traffic caps to limit consumers' usage due to insufficient capacities. The sponsored data plan proposed by AT&T provides a method for specific CPs to traverse the traffic caps. A potentially higher revenue may support the investment for larger capacities and thus improve QoS. For the time being, AT&T only provides the toll-free services for its sponsored data plan, probably because consumers have a strong preference to a simple data plan. Our two-class service model (i.e., ordinary class and sponsored class) is built based on the data plans at status quo.

In fact, the sponsored data plan does not differentiate services from CPs in the ordinary and the sponsored classes. In other words, the quality of services is the same for both classes. This is different from PMP [60] or Public Option ISP [54]. However, CPs in the ordinary class face serious levels of competition. This is, in some sense, like the bad QoS challenge faced by CPs in the lower charged channel of PMP [60] or the public option ISP [54]. The ordinary class is mainly preferred by CPs with lower per unit revenue, e.g., startup companies. Content providers with higher per unit revenue, e.g., Google, prefer joining the sponsored class. This may also happen to CPs that are sensitive to QoS since they may fail to compete for the traffic from end users' scare traffic cap. Although the sponsored data plan opens a door for CPs to increases their traffic demand, it also brings the risk that when competition happens, it benefits more to CPs with higher per

unit revenue, instead of those with better technology support.

The sponsored data plan brings higher revenue to ISPs. However, ISPs may prefer high prices for the sponsored data and small traffic caps (Theorem 4.4.2) under current wireless data networks. This selfish strategy hurts both consumers and CPs. Fortunately, the damage for consumers is negnigible when the majority capacity is allocated to the ordinary class (Figures 4.6(c) and 4.7(c)). ISPs can also relieve from the concerns of poor QoS caused by the sponsored data plan (Figures 4.6(d) and 4.7(d)). Suitable revenue encourages ISPs to provide more investment to extend the capacity, which benefit both consumers and CPs. Despite the potential risks, we believe that a fair and transparent sponsored data plan would provide a unified platform for competition among CPs and create a healthy ecosystem in wireless data networks.

Although we build our two-class service model based on the currently dominated data plans and capture the interactions among various components, our work has several limitations. First, our model may not capture the short-term off-equilibrium that usually happens in practice due to some players' non-rational or non-optimal decisions. Second, our two-class model only focuses on a single ISP and its fixed end users. We set up this model not only for mathematical simplicity, but also capture one ISP's monopoly access power for a majority of CPs even in the market with multiple ISPs. Current long-term contracts also limit end users' transition from one ISP to another. However, it is still interesting to explore the competitive market with multiple ISPs. Finally, our numerical evaluations are limited to capture qualitative trends. Carrying out real experiment or detailed validation could be very challenging, since it is quite difficult to obtain an accurate estimation on a number of parameters in our model. Despite these limitations, we still believe our analysis has captured some important insights and might help the scheme and regulation designs for future wireless data pricing markets.

4.6 Summary

In this chapter, we propose a two-class service model, analyze the interaction among end users, content providers and the ISP, and study the impact of the sponsored data plan on the Internet service market. In particular, we focus its impact on the quality of service, the profit of CPs and the ISP, and on shaping the users' traffic consumption behaviors. Our interesting findings include: 1) when the ISP's capacity is sufficient, the sponsored data plan benefits consumers and CPs in the short run, but the ISP does not have incentives to further improve its service in the long run; 2) when the ISP's capacity is insufficient, the ISP and end users may both benefit from the scheme, while the ISP and CPs always compete for the revenue; and 3) the sponsored data plan may enlarge the unbalance in revenue distribution between different CPs and result in unfair competition. Our findings provide important insights to designing sponsored data plans and potentially necessary regulations. We believe that given proper regulations, a fair and transparent sponsored data plan would be a promising trend of pricing models for future wireless data networks.

CHAPTER 5

TDS: TIME-DEPENDENT SPONSORED DATA PLAN FOR WIRELESS DATA TRAFFIC MARKET

5.1 Overview

In the past years, we have witnessed a surge of smart mobile devices, such as smart phones, smart pads, etc. Together with the applications, such as Dropbox, YouTube, the amount of wireless data traffic has grown tremendously. This poses huge burden to the Internet service providers (ISPs) since supporting such demand-supply gap requires large investments.

The ISPs are only one stakeholder in the Internet. The two-sided Internet market can be captured in Fig. 5.1, where ISPs are in the middle, end users (EUs) are on one side and content providers (CPs) are on the other side. Facing the surging demands, usage-based plans start prevailing in wireless data markets over flat-rate unlimited plans. For example, Verizon Wireless charges users for $20 per month for 2 GB amount of data [75]. Such usage-based plans are backed by FCC [59], yet they raise concerns from the CPs because they may intrinsically limit users' willingness to consume data content from the CPs, whose revenue heavily depends on user views. One core problem is the one-sided charge for end users, i.e., ISPs, in particular, the last-mile access ISPs, charge the users as their primary revenue resources. This leads to an unbalanced finance model as neither users want to increase their data consumption and pay more nor ISPs want to reduce their price. New pricing models have been proposed to vitalize the Internet market, among which the sponsored data plan (SDP) [4, 8] attracts special interests from both industry and academia.

SDP, or also called tool-free service, means that an ISP and CPs sign some form of contract, such that when end users access the contents from CPs joining the SDP, their

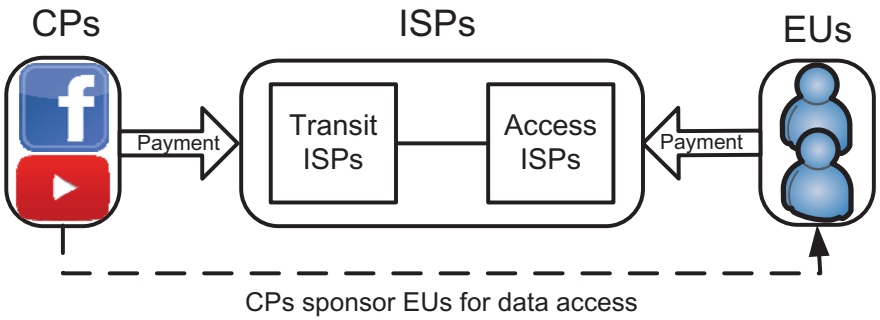


Figure 5.1: Two-sided Internet market

traffic from/to these CPs will not be charged by the ISP. Instead, CPs will pay for that volume of traffic for end users to the ISP. Since its birth, SDP shows great potential to becoming a major charging pattern over the wireless data network. Intrinsically, SDP balances the finance model of the CPs towards ISPs and users as it creates a positive cycle among users, CPs and ISPs: end users are willing to consume more traffic sponsored by CPs; CPs can attract more users and thus more advertisement income; ISPs can obtain more revenue by charging the CPs. Thus, CPs, ISPs and end users may all benefit from this strategy.

SDP also has been in practice. For example, AT&T announced its sponsored data program in January 2014 after a long time of planning [8]. Its sponsored data partner, Syntonic Wireless, launched a toll-free content store six months later [26]. Google has also joined with India's Bharti Airtel to offer free access to certain Google-based services such as Gmail, Google+ and first page of websites via Google search without ringing up data charges [27].

There are emerging studies [51, 83, 85] on SDP, and the research foci are the competition, benefit, equilibrium, fairness, the possible regulations needed, etc., of the Internet market under SDP. These studies have confirmed that SDP can lead to a more balanced finance model for the Internet market. None of the aforementioned works, however, study how to sponsor data. We argue that studies on appropriate sponsoring methods are also important, and even affect the overall success of SDP. In this chapter, we propose and study time-dependent sponsoring (TDS). In TDS, CPs can decide the fraction of traffic to sponsor to end users for a given time, and that the fraction may vary over time.

The main novelty of TDS is its potential to improve resource utilization.

It is non-trivial to analyze TDS. The key challenges include: 1) how to model user behavior as different users may behave differently under TDS, 2) the demands under different times can be correlated due to traffic migration under different subsidizations, 3) the interaction among users, CPs and ISPs is complex, and there need appropriate models for the overall game, comprehensive discussions and interpretations.

In this chapter, we provide a rigid study on TDS. We model the users as strategic users [11, 72], where they may delay their data consumption in exchange for a better price. We establish a Stackelberg game model to capture the interactions among a monopoly ISP, a set of CPs and an arbitrary number of strategic users. We formulate the ISP' s and CPs' decisions (i.e., the market equilibrium) under TDS as optimizations. We show that the sponsoring decision of CPs under TDS is a non-convex optimization due to correlated demands under different times. We propose a dynamic programming based algorithm and solve it in polynomial time. Our main findings include:

- TDS improves the CPs' bandwidth utilization, CPs' profit and consumers' welfare for slightly patient strategic users; but may result in controversial effects for highly patient strategic users;
- When CPs provide different subsidizations to different groups, the CPs' bandwidth utilization can be improved significantly and so are CPs' profit and consumers' welfare;
- Well designed TDS reduces the waste of ISP's capacity and thus improves social welfare and ISP's profit, compared with SDP.

This is the outline of this chapter. In Sec. 5.2, we set up a Stackelberg game model to capture the interactions between users, CPs and the ISP. Sec. 5.3 and 5.4 analyze CPs' and the ISP's optimal decisions respectively, as well as their impacts to the market. Finally, Sec. 5.5 summaries this chapter.

5.2 General Model

In this section, we analyze the market with three parties: a set of CPs $\mathcal{N}$, a monopolistic ISP and a set of end users $\mathcal{M}$. The CPs provide services to end users. We assume that one CP supplies only one service. For a CP providing multiple services, we treat it as multiple virtual CPs. The ISP provides Internet access services to CPs and end users. We define the capacity of the ISP as the bandwidth of the bottleneck for the connection services between CPs and end users, denoted as μ. We use a triple $(\mathcal{N}, \mu, \mathcal{M})$ to represent the whole system.

5.2.1 Consumers' Traffic Demand

We consider the finite time horizon $[0, T]$ with time slots $\{1, \cdots, L\}$ and time length $\Delta \triangleq T/L$ for each slot, where L is the number of slots.[1] End users can arrive at any slot and consume the traffic at the same slot or delay their consumption at one later slot.[2] We denote the *maximal waiting time* as $K \in \{0, \cdots, L-1\}$. In particular, when $K = 0$, end users are impatient to delay their traffic usage of each service. We define the *waiting period* for the users with maximal waiting time K arriving at slot t as $S_t^K \triangleq \big\{t, \cdots, \min\{t+K, L\}\big\}$. Different users may have different maximal waiting time and thus different waiting periods. We denote the fraction of end users with maximal waiting time K as $g(K)$ that satisfies $\sum_{K=0}^{L-1} g(K) = 1$. For the service provided by CP $s \in \mathcal{N}$, short for service s, we denote the population of its users that arrive at slot t as m_t^s. We also denote the average traffic usage during one slot for service s as δ^s and thus the total traffic at slot t becomes $\theta_t^s \triangleq m_t^s \delta^s$ if they consume the traffic immediately.

The prices charged to CPs and end users by the ISP are independent of time. CPs are charged according to the usage of bandwidth. For instance, Netflix pays the CDNs on a per-megabit-per-second-sustained model so as to guarantee the quality of service [66]. Instead of the bandwidth, the ISP charges end users by the traffic usage amount in wire-

[1] For mathematical simplicity, we set $\Delta = 1$ and $T = L$ if not otherwise specified.

[2] Here, we assume that each user only consumes the traffic of one slot. For the user with multi-slot traffic demand, we treat it as multiple users.

less data networks. Under the neutrality rules, we denote p as the price of per unit traffic for all end users $\mathcal{M}$ and q as the price of per unit bandwidth during $[0, T]$ for all CPs $\mathcal{N}$. The price charged to end users can be subsided by CPs. Each CP can provide different subsidizations to its users under different time slots. For any CP $s \in \mathcal{N}$, we denote its subsidization of per unit traffic for its users as h_t^s under time slot t.

For the users of service s arriving at slot t, they can consume their traffic during waiting period S_t^K and obtain a non-negative value for each unit. We assume valuation per unit traffic for service s follows the probability density function $f_s(\cdot)$ and thus the fraction of users below per unit valuation v for service s becomes $F_s(v) \triangleq \int_0^v f_s(x)dx$. Then the utility for one user of service s with valuation per unit traffic v, waiting period S_t^K and choosing slot $k \in S_t^K$, denoted as $u_t^s(k)$, can be expressed as:

$$u_t^s(k) = v - (p - h_k^s). \tag{5.1}$$

The users of service s consume the traffic if and only if they can obtain non-negative utilities, i.e., $\max_{k \in S_t^K}\{u_t^s(k)\} \geq 0$. It indicates that only the users with valuation $v \geq p - \max_{k \in S_t^K}\{h_k^s\}$ consume the traffic of service s. Each user of service s chooses the optimal time that maximizes its utility. Denote the optimal time slot for the users of service s as ϖ_t^s. Then, we have:

$$\varpi_t^s = \arg\max_{k \in S_t^K}\{h_k^s\}. \tag{5.2}$$

It indicates that users always choose the slot with the maximal subsidization. When two slots have the same utility, we assume the users prefer the earlier slot to break the tie. Then, given the subsidization of service s, i.e., $\mathbf{h}^s = (h_1^s, \cdots, h_L^s)$ and the price charged by the ISP to end users for per unit traffic, i.e., p, the potential demand for service s during time slot t, denoted as ρ_t^s can be expressed as:

$$\rho_t^s(\mathbf{h}^s) = \sum_{K=0}^{L-1} \sum_{l: t \in S_l^K} g(K)\theta_l^s \mathcal{I}\{t = \varpi_l^s\}, \tag{5.3}$$

where $\mathcal{I}$ is an indicator function. Then, the actual demand, denoted as D_t^s, becomes:

$$D_t^s(\mathbf{h}^s, p) = \big(1 - F_s(p - h_t^s)\big)\rho_t^s(\mathbf{h}^s). \tag{5.4}$$

5.2.2 Utility of CPs

We use r^s to denote the per unit revenue of CP s for providing services. Different CPs may have much different per unit revenues. For example, Google search has a much higher per unit revenue than YouTube. The revenue can be generated by advertisements (e.g., YouTube), or value-added services (e.g., Tencent), or other e-commerce (e.g., Amazon). The cost of CP s consists of two parts: 1) the cost by sponsoring its users, i.e., $\mathbf{h}^s$ for per unit traffic, and 2) the bandwidth cost charged by the ISP, i.e., q for per unit bandwidth. We denote the required bandwidth for CP s as λ^s.[3] Thus, the utility of CP s, denoted by Φ^s, is:

$$\Phi^s(\mathbf{h}^s, \lambda^s) = \sum_{t=1}^{L} (r^s - h_t^s) D_t^s(\mathbf{h}^s, p) - q\lambda^s. \tag{5.5}$$

Given the prices charged to end users for per unit traffic, i.e., p, and to CPs for per unit bandwidth, i.e., q, the optimal subsidization and required bandwidth for CP s can be determined by maximizing its utility function:

$$\begin{aligned} \text{OPT-1:} \max_{\{\mathbf{h}^s, \lambda^s\}} \quad & \Phi^s(\mathbf{h}^s, \lambda^s) \\ \text{s.t.} \quad & D_t^s(\mathbf{h}^s, p) \le \lambda^s, \forall t \in \{1, \cdots, L\}, && (5.6) \\ & \mathbf{0}_{1\times L} \preceq \mathbf{h}^s \preceq r^s \mathbf{1}_{1\times L}. && (5.7) \end{aligned}$$

If the price per unit bandwidth charged to CPs by the ISP is free or ignorable, i.e., $q = 0$, any CP $s \in \mathcal{N}$ sets large enough required bandwidth and single subsidization h^s to all time slots that satisfies $h^s = \arg\max_{h^s \ge 0}(r^s - h^s)\big(1 - F_s(p - h^s)\big)$. That makes CP s achieve its *maximum possible revenue*, i.e., $\max_{h^s \ge 0}(r^s - h^s)\big(1 - F_s(p - h^s)\big) \sum_{k=1}^{L} \theta_k^s$. We define the monopoly revenue function as $H_s(h^s) = (r^s - h^s)\big(1 - F_s(p - h^s)\big)$. To simplify our analysis, we make the following assumption.

Assumption 5.2.1 (Unimodal Property)**.** *There exist some monopoly price, denoted as h_M^s that $H_s(\cdot)$ is increasing for all $h^s < h_M^s$ and decreasing for all $h^s > h_M^s$.*

[3] When the demand of one CP is higher than the required bandwidth, the traffic of this CP will be throttled by the ISP if without extra payment.

Assumption 1 is satisfied by a wide range of distribution, e.g., uniform distribution, exponential distribution. Under assumption 1 and ignorable price per unit bandwidth, the subsidizations for all slots are just the monopoly price if $h_M^s \geq 0$ and zero otherwise. We also assume that the monopoly price is non-decreasing with p. When ISP charge higher price per unit traffic, CPs will not reduce their subsidization. In practice, ISPs always charge each CP a non-ignorable price q per unit bandwidth. Under such case, OPT-1 is a non-convex optimization problem, demonstrated in the following example.

Example 5.2.1. *Consider the time horizon with three time slots* $\{1, 2, 3\}$. *The users' valuations for the service provided by one particular CP are uniformly distributed within* $[0, 1]$ *and its users' populations are* $m_1 = 1$ *at slot* $\{1\}$ *and* $m_2 = 1$ *at slot* $\{2\}$. *Each user can delay at most one slot, i.e.,* $K = 1$. *The revenue per unit traffic for this service is* $r^s = 1$ *and the prices charged to end users and this CP by the ISP are* $p = 1$ *and* $q = 1$ *accordingly. The solution with subsidization* $\boldsymbol{h}_1^s = (0.6, 0.5, 0.2)$ *and required bandwidth* $\lambda_1^s = 0.6$ *and the solution with subsidization* $\boldsymbol{h}_2^s = (0.2, 0.5, 0.6)$ *and required bandwidth* $\lambda_1^s = 0.6$ *are both the feasible solutions for OPT-1. However, the average of these two solutions, i.e.,* $\boldsymbol{h}_3^s = (0.4, 0.5, 0.4)$ *and* $\lambda_3^s = 0.6$*, is infeasible since two kinds of users all choose the slot 2 with total demand 1 higher than the required bandwidth 0.6. Thus, the set of feasible prices of OPT-1 is non-convex.*

5.2.3 Utility of the ISP

We use the ISP's revenue to represent its utility, which is originally from two sources: 1) the unit price charged to CPs for the connection services, i.e., q per unit bandwidth, and 2) the unit price charged to end users, i.e., p per unit traffic. With the TDS strategy, the price charged to end users can be partly subsidized by CPs but the total revenue per unit traffic keeps the same. Thus, the payoff[4] (or the profit) of the ISP, denoted by Π, is:

$$\Pi(p, q) = p \sum_{s=1}^{N} \sum_{t=1}^{L} D_t^s(\mathbf{h}^s, p) + q \sum_{s=1}^{N} \lambda^s. \tag{5.8}$$

[4]We ignore the cost of per unit traffic since the main cost is just fixed cost for the ISP and the marginal cost is ignorable.

Then, the ISP decides its optimal prices by the following optimization:

$$\begin{aligned} \text{OPT-2:} \max_{\{p,q\}} \quad & \Pi(p,q) \\ \text{s.t.} \quad & \sum_{s=1}^{N} D_t^s(\mathbf{h}^s, p) \leq \mu, \forall t \in \{1, \cdots, L\}, \\ & p \geq 0, \;\; q \geq 0. \end{aligned}$$

5.2.4 A Two-stage Stackelberg Game

We model the interactive behavior of the ISP and the set of CPs $\mathcal{N}$ as a two-stage Stackelberg game in the system $(\mathcal{N}, \mu, \mathcal{M})$. In particular, we have the following settings:

- *Players*: The ISP and the set of CPs $\mathcal{N}$.
- *Strategies*: The ISP decides the unit prices charged to end users for traffic usage, and to CPs for bandwidth usage, i.e., the ISP's strategy profile is $s_I \in \{(p,q) : p \geq 0, q \geq 0\}$. Each CP decides the price subsidization to its users, e.g., $\mathbf{h}^s$ for CP s and the required bandwidth λ^s.
- *Rules*: The ISP is the first mover who decides its prices charged for bandwidth usage and traffic usage and announces them to CPs and end users. CPs are the second movers and decide their optimal price subsidization and required bandwidth. Each CP makes its own decision independently.
- *Outcome*: The outcome is determined by backward induction. In particular, given any ISP's strategy, each CP chooses its optimal subsidization and required bandwidth to maximize its utility, i.e., OPT-1. Based on this knowledge, the ISP decides its optimal prices charged for bandwidth usage and traffic usage that maximize its utility, i.e., OPT-2.

Note that one assumption about the Stackelberg game is that the ISP is the first-mover and the CPs are the second movers. This is the reality in many countries or regions. ISPs usually know ex ante that CPs would observe their actions, e.g., new pricing strategy,

and make optimal decisions based on their actions. When the ISP fixes its price charged to CPs and to end users, the CPs decide their optimal price subsidization independently. The decision of one particular CP would not be affected by other CPs' decisions. Thus, we can analyze each CP's price subsidization separately. Based on these, we first analyze one particular CP's subsidization in Section 5.3 and then the ISP's optimal choice in section 5.4.

5.3 The CPs' Subsidization

In this section, we analyze the CPs' optimal strategy, i.e., the second stage of the Stackelberg game. We consider the optimal subsidization and required bandwidth of one particular CP, i.e., CP s, under impatient users, i.e., $K = 0$, and patient users, i.e., $K \geq 1$. In both cases, we first analyze the optimal subsidization given the required bandwidth. After that, we study the optimal required bandwidth. Before the analysis of CPs for the two cases, we first consider some general characteristics for OPT-1.

Lemma 5.3.1. *Given the required bandwidth λ^s, the optimal price subsidization h_t^* in slot t is chosen from the following cases: 1) $h_t^* = 0$; 2) $h_t^* = h_M^s$; and 3) $h_t^* = h_{\hat{t}}^s$ such that $D_{\hat{t}}^s(\boldsymbol{h}^*, p) = \lambda^s$ and $h_{\hat{t}}^s \in [0, h_M^s]$.*

Proof. Obviously, $0 \leq h_t^* \leq h_M$. We prove the lemma by contradiction. Consider one time t_0 not satisfying the three conditions. Then, $0 < h_{t_0}^* < h_M$ and $D_{t_0}^s(\mathbf{h}^*, p) < \lambda^s$. Denote the ranking under the optimal subsidizations as $\mathcal{R}$, then $(1-F(p-h_{t_0}))\rho_{t_0}(\mathcal{R}) < \lambda^s$. Denote the set $S = \{t \in \{1, \cdots, L\} | h_t^* = h_{t_0}^*\}$. Let $\delta = h_M - h_{t_0}$ if $h_{t_0}^* \geq h_t^*$ for all $t \in \{1, \cdots, L\}$ and $\delta = \min_{\{t_1 | h_{t_1} > h_{t_0}\}} h_{t_i} - p_{t_0}$ otherwise. Consider the new subsidization:

$$h_t' = \begin{cases} h_t^* + \epsilon & \text{if } t \in S, \\ h_t^* & \text{otherwise,} \end{cases} \tag{5.9}$$

where $0 < \epsilon < \delta$. Then, we have $\mathcal{R}' = \mathcal{R}$ and $h_t' \geq h_t^*$ for all t. We choose small enough ϵ such that $(1 - F(p - h_{t_0}'))\rho_{t_0}(\mathcal{R}) \leq \lambda^s$ for all $t_0 \in S$. Note that it is also true for $t_0 \notin S$. Thus, h_t' is a feasible solution. Then, we consider the revenue obtain for these two subsidizations. Since $H(\cdot)$ is an increasing function when $h \in [0, h_M]$, we have

$H(h'_t) > H(h^*_t)$ for $t \in S$ and $H(h'_t) = H(h^*_t)$ for $t \notin S$. Thus, the revenue obtained by h'_t is higher than h^*_t, contradicting the assumption that h^*_t is the optimal one. □

Lemma 1 demonstrates the three types of value for the optimal price subsidization. Relatively large traffic demand makes no subsidization, i.e., $h^*_t = 0$ while relatively small traffic demand makes highest subsidization, i.e., $h^*_t = h^s_M$. Except these two cases, partial subsidization should be adopted such that the required bandwidth can be fully utilized, i.e., $D^s_t(\mathbf{h}^*, p) = \lambda^s$. However, it is difficult even to know which case the traffic demand belongs to since it is determined by the relative subsidizations under different times. To simplify the analysis, we introduce the definition of preference ranking as follows.

Definition 5.3.1 (Preference Ranking)**.** *Given any price subsidization* ***h****, the preference ranking* $\mathcal{R} = \{R_1, \cdots, R_L\}$ *is the permutation of* $\{1, \cdots, L\}$ *and satisfies:*

$$R_i \begin{cases} < R_j & \text{if } h_i < h_j, \\ > R_j & \text{if } h_i \geq h_j, \end{cases} \tag{5.10}$$

for any $i \in \{1, \cdots, L-1\}$ *and* $j \in \{i+1, \cdots, L\}$.

Definition 5.3.1 states that the users prefer higher subsidization and early time slot. When the preference ranking $\mathcal{R}$ is given, the traffic demand can also be determined, i.e., $D^s_t(\mathbf{h}, p; \mathcal{R}) = (1 - F_s(p - h_t))\rho^s_t(\mathcal{R})$. With this definition, we can analyze the optimal subsidization further by proposition 5.3.1.

Proposition 5.3.1. *Given the required bandwidth* λ^s *and preference ranking* $\mathcal{R}$*, the optimal price subsidization is:*

$$h^*_t = \begin{cases} h^s_M & \text{if } \rho^s_t < \frac{\lambda^s}{1-F_s(p-h^s_M)}, \\ \max\left\{0, p - F_s^{-1}(1 - \lambda^s/\rho^s_t)\right\} & \text{if } \rho^s_t \geq \frac{\lambda^s}{1-F_s(p-h^s_M)}, \end{cases} \tag{5.11}$$

where $F_s^{-1}(\cdot)$ *is the inverse function of* $F_s(\cdot)$.

Proof. According to the lemma 1, the optimal subsidization satisfying one of the three conditions. We first consider the case $\rho^s_t \geq \frac{\lambda^s}{1-F_s(p-h^s_M)}$, that means $h^*_t \neq h^s_M$. If $p - F_s^{-1}(1 - \lambda^s/\rho^s_t) < 0$, then $\rho^s_t(1 - F_s(p - h^*_t)) > \lambda^s$ when $h^*_t \in [0, h_M]$. That means

only condition one can be satisfied, i.e., $h_t^* = 0$. If $p - F_s^{-1}(1 - \lambda^s/\rho_t^s) \geq 0$, then $\rho_t^s(1 - F_s(p - h_t^*)) = \lambda^s$ when $h_t^* = p - F_s^{-1}(1 - \lambda^s/\rho_t^s)$. Note the revenue obtained by $h_t^* = p - F_s^{-1}(1 - \lambda^s/\rho_t^s)$ is higher than the case one. Thus, the optimal subsidization is $h_t^* = \max\{0, p - F_s^{-1}(1 - \lambda^s/\rho_t^s)\}$. We then consider $\rho_t^s < \frac{\lambda^s}{1-F_s(p-h_M^s)}$. Then, we have $\rho_t^s(1 - F_s(p - h_t^*)) < \lambda^s$ when $h_t^* \in [0, h_M]$. That means only case one and two should be satisfied. Note that $h_t^* = h_M^s$ has higher revenue than $h_t^* = 0$, thus the optimal subsidization is $h_t^* = h_M^s$. Hence, we finish the proof. □

5.3.1 Impatient Users ($K = 0$)

When $K = 0$, all users are impatient to delay their traffic usage, i.e., $\rho_t^s = \theta_t^s$. This is usually real for services such as live telecast videos. Then, given the required bandwidth λ^s, the optimal subsidization becomes:

$$h_t^* = \begin{cases} h_M^s & \text{if } \theta_t^s < \frac{\lambda^s}{1-F_s(p-h_M^s)}, \\ \max\left\{0, p - F_s^{-1}(1 - \lambda^s/\theta_t^s\right\} & \text{if } \theta_t^s \geq \frac{\lambda^s}{1-F_s(p-h_M^s)}. \end{cases} \tag{5.12}$$

With the optimal subsidization, we then analyze the optimal required bandwidth. We define $\tilde{\theta}_i = \big(1 - F_s(p - h_M^s)\big)\theta_i^s$ and rearrange the time slots such that $\tilde{\theta}_i < \tilde{\theta}_j$ if $i < j$. We also add one dummy slot $t = 0$ with traffic demand $\tilde{\theta}_0 = \big(1 - F_s(p)\big)\theta_1^s$. Note that when $\lambda^s < \tilde{\theta}_0$, $\mathbf{h} = \mathbf{0}$; when $\lambda^s > \tilde{\theta}_L$, $\mathbf{h} = h_M\mathbf{1}$. Then, the optimal required bandwidth, denoted as λ^*, should be within interval $[\tilde{\theta}_0, \tilde{\theta}_L]$. We divide this interval into several subintervals $[\tilde{\theta}_{l-1}, \tilde{\theta}_l](l \in \{1, \cdots, L\})$. If $\lambda^s \in [\tilde{\theta}_{l-1}, \tilde{\theta}_l]$, then the bandwidth is under utilized for any $t \in \{1, \cdots, l-1\}$ and fully utilized for any $t \in \{l, \cdots, L\}$. Denote the profit[5] during relabeled slots $\{i, \cdots, j\}$ as $\phi_{i,j}$. Then, we can divide the total profit of CP s into two parts: 1) the profit from all under utilized slots, i.e.,

$$\phi_{1,l-1} = (r^s - h_M^s)\sum_{i=1}^{l-1}\tilde{\theta}_i, \tag{5.13}$$

and 2) the profit from all fully utilized slots, i.e.,

$$\phi_{l,L}(\lambda^s) = \lambda^s\sum_{i=l}^{L}\left[r^s - \max\{0, p - F_s^{-1}(1 - \lambda^s/\theta_i^s)\}\right]. \tag{5.14}$$

[5]The profit here refers to the CPs' revenue minus the cost of traffic subsidization.

Thus, we can simplify the OPT-1 as:

$$\begin{aligned} \text{OPT-3: } \max_{\{\lambda^s, l\}} \quad & \phi_{1,l-1} + \phi_{l,L}(\lambda^s) - q\lambda^s \\ \text{s.t.} \quad & \tilde{\theta}_{l-1} \leq \lambda^s \leq \tilde{\theta}_l, l \in \{1, \cdots, L\}. \end{aligned} \tag{5.15}$$

For the problem of determining the optimal required bandwidth, we introduce a new variable l such that we can separate the domain region $[\tilde{\theta}_0, \tilde{\theta}_L]$ into several ones, i.e., $[\tilde{\theta}_{l-1}, \tilde{\theta}_l](l \in \{1, \cdots, L\})$. The above optimization can be solved by first fixing l and obtaining local optimal required bandwidth λ_l^* and then choosing the global optimal λ^* from $\{\lambda_1^*, \cdots, \lambda_L^*\}$. Note that even when we fix l, the above optimization may be non-convex. The following lemma demonstrates the conditions when the problem is convex.

Lemma 5.3.2. *If the monopoly function $H_s(\cdot)$ is concave during $[0, h_M^s]$ and $F_s(\cdot)$ is a concave function, then the above optimization is convex when fixing l.*

Proof. To prove that the simplified OPT-1 is a convex optimization problem, we only need to prove that $\phi_{l,L}(\lambda^s)$ is a concave function during $[\tilde{\theta}_{l-1}, \tilde{\theta}_l]$. Simplify $\phi_{l,L}(\lambda^s)$ and we have

$$\begin{aligned} \phi_{l,L}(\lambda^s) &= \lambda^s \textstyle\sum_{i=l}^{L} \left[r^s - \max\{0, p - F_s^{-1}(1 - \lambda^s/\theta_i^s)\}\right] \\ &= -\textstyle\sum_{i=l}^{L} \max\{-\lambda^s r^s, \lambda^s(p - r^s - F_s^{-1}(1 - \lambda^s/\theta_i^s))\}. \end{aligned} \tag{5.16}$$

Define $G(\lambda) = \lambda^s(p - r^s - F_s^{-1}(1 - \lambda^s/\theta_i^s))$. If $G(\lambda)$ is a convex function, then the objective function is a concave function. Denote $h(\lambda) = p - F_s^{-1}(1 - \lambda^s/\theta_i^s)$, and we have $G(\lambda) = -(r^s - h(\lambda))(1 - F_s(p - h(\lambda)))\theta_i^s = -H(h(\lambda))$. Since $H(h)$ is increasing and concave with h in $[0, h_M]$, then $G(h)$ is decreasing and convex with h in $[0, h_M]$. Then we only need to prove that $h(\lambda)$ is concave. Note that $h'' = \frac{F_s''(F_s^{-1}(1-\lambda^s/\theta_i^s))}{(F_s'(F_s^{-1}(1-\lambda^s/\theta_i^s)))^3(\theta_i^s)^2}$. Since $F_s(\cdot)$ is a concave function, $h(\lambda)$ is also a concave function. Thus, we get the proof. □

The condition of concave monopoly function and $F_s(\cdot)$ makes the revenue $\phi_{l,L}(\lambda)$ from fully utilized slots concave. This makes the optimization problem OPT-3 convex when fixing l. Besides that, the condition also makes the total revenue $\Phi^s(\lambda)$ concave during whole interval $[\tilde{\theta}_0, \tilde{\theta}_L]$. Under the condition of concave monopoly function during

$[0, h_M^s]$ and concave function for $F_s(\cdot)$, we can obtain the optimal required bandwidth in the following theorem.

Theorem 5.3.1. *If there exist l such that $\frac{\partial \phi_{l+1,L}}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon} \leq q \leq \frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon}$ for any small ϵ, then $\lambda^* = \tilde{\theta}_l$; otherwise, there exist one l and $\lambda^* \in [\tilde{\theta}_{l-1}, \tilde{\theta}_l]$ such that $\frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s=\lambda^*} = q$.*

Proof. We first prove that the function $\Phi^s(\lambda)$ is concave during the whole interval $[\tilde{\theta}_0, \tilde{\theta}_L]$. According to lemma 5.3.2, the simplified optimization problem is convex. Then, $\Phi^s(\lambda)$ is concave during each interval $[\tilde{\theta}_{l-1}, \tilde{\theta}_l]$. Thus, we only need to prove that at each end point, $\Phi^s(\lambda)$ is also concave, i.e., $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq \frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon}$. That means $\frac{\partial \phi_{l,L}(\lambda^s)}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq \frac{\partial \phi_{l+1,L}(\lambda^s)}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon}$. It follows that $\frac{\partial \phi_{l,l+1}(\lambda^s)}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq 0$. Note that we have

$$\phi_{l,l+1}(\lambda^s) = -\max\{-\lambda^s r^s, G(\lambda)\}. \tag{5.17}$$

We also have

$$\begin{aligned} \frac{\partial G(\lambda)}{\partial \lambda}|_{\lambda^s=\tilde{\theta}_l-\epsilon} &= -\frac{\partial H(h)}{\partial h}|_{h=h(\tilde{\theta}_l-\epsilon)} \frac{\partial h}{\lambda}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \\ &= -\frac{\partial H(h)}{\partial h}|_{h=h_M-\epsilon_1)} \frac{\partial h}{\lambda}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \leq 0. \end{aligned} \tag{5.18}$$

Thus, $\frac{\partial \phi_{l,l+1}(\lambda^s)}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq 0$ and the function $\Phi^s(\lambda)$ is concave during the whole interval $[\tilde{\theta}_0, \tilde{\theta}_L]$.

If there exist one l and $\lambda^* \in [\tilde{\theta}_{l-1}, \tilde{\theta}_l]$ such that $\frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s=\lambda^*} = q$, this λ^* also makes $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\lambda^*} = 0$ and thus is the global optimal in $[\tilde{\theta}_{l-1}, \tilde{\theta}_l]$. Otherwise, there exist one end point $\tilde{\theta}_l$, that makes $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq 0$ and $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon} \leq 0$ for any small ϵ. This $\tilde{\theta}_l$ is the global optimal since $\Phi^s(\lambda)$ is concave during the whole interval $[\tilde{\theta}_0, \tilde{\theta}_L]$. Note that $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq 0$ and $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon} \leq 0$ is equivalent to $\frac{\partial \phi_{l+1,L}}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon} \leq q \leq \frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon}$. Thus, we get the proof. □

Theorem 5.3.1 demonstrates the optimal required bandwidth is either the end point of some subinterval, e.g., $\tilde{\theta}_l$, that makes $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l-\epsilon} \geq 0$ and $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\tilde{\theta}_l+\epsilon} \leq 0$ for any small ϵ, or the point within some subinterval $[\tilde{\theta}_{l-1}, \tilde{\theta}_l]$ that makes $\frac{\partial \Phi^s}{\partial \lambda^s}|_{\lambda^s=\lambda^*} = 0$, i.e., $\frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s=\lambda^*} = q$.

Beside the analysis of CP's optimal subsidization under impatient users for the discrete time in previous discussions, we also analyze the optimal strategy under continuous time, i.e., $\Delta \to 0$. We define $G(x) = \sum_{t:\tilde{\theta}_t/\Delta<x} \tilde{\theta}_t/\Delta$ and assume that it is continuous and second-order differentiable when $\Delta \to 0$. Note that $\frac{\partial G(x)}{\partial x} \geq 0$ and $\frac{\partial^2 G(x)}{\partial^2 x} \geq 0$, as $\Delta \to 0$. Here, x plays the same role with l in optimization OPT-3. It separates the demands into two parts, i.e., under utilized demands and over utilized demands. If we know the optimal separation x^*, then the optimal $\lambda^* = G'(x^*)$. Define $\underline{\theta} = \min\{\tilde{\theta}_t\}/\Delta$ and $\overline{\theta} = \max\{\tilde{\theta}_t\}/\Delta$. Then, as $\Delta \to 0$, the optimization problem OPT-1 becomes:

$$\begin{aligned} \max_{x} \quad & (r^s - h_M^s)G(x) + \big[(\overline{\theta} - x)(r^s - h_M^s) - q\big]G'(x) \\ \text{s.t.} \quad & \underline{\theta} \leq x \leq \overline{\theta}. \end{aligned} \tag{5.19}$$

Theorem 5.3.2. *As $\Delta \to 0$, the optimal required bandwidth is:*

$$\lambda^* = \begin{cases} G'(\underline{\theta}) & \text{if } \overline{\theta} - \underline{\theta} < \frac{q}{r^s - h_M^s}, \\ G'(\overline{\theta} - \frac{q}{r^s - h_M^s}) & \text{if } \overline{\theta} - \underline{\theta} \geq \frac{q}{r^s - h_M^s}. \end{cases} \tag{5.20}$$

Proof. Considering the first-order derivative of CP's profit, we have:

$$\frac{\partial \Phi^s}{\partial x} = \big[(\overline{\theta} - x)(r^s - h_M^s) - q\big]\frac{\partial^2 G(\lambda^s)}{\partial x}. \tag{5.21}$$

Note that $\frac{\partial^2 G(\lambda^s)}{\partial x} \geq 0$. Let $\frac{\partial \Phi^s}{\partial x} = 0$, then we have $x = \overline{\theta} - q/(r^s - h_M^s)$. If $\overline{\theta} - q/(r^s - h_M^s) < \underline{\theta}$, then we have $x^* = \underline{\theta}$. Otherwise, $x^* = \overline{\theta} - q/(r^s - h_M^s)$. Since $\lambda^* = G'(x^*)$, we get the proof. □

The theorem states that CP s keeps the minimum required bandwidth $G'(\underline{\theta})$ when the gap between peak demand and valley demand is lower than some threshold, i.e., $\frac{q}{r^s - h_M^s}$. Under such required bandwidth, the bandwidth is fully utilized at each time. When the gap is larger than the threshold, the required bandwidth increases with the maximum demand. In addition, higher price charged to bandwidth, i.e., q, results in the larger threshold and thus lower required bandwidth.

5.3.2 Patient Users ($K \geq 1$)

When we consider the case of patient users, i.e., $K \geq 1$, the price subsidization may not be determined easily even given the required bandwidth. This problem is non-convex, as we demonstrated in the example of previous section. In this section, we design a dynamic algorithm to obtain the optimal price subsidization with polynomial time complexity. Before we begin the analysis of our dynamic algorithm, we first introduce the definition of peak slot.

Definition 5.3.2 (peak slot)**.** *A time slot t is a peak slot in period $\{i, \cdots, j\}$ if $R_t > R_s$ for any $s \in \{i, \cdots, j\} \setminus \{t\}$.*

A peak slot is more preferred by users than other slots in period $\{i, \cdots, j\}$. This peak slot helps us design the dynamic algorithm that separates the original problem into subproblems.

Consider the optimal subsidization $\mathbf{h}^*$ with preference ranking $\mathcal{R}^*$. Suppose slot t is the peak slot in period $\{1, \cdots, L\}$ under preference ranking $\mathcal{R}^*$ and thus $R_t^* = L$. Then, the users with maximal waiting time K arriving at any slot during $\big\{\max\{t-K, 1\}, \cdots, t-1\big\}$ delay their service to slot t and thus the potential demand for slot t is $\rho_t^s(\mathcal{R}^*) = \sum_{K=0}^{L-1} \sum_{l:t \in S_l^K} g(K)\theta_l^s$. If we know the peak slot t and h_t^*, we can separate the original problem of maximizing the profit obtained from period $\{1, \cdots, L\}$ into two subproblems: maximizing the profit obtained from period $\{1, \cdots, t-1\}$ and the profit obtained from period $\{t+1, \cdots, T\}$. Although these two subproblems are independent, they are not strictly the subproblems of original problem if not defined carefully. This is because the population of users with maximal waiting time K arriving at any slot $l \in \big\{\max\{t-K, 1\}, \cdots, t-1\big\}$ is zero, instead of $g(K)\theta_l^s$ for the first subproblem.

In our dynamic algorithm, we consider the subproblem of maximizing the profit obtained from $\{i, \cdots, j\}$, denoted as $W(i, j, \bar{h})$, with traffic generated by users with maximal waiting time K arriving at slot l: 1) $\bar{\theta}_l^K = 0$ if $j \neq T$ and $\max\{i, j-K+1\} \leq l \leq j$, and 2) $\bar{\theta}_l^K = g(K)\theta_l^s$ otherwise. Note that the subsidization during period $\{i, \cdots, j\}$

should also be weakly lower than $\bar{h}$. We also extend the definition of waiting period to some subperiod $\{i, \cdots, j\}$, instead of whole $\{1, \cdots, L\}$, denoted as $S_t^K(i,j) \triangleq \big\{t, \cdots, \min\{t+K, j\}\big\}$. Then, we have:

$$\begin{aligned} W(i,j,\bar{h}) = \quad & \max_{\{\mathbf{h}\}} \sum_{t=i}^{j} (r^s - h_t)\bar{D}_t^s(\mathbf{h}, p) \\ \text{s.t.} \quad & \bar{D}_t^s(\mathbf{h}, p) \leq \lambda^s, \forall t \in \{1, \cdots, L\}, \qquad (5.22) \\ & \mathbf{0}_{1\times L} \preceq \mathbf{h} \preceq v^s \mathbf{1}_{1\times L}, \qquad (5.23) \\ & h_t \leq \bar{h}, \forall t \in \{i, \cdots, j\}. \qquad (5.24) \end{aligned}$$

where $\bar{D}_t^s = \big(1 - F_s(p - h_t)\big) \sum_{K=0}^{L-1} \sum_{l:t\in S_l^K(i,j)} \bar{\theta}_l^K \mathcal{I}\{t = \varpi_l^s\}$. Note that given the required bandwidth λ^s, the optimization problem of OPT-1 is equal to $W(1, T, h_M^s)$ if we ignore the bandwidth cost. Note that $W(i,i,\bar{h}) = g(0)\theta_i^s \max_{h\in[0,\bar{h}]} H_s(h)$. Then, we have:

$$\begin{aligned} W(i,j,\bar{h}) = \max_{k\in\{i,\cdots,j\}} \Big\{ & \max_{h\in[0,\bar{h}]} \big\{ W(i,k-1,h) \\ & + \gamma_{i,j}^k(h) + W(k+1,j,h) \big\} \Big\}, \qquad (5.25) \end{aligned}$$

where $\gamma_{i,j}^k(h)$ is

$$\gamma_{i,j}^k(h) = \begin{cases} \min\{\sum_{K=0}^{L-1} \sum_{l:t\in S_l^K(i,k)} \bar{\theta}_l^K H_s(h), \lambda^s\} \\ \text{if } h \leq \max\big\{0, p - F^{-1}(1 - \lambda^s/\bar{\rho}(i,j,k))\big\}, \\ -\infty \quad \text{otherwise,} \end{cases} \qquad (5.26)$$

and $\bar{\rho}(i,j,k) = \sum_{K=0}^{L-1} \sum_{l:t\in S_l^K(i,k)} \bar{\theta}_l^K$.

To understand why the recursion in Eq. 5.25 holds, we consider the optimal subsidization of CP s and assume that $k \in \{i, \cdots, j\}$ is the peak slot with corresponding subsidization $h_k \leq \bar{h}$. Then, the users with maximal waiting time K arriving at any slot during $\big\{\max\{i, k-K\}, \cdots, k\big\}$ delay their services to slot k and thus the demand at slot k is $\sum_{K=0}^{L-1} \sum_{l:t\in S_l^K(i,k)} \bar{\theta}_l^K \big(1 - F_s(p - h_k)\big)$ constrained by the required bandwidth. Thus, we can obtain the profit at slot k as $\gamma_{i,j}^k(h_k)$. Due to the users of delaying services to slot k, the traffic generated by the users with maximal waiting time K arriving at any slot during $\{i, \cdots, k-1\}$ becomes: 1) $\bar{\theta}_l^K = 0$ if $\max\{i, k-K\} \leq l \leq k-1$, and 2) $\bar{\theta}_l^K = g(K)\theta_l^s$ otherwise. The profit obtained from period $\{i, \cdots, k-1\}$ with the traffic

demand $\bar{\theta}_l^K$ and prices weakly smaller than h_k is exactly $W(i, k-1, h_k)$ according to the definition. Then, we consider the period during $\{k+1, \cdots, j\}$. Note that in original period $\{i, \cdots, j\}$, we already have: 1) $\bar{\theta}_l^K = 0$ if $\max\{i, j-K+1\} \leq l \leq j$, and 2) $\bar{\theta}_l^K = g(K)\theta_l^s$ otherwise. The traffic demand during period $\{k+1, \cdots, j\}$ still holds unchanged after separation. Then, the profit obtained from period $\{k+1, \cdots, j\}$ with prices weakly smaller than h_k can be also given as $W(k+1, j, h_k)$. Thus, we have $W(i, j, \bar{h}) = W(i, k, h_k) + \gamma_{i,j}^k(h_k) + W(k, j, h_k)$.

Note that $W(i, j, h)$ is non-decreasing with h. This is also true for $\gamma_{i,j}^k(h)$ if $h \leq \max\left\{0, p - F^{-1}\big(1 - \lambda^s/\bar{\rho}(i, j, k)\big)\right\}$. Then, the optimal subsidization during $[0, \bar{h}]$ in Eq. 5.25 is:

$$\bar{h}_{i,j}^k = \min\left\{\overline{h}, \max\left\{0, p - F^{-1}\big(1 - \lambda^s/\bar{\rho}(i, j, k)\big)\right\}\right\}. \tag{5.27}$$

Thus, we can simplify the Eq. 5.25 as:

$$W(i, j, \overline{h}) = \max_{k \in \{i, \cdots, j\}} \left\{W(i, k, \bar{h}_{i,j}^k) + \gamma_{i,j}^k(\bar{h}_{i,j}^k) + W(k, j, \bar{h}_{i,j}^k)\right\}. \tag{5.28}$$

Lemma 5.3.3. *The solution space of subsidization h_t for any $t \in \{1, \cdots, L\}$ is in $\Gamma = \bigcup_{i,j,k}\{h_{i,j}^k\} \cup \{0\} \cup \{h_M\}$ with cardinality $O(L^3)$, where $h_{i,j}^k = \max\left\{0, p - F^{-1}(1 - \lambda^s/\bar{\rho}(i, j, k))\right\}$.*

Proof. According to equation 5.27, the solution space of subsidization can only be in the set Γ. Note that $i \in \{1, \cdots, L-1\}$, $j \in i, \cdots, L$ and $k \in \{i, \cdots, j\}$. Thus, we have the $|\Gamma|$ should be at most L^3. Thus, we get the proof. □

This lemma demonstrates that without solving the OPT-1 with the fixed required bandwidth, we can limit the potential solution from $[0, h_M]$ to the polynomial set Γ. With this polynomial solution space, we design the dynamic algorithm $DynamicAlg()$ based on the above analysis.

Theorem 5.3.3. *The optimal subsidization for OPT-1 given the required bandwidth λ^s can be obtained in time complexity of $O(L^6)$.*

Algorithm 2 DynamicAlg(i,j,$\bar{h}$)

1: **if** $W(i,j,\bar{h}) > -\infty$ **then**
2: **return** $W(i,j,\bar{h})$
3: **end if**
4: **if** $i = j$ **then**
5: $W(i,j,\bar{h}) = g(0)\theta_i^s H_s(\bar{h})$
6: **else**
7: **for** $k = i$ to j **do**
8: $h_k = \min\left\{\bar{h}, \max\{0, p - F^{-1}(1-\lambda^s/\bar{\rho}(i,j,k))\}\right\}$
9: $\gamma_{i,j}^k = (r^s - h_k)\min\{(1-F_s(p-h_k))\bar{\rho}(i,j,k), \lambda^s\}$
10: q =DynamicAlg$(i,k-1,h_k) + \gamma_{i,j}^k$+DynamicAlg$(k+1,j,h_k)$
11: **if** $W(i,j,\bar{h}) < q$ **then**
12: $W(i,j,\bar{h}) = q$
13: **end if**
14: **end for**
15: **end if**
16: **return** $W(i,j,\bar{h})$

Proof. Given the required bandwidth λ^s, the optimal subsidization for CP s is equivalent to optimal solution for $W(1,T,h_M)$. We first analyze the time complexity of calculating the optimal value for each $W(i,j,h)$. After that we analyze the time complexity of optimal subsidization from each $W(i,j,h)$.

To calculate the time complexity of each $W(i,j,h)$, we need to calculate the value $\rho(i,j,k)$. There are at most $O(L^3)$ values for $\rho(i,j,k)$ and each one takes at most $O(L^2)$. Thus, all values for $\rho(i,j,k)$ can be computed by $O(L^5)$ times. According to lemma 5.3.3, the solution space for OPT-1 has cardinality $O(L^3)$. Thus, there exists $O(L^5)$ number of values for $W(i,j,h)$. For the calculation of each $W(i,j,h)$, it takes $O(L)$ time (line 7-13), since $\rho(i,j,k)$ have been determined first. Thus, the total time complexity for all $W(i,j,h)$ is $O(L^6)$.

We then analyze the time complexity of obtaining optimal subsidization from each $W(i,j,h)$. We can obtain each optimal subsidization h_k from each recursion Eq. 5.28. Each optimal subsidization h_k can be obtained with time complexity of $O(L)$ and there exist $O(L)$ subsidizations, i.e., $h_k(k \in \{1,\cdots,L\})$. Thus, the total time complexity is $O(L^5 + L^6 + L^2) = O(L^6)$. □

The above theorem demonstrates that we can obtain the optimal subsidization for the non-convex optimization OPT-1 given the required bandwidth λ^s via polynomial time

$O(L^6)$. With this dynamic algorithm, we then search the optimal required bandwidth by line search algorithm $LSearchAlg()$.

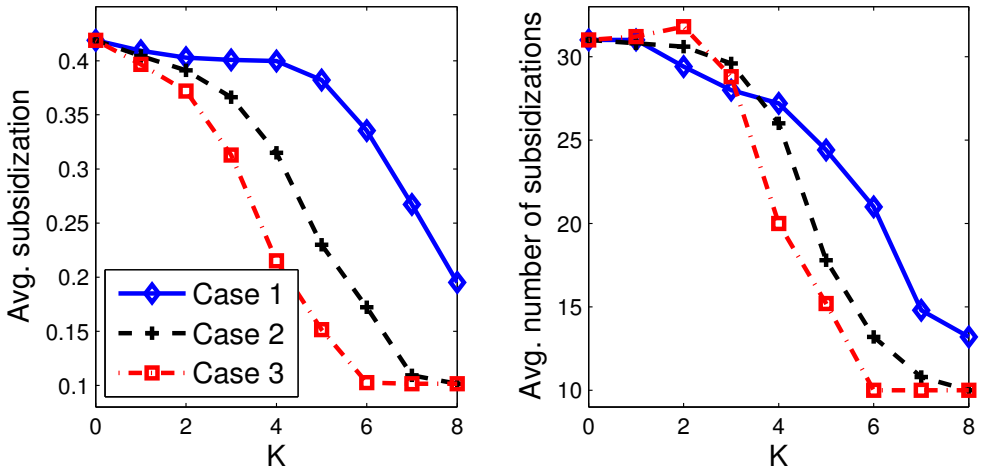
Algorithm 3 LSearchAlg()

1: Initialize $\lambda[0]$;
2: Calculate $\{W(i,j,h;\lambda[0])\}$ by calling $DynamicAlg()$;
3: Calculate optimal ranking $\mathcal{R}'_{[0]}$ induced by $\{W(i,j,h;\lambda[0])\}$;
4: $t \leftarrow 0$;
5: **do**
6: Calculate $\{\rho_k(\mathcal{R}_{[t]})\}$;
7: Given $\{\rho_k(\mathcal{R}_{[t]})\}$, calculate optimal $\lambda'[t]$ by solving OPT-3;
8: $\lambda[t+1] \leftarrow \lambda[t] + g[t](\lambda'[t] - \lambda[t])$;
9: Calculate $\{W(i,j,h;\lambda[t+1])\}$ by calling $DynamicAlg()$;
10: Calculate optimal ranking $\mathcal{R}'_{[t]}$ induced by $\{W(i,j,h;\lambda[t+1])\}$;
11: $\mathcal{R}_{[t+1]} \leftarrow \mathcal{R}'_{[t]}$;
12: $t \leftarrow t+1$;
13: **until** $t < T$ or $\mathcal{R}_{[t]} == \mathcal{R}_{[t-1]}$
14: **return** $\lambda[t]$.

The algorithm starts with initializing required bandwidth $\lambda[0]$ and calculating optimal ranking $\mathcal{R}'_{[0]}$ (line 1 to 4). In each step t, after obtaining the potential demand $\{\rho_k(\mathcal{R}_{[t]})\}$, the algorithm calculates the optimal required bandwidth $\lambda'[t]$ by solving OPT-3 (line 6 to 7). Then it updates the required bandwidth $\lambda[t+1]$ based on $\lambda[t]$ and the induced required bandwidth $\lambda'[t]$ (line 8). The step size parameter $g[t]$ can be time-static or decreasing in t. After that, the algorithm updates the optimal ranking $\mathcal{R}_{[t+1]}$ (line 9 to 11). The algorithm terminates when the round time approaches the maximal number, i.e., T, or the outcome is stable (line 13).

5.3.3 Numerical Illustrations

To understand the CPs' subsidization more intuitively, we illustrate a numerical example as follows. We divide users into impatient and patient users with population ratio $m_1 : m_2$. This population ratio is characterized by three cases: 1) $m_1 : m_2 = 3 : 1$, and 2) $m_1 : m_2 = 1 : 1$, and 3) $m_1 : m_2 = 1 : 3$. Case 1 captures the scenario of dominated impatient populations and case 3 captures the scenario of dominated patient populations. We consider the homogeneous maximal waiting time for patient users and set $K = 4$ by default. The per unit valuation of users for the traffic follows uniform distribution $U([0,2])$. We analyze the traffic of two days with one hour as a slot. The potential traffic



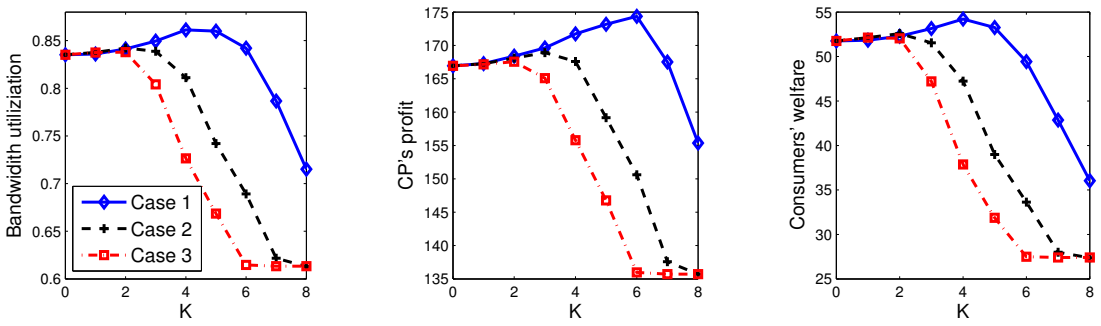
(a) (b)

Figure 5.2: Subsidizations with various K

pattern of each day is set as: $\theta_t = t + \delta$ if $t \leq 12$ and $\theta_t = 25 - t + \delta$ if $13 \leq t \leq 24$, where δ follows $G(0, 1)$. The revenue per unit traffic for this service is $r = 2$ and the prices charged to users and this CP by the ISP are $p = 1.5$ and $q = 10$ accordingly. We set the required bandwidth to be $\lambda = 3$ unless otherwise specified.

Effects of Strategic Behaviors

We first illustrate the average subsidization and the average number of subsidizations with various K from $\{0, 1, \cdots, 8\}$, shown in Fig. 5.2. Figure 5.2(a) demonstrates the effects of maximal waiting time on the average subsidization. As users have larger K, the average subsidization from the CP decreases. Moreover, when the population ratio of patient users is higher, the average subsidization decreases faster. Similar things happen to the average number of subsidizations, shown in Fig. 5.2(b). The intuition is that when the CP provides a slightly higher subsidization in some slot, patient users delay their traffic to this slot, resulting in low incentives to subsidize such high price. When K is larger, the behavior of strategic users becomes more homogeneous, i.e., more patient users delaying their traffic simultaneously. We call this phenomenon as *homogeneous strategic behavior*. To maintain such few number of subsidizations and satisfy the bandwidth limitation, the CP needs to decrease the subsidization in some slots.



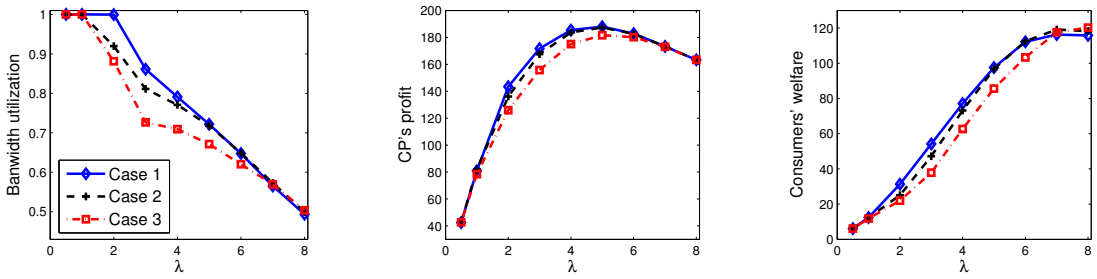
(a) (b) (c)

Figure 5.3: Effects of maximal waiting time K

Then, we illustrate how strategic behavior of users affects the bandwidth utilization, CP's profit and consumers' welfare in Fig. 5.3. Here, we define the consumers' welfare as the difference between the value of consumers for the service and their payment. Intuitively, when the consumers are more patient, the bandwidth should be better utilized. After all, the CP provides high subsidizations in the slots with low potential demand, nicely attracting more patient users from the slots with high potential demand, and thus, utilizes the bandwidth in an efficient way. Indeed, this conjecture can be verified to some extend when the maximal waiting time is not high, e.g., $K \leq 4$ for case 1 and $K \leq 2$ for case 2 and 3, shown in Fig. 5.3(a). However, we always observe the opposite effect, i.e., more wasted bandwidth. As users become more patient (larger K or smaller $m_1 : m_2$ or both), the CP is forced to provide fewer and lower subsidizations, and thus resulting in inefficient usage of bandwidth. This phenomenon hurts both the CP and the consumers (Fig. 5.3(b) and Fig. 5.3(c)). For instance, when $m_1 : m_2 = 1 : 3$ and $K = 6$, the CP's profit and the consumers' welfare reduce 19% and 47% accordingly, compared with the case of $K = 0$. In later simulations, we demonstrate how the strategy of staggered subsidizations avoids this phenomenon.

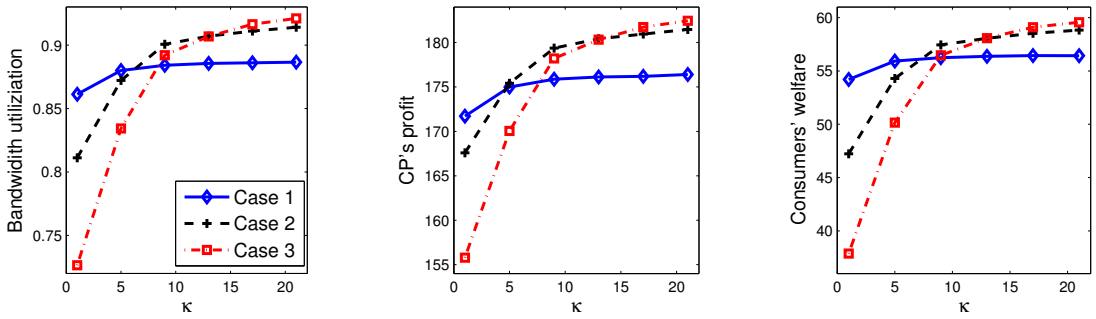
Effects of Required Bandwidth

We demonstrate the effects of required bandwidth on the bandwidth utilization, CP's profit and consumers' welfare in Fig. 5.4. Figure 5.4(a) shows that when the required bandwidth is small enough, e.g., $\lambda = 1$, the bandwidth can be always fully utilized. As



(a) (b) (c)

Figure 5.4: Effects of required bandwidth λ



(a) (b) (c)

Figure 5.5: Effects of number of groups κ

the increase of required bandwidth, the bandwidth utilization decreases. However, the traffic demand always increases with the required bandwidth. Figure 5.4(b) illustrates that the CP's profit increases with the required bandwidth, when the required bandwidth is small, e.g., $\lambda = 1$, due to the increase of demand. However, as the required bandwidth is large, e.g., $\lambda = 5$, the CP's profit decreases with the required bandwidth, due to less benefit of traffic demand but higher bandwidth cost. The optimal required bandwidth for this CP is around $\lambda = 4.7$. Figure 5.4(c) shows that the consumers' welfare always increases with the required bandwidth.

Effects of Staggered Subsidizations

As we discuss previously, homogeneous strategy behavior leads to inefficient bandwidth utilization when consumers are patient. To avoid this problem, we adopt staggered subsidizations, i.e., separating consumers into multiple groups with different subsidizations. More specifically, we divide consumers into κ groups and decide the sub-

sidization sequently. The subsidization of one group is set according to the algorithm $DynamicAlg()$, but with bandwidth limitation left from previous groups, instead of the required bandwidth.

We demonstrate the effects of various numbers of groups on the bandwidth utilization, CP's profit and consumers' welfare in Fig. 5.5. Figure 5.5(a) shows that the bandwidth utilization increases significantly with the number of groups. In addition, more population ratio of patient consumers means less wasted bandwidth. For instance, when $m_1 : m_2 = 3 : 1$ and $\kappa = 21$, the bandwidth utilization improves from 86% to 89%, compared with the case of no staggered subsidization; while the improvement is from 73% to 92% for case 3, i.e., $m_1 : m_2 = 1 : 3$. Note that this increase is not obvious when the number of groups is large, e.g, $\kappa = 5$ for case 1. Thus, we only need to decide subsidizations for small number of groups so as to avoid the homogeneous strategy behavior for patient consumers. Moreover, the efficient usage of bandwidth due to staggered subsidizations benefits both consumers and the CP, shown in Fig. 5.5(b) and Fig. 5.5(c).

Remarks: When we analyze the CPs with strategic users, the maximal waiting time effects CPs' optimal subsidization significantly. Higher maximal waiting time indicates lower number of subsidizations and also lower average subsidization. In addition, higher maximal waiting time does not always improve the bandwidth utilization and increase CPs' profit. This only happens when the users are slightly patient, e.g., $K = 4$ and $m_1 : m_2 = 3 : 1$. Highly patient users lead to serious homogeneous strategic behavior and thus increase the waste of bandwidth usage. This homogeneous strategic behavior can be avoided when CPs stagger the subsidizations for different groups. Higher number of groups indicates much higher bandwidth utilization and thus higher CPs' profit and consumers' welfare.

5.4 Monopolistic ISP's Strategy

In previous section, we analyze the second stage of the Stackelberg game, i.e., CPs' optimal subsidization and required bandwidth. In this section, we consider the first stage

of the Stackelberg game, i.e., ISP's strategy. We analyze the effects of ISP's strategy under two cases: homogeneous CPs and heterogeneous CPs.

5.4.1 Homogeneous CPs

In this subsection, we analyze the effects of ISP's strategy when CPs are homogeneous, i.e., each CP has the same revenue per unit ($r^s = r$) and the same traffic pattern ($\theta_t^s = \theta_t$). The homogeneous CPs provide the same subsidization and required bandwidth, and thus simplify the analysis. When the CPs are homogeneous, we can simplify the ISP's optimization problem OPT-2 as:

$$\text{OPT-4: } \max_{\{p,q\}} \quad p\sum_{t=1}^{L} D_t(\mathbf{h}, p) + q\lambda(p,q)$$

$$\text{s.t.} \quad \lambda(p,q) \leq \mu/N, \tag{5.29}$$

$$p \geq 0, \;\; q \geq 0. \tag{5.30}$$

With the above simplification, we demonstrate the effects of expending the capacity in the following theorem.

Theorem 5.4.1 (Effects of capacity)**.** *Denote the ISP's optimal strategy in the system* $(\mathcal{N}, \mu, \mathcal{M})$ *and* $(\mathcal{N}, \mu', \mathcal{M})$ *as* (p, q) *and* (p', q') *respectively. If* $\mu' \geq \mu$*, then* $\Pi' \geq \Pi$*. Moreover, if* $\mathcal{R}' = \mathcal{R}$*, then* $q' \leq q$ *or* $p' \leq p$ *or both.*

Proof. Since $\mu' \geq \mu$, (p, q) should also be the feasible solution. Thus, $\Pi' \geq \Pi$. Moreover, if the capacity is the constraint, then $\Pi' > \Pi$.

We assume that $(p', q') \succ (p, q)$. Since the rankings are the same, then we have $\lambda(p', q') \leq \lambda(p, q) \leq \mu/N$ (Theorem 5.4.2). That means (p', q') is also the optimal solution for the system $(\mathcal{N}, \mu, \mathcal{M})$. Note that (p, q) is also the optimal one. It follows $(p', q') = (p, q)$, which is controversial to the assumption. Thus, we finish the proof. □

Theorem 5.4.1 demonstrates that when the capacity is extended, the ISP decreases the price per unit bandwidth charged to CPs or the price per unit traffic charged to users or

both so as to obtain higher profit. In addition, CPs and the ISP can all benefit from the capacity extension. Then, we consider the effects of ISP's strategy on CPs.

Theorem 5.4.2 (Effects of ISP's strategy)**.** *Given any two strategies satisfying* $(p', q') \succeq (p, q)$*, we have* $\Phi' \leq \Phi$*. In addition, if* $\mathcal{R}' = \mathcal{R}$*, then we have* $\lambda' \leq \lambda$ *and* $h'_t \leq h_t$ *for any* $t \in \{1, \cdots, L\}$.

Proof. Denote the optimal ranking under ISP's strategy (p', q') as $\mathcal{R}'$. Then, we consider the CPs' profit, denoted as Φ^+, under ISP's strategy (p, q) with the same ranking $\mathcal{R}'$. If $\Phi^+ \geq \Phi'$, then $\Phi \geq \Phi'$ since $\Phi \geq \Phi^+$. Given the ranking $\mathcal{R}'$, the optimal required bandwidth is equivalent to OPT-3 with potential demand $\rho_i(\mathcal{R}')$. Since Φ^s is concave function, then the first-order derivative is a decreasing function, i.e., $\frac{\partial \Phi}{\partial \lambda}|_{\lambda \in [\theta_{l-1}, \theta_l]} = \frac{\partial \phi_{l,L}}{\partial \lambda} - q$. The zero point can be either 1) some end point $\lambda^* = \tilde{\theta}_l$ that satisfies $\frac{\partial \phi_{l+1,L}}{\partial \lambda^s}|_{\lambda^s = \tilde{\theta}_l + \epsilon} \leq q \leq \frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s = \tilde{\theta}_l - \epsilon}$ for any small ϵ or 2) some inner point that satisfies $\frac{\partial \phi_{l,L}}{\partial \lambda^s}|_{\lambda^s = \lambda^*} = q$. When we increase the price per bandwidth q, the function $\frac{\partial \Phi}{\partial \lambda}$ shifts left and thus the zero point for the function $\frac{\partial \Phi}{\partial \lambda}$ also shifts left. Thus, if $q' \geq q$, then $\lambda' \leq \lambda$. Since h_t^* increases with λ, then $h' \leq h$.

Then, we consider the case of $p' \geq p$. We want to prove that for any λ, $\frac{\partial \Phi'}{\partial \lambda} \leq \frac{\partial \Phi}{\partial \lambda}$. Note that

$$\phi_{l,L}(\lambda^s) = -\max\{-L\lambda^s r^s, \sum_{i=l}^{L} G_i(\lambda)\}, \tag{5.31}$$

and

$$\frac{\partial G_i(\lambda)}{\partial \lambda} = -\frac{\partial H(h_i)}{\partial h_i}\frac{\partial h_i}{\partial \lambda} \leq 0, \tag{5.32}$$

where $G_i(\lambda) = -H(h_i(\lambda))\theta_i^s$ and $h_i = p - F_s^{-1}(1 - \lambda^s/\theta_i^s)$.

Since $H(\cdot)$ is concave in $[0, h_M]$, then $\frac{\partial^2 H(h)}{\partial^2 h} = -(2F'(p-h)+(r-h)F''(p-h)) \leq 0$. In addition, $\frac{\partial h_M(p)}{\partial p} = \frac{F'(p-h)+(r-h)F''(p-h)}{2F'(p-h)+(r-h)F''(p-h)} \leq 0$. Then, we have $F'(p-h)+(r-h)F''(p-h) \leq 0$. Note that $\frac{\partial G_i}{\partial h \partial p} = -\frac{\partial h_i}{\partial \lambda}\frac{\partial H}{\partial h_i \partial p} = -\frac{\partial h_i}{\partial \lambda}(F'(p-h)+(r-h)F''(p-h)) \geq 0$. Then, we have $\frac{\partial \Phi'}{\partial \lambda} \leq \frac{\partial \Phi}{\partial \lambda}$. It indicates that the zero point shifts left as the increase of p. Thus, $\lambda' \leq \lambda$. According to Eq. 5.12, the optimal h_t^* increases with p.

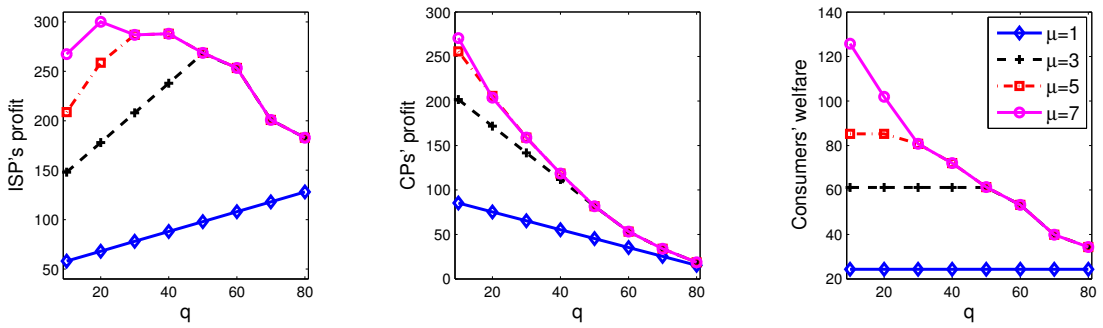
Note that for both cases, we have $\frac{\partial \Phi'}{\partial \lambda} \leq \frac{\partial \Phi}{\partial \lambda}$. Thus, we have $\Phi' = \int_{s=0}^{\lambda'} \frac{\partial \Phi'}{\partial s} ds \leq \int_{s=0}^{\lambda'} \frac{\partial \Phi}{\partial s} ds = \Phi$. Also note that $\frac{\partial \Phi'}{\partial \lambda}$ and $\frac{\partial \Phi}{\partial \lambda}$ may not exist in some points. Since this non-differentiable point set is finite, they cannot affect the result of this proof. □

Theorem 5.4.2 illustrates that when the ISP increases its price per unit traffic or price per unit bandwidth or both, the CPs' profit decreases. Moreover, when the rankings of subsidization are the same, CPs decrease their subsidization and their required bandwidth.

To understand the effects of ISP's strategy intuitively, we illustrate an example with the same traffic pattern, users' valuation, CPs' revenue as in the subsection IV-C. We set the population ratio of impatient users over patient users as $m_1 : m_2 = 1 : 3$ and the homogeneous maximal delay time as $K = 4$. The prices charged to users and CPs by the ISP are $p = 1$ and $q = 30$ accordingly unless otherwise specified.

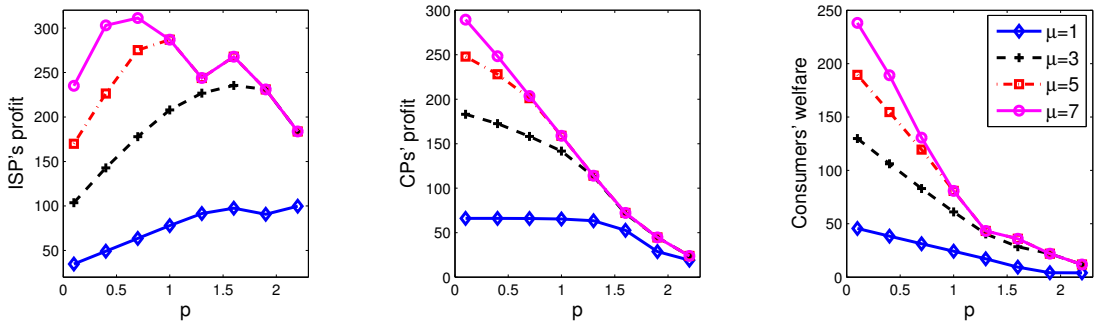
We first demonstrate the effects of various prices per unit bandwidth on the ISP's, CPs' profit and social welfare in Fig. 5.6. Figure 5.6(a) indicates that the ISP's profit increases with q when q is small, e.g., $q = 10$, and decreases with q when q is large, e.g., $q = 50$. Under suitable q, e.g., $q = 20$ under $\mu = 7$, the ISP's profit reaches the highest. In addition, larger capacity also indicates the higher ISP's profit. The small capacity, e.g. $\mu = 3$, may constrain the decision for the optimal price per unit bandwidth. Figure 5.6(b) shows that as the increase of q, the CPs' profit decreases significantly. For instance, when q increases from 10 to 40, the CPs' profit decreases from 271 to 118, nearly 56%. Larger capacity indicates the higher CPs' profit. Moreover, when the capacity is larger, the CPs' profit also decreases faster as q increases. Figure 5.6(c) demonstrates that as the increase of q, the consumers' welfare decreases. Note that when q is small, e.g., $q = 20$ under $\mu = 3$, the consumers' welfare keeps the same as the increase of q since the demand is constrained by the capacity. Moreover, larger capacity indicates higher consumers' welfare.

We then demonstrate the effects of various prices per unit traffic on the ISP's, CPs' profit and consumers' welfare in Fig. 5.7. Figure 5.7(a) states that suitable price per unit traffic charged to users, e.g., $p = 0.7$, maximizes the ISP's profit. Larger capacity indicates



(a) (b) (c)

Figure 5.6: Effects of prices per unit bandwidth



(a) (b) (c)

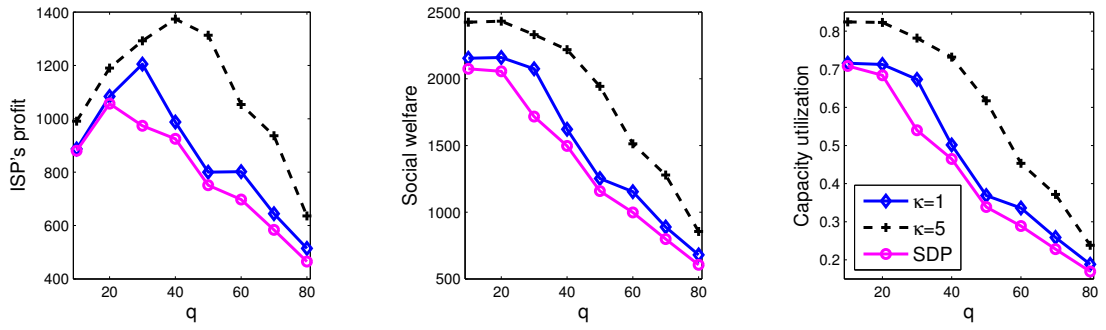
Figure 5.7: Effects of prices per unit traffic

higher ISP's profit. Moreover, the ISP's profit increases with p when the capacity becomes the constraint, e.g., $\mu = 1$. Figure 5.7(b) illustrates that CPs' profit decreases as the increase of p. Note that the CPs' profit decreases much slower under smaller p than larger p as the increase of p. For instance, when $\mu = 1$, the CPs' profit almost keeps the same as p increases from 0.1 to 1.3 but decreases 55% as p increases from 1.3 to 1.9. Figure 5.7(c) indicates that consumers' welfare decreases significantly as the increase of p. When the capacity is larger, the consumers' welfare decreases faster.

5.4.2 Heterogeneous CPs

In this subsection, we analyze the effects of ISP's strategy on the capacity utilization[6], ISP's profit, social welfare under heterogeneous CPs via simulations. We define the so-

[6]Capacity utilization here refers to the total utliziation of ISPs capacity, different from bandwidth utilization which refers to the utilization of CPs required bandwidth.



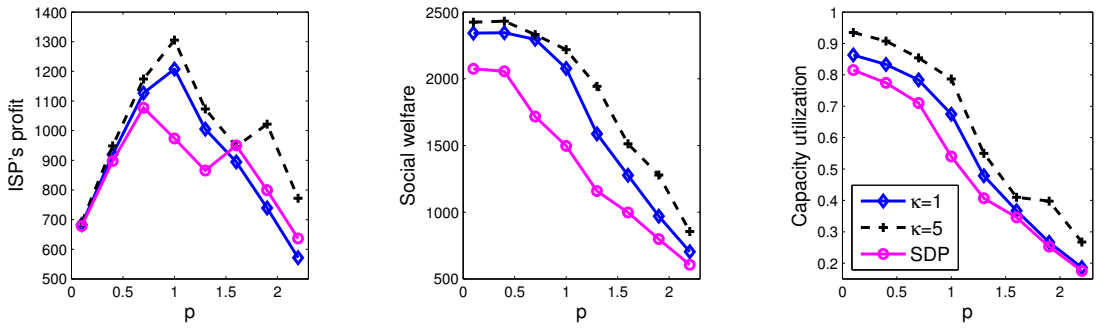
(a) (b) (c)

Figure 5.8: Effects of q with heterogeneous CPs

cial welfare as the summation of CPs' profit, ISP's profit and consumers' welfare. We consider five heterogeneous CPs in the simulations. Each CP s has a random phase displacement ξ^s from $G(0, 2^2)$ and random amplitude of traffic pattern γ^s from $U[0.5, 1.5]$, i.e., $\theta_t^s = \gamma^s(t + \xi^s) + \delta$. The revenue per unit traffic for service $s \in \{1, \cdots, 5\}$ is set to be $r^s = 0.5 \times s$. The rest settings are the same with subsection 5.4.1.

We demonstrate the effects of prices per unit bandwidth on the ISP's profit, social welfare and capacity utilization in Fig. 5.8. Figure 5.8(a) states that when q is small, the ISP's profit for TDS increases with q; while the ISP's profit for TDS decreases with q when q is large. In addition, the ISP's profit for TDS is always higher than that for SDP. Higher number of groups indicates much higher ISP's profit. For instance, when $q = 40$, the ISP's profit for TDS with $\kappa = 1$ is 7% larger than that for SDP while 48% for TDS with $\kappa = 5$ larger than that for SDP. Figure 5.8(b) shows that as the increase of q, the social welfare decreases. Figure 5.8(c) demonstrates that the capacity utilization has the similar trends with the social welfare. As the increase of q, the capacity utilization decreases. The intuition is that higher q makes the required bandwidth for each CP smaller and thus results in more waste of unused capacity.

We then illustrate the effects of prices per unit traffic on ISP's profit, social welfare and capacity utilization in Fig. 5.9. Figure 5.9(a) demonstrates that either large or small p makes the ISP's profit small. Generally, the ISP's profit under TDS is larger than TDP, 3% on average when $\kappa = 1$ and 15% on average when $\kappa = 5$. Figure 5.9(b) shows that



(a) (b) (c)

Figure 5.9: Effects of p with heterogeneous CPs

as the increase of p, the social welfare decreases. When p is large, the social welfare decreases faster. In addition, the social welfare for TDS is always higher. Figure 5.9(c) demonstrates that the capacity utilization also has the similar trends with the social welfare. As the increase of p, the capacity utilization decreases. In addition, the capacity utilization for TDS is also higher.

Remarks: The capacity has great impacts on ISP's, CPs' profit and consumers' welfare. Larger capacity always means higher ISP's, CPs' profit and consumers' welfare. In addition, ISP's optimal prices charged to CPs and end users all decrease with the capacity. When we compare TDS with SDP, we find that well designed TDS always has better performance than SDP with the metrics, ISP's profit, social welfare and capacity utilization. Higher number of groups indicates better performance of TDS. Moreover, the social welfare and capacity utilization are controversial to the ISP's strategy. Higher prices charged to CPs and end users always indicate lower social welfare and capacity utilization.

5.5 Summary

In this chapter, we proposed and studied time-dependent sponsoring (TDS), i.e., each CP can subsidize the data consumption of its users differently at different time. We built a Stackelberg game to model the interactions of strategic users, CPs and a monopoly ISP.

Our main findings are: 1) TDS improves the CPs' bandwidth utilization, CP's profit and consumers' welfare for slightly patient strategic users; but may result in controversial effects for highly patient strategic users, and 2) when CPs provide different subsidizations to different groups, the bandwidth usage can be improved significantly and so are CPs' profit and consumers' welfare, and 3) well designed TDS reduces the waste of capacity and thus improves social welfare and ISP's profit, compared with SDP.

CHAPTER 6
CONCLUSION AND FUTURE WORK

6.1 Conclusion

The surging traffic demand in wireless data networks due to the popularity of smart mobile devices and bandwidth-intensive applications poses challenges for the network operators to improve or even maintain network quality of the system. On the other hand, it also brings huge financial burden to the Internet service providers (ISPs) since supporting such demand-supply gap requires large investments. The pricing of data traffic and other services is central to the core challenges of network management, growth sustainability and monetization supports. In this thesis, we analyze the two main pricing proposals in smart data pricing, i.e., *time dependent pricing* and *sponsored data plan*. We try to understand the rationale behind the two pricing models, as well as their impacts to the wireless data market, in particular, who will benefit and who will be hurt from these schemes. We also propose and analyze a new pricing proposal, *time dependent sponsoring*, that combines the advantages of *time dependent pricing* and *sponsored data plan*.

- First, we focus on the *time dependent pricing*, which is a promising pricing method to relieve the congestion caused by the surging traffic demand. TDP captures the time-variation characteristic of demand by charging users dynamically over time and has the potential to even out time-of-the-day fluctuations in bandwidth consumption. We explore the design space of time dependent pricing. In particular, we focus on a number of schemes, e.g., the usage-based scheme, the flat-rate scheme, and a mixture of them which we called a cap scheme. Our main findings include: 1) the ISP obtains a higher profit with usage-based (or flat-rate) scheme

if the capacity is insufficient (or sufficient); 2) the usage-based scheme usually achieves a higher consumer surplus and more efficient traffic utilization than the flat-rate scheme; and 3) the cap scheme is strongly preferred by the ISP to further increase its revenue.

- Second, we analyze the *sponsored data plan*, a recent pricing proposal, i.e., when accessing contents from a particular CP, end users do not need to pay for that volume of traffic consumed, but the CP will sponsor for this data consumption. We build a two-class service model to analyze the consumers' traffic demand under the sponsored data plan with consideration of QoS. We use a two-stage Stackelberg game to characterize the interaction between CPs and the ISP and reveal a number of important findings. Our conclusions are: 1) When the ISP's capacity is sufficient, the sponsored data plan benefits consumers and CPs in the short run, but the ISP does not have incentives to further improve its service in the long run. 2) When ISP's capacity is insufficient, the ISP and end users may achieve a win-win trade, while the ISP and CPs always compete for the revenue. 3) The sponsored data plan may enlarge the unbalance in revenue distribution between different CPs; CPs with higher unit income and poorer technology support are more likely to prefer the sponsored data plan.

- Third, we propose and study one new smart data pricing scheme, *time dependent sponsoring*, i.e., content providers can decide when and how much to sponsor their traffic. The main novelty of TDS is its potential to improve Internet resource utilization by migrating data traffic from peak times to valley times. We formulate a Stackelberg game model to study the interactions between the ISP, CPs and users, and derive the optimal sponsoring fractions over different times under TDS. In particular, we develop a dynamic programming algorithm to solve a non-convex optimization in sponsoring decisions. We find that: 1) TDS improves the CPs' bandwidth utilization, CP's profit and consumers' welfare for slightly patient strategic users; but may result in controversial effects for highly patient strategic users, and

2) when CPs provide different subsidizations to different groups, the bandwidth usage can be improved significantly and so are CPs' profit and consumers' welfare, and 3) well designed TDS reduces the waste of capacity and thus improves social welfare and ISP's profit, compared with the sponsored data plan.

6.2 Future Work

Adopting smart data pricing to manage the network congestion is a vast research area, thus there exist a number of problems waiting for being solved. After the literature review of most related works, we present several future work here:

- First, when we build a game-theoretic model to analyze different schemes in time dependent pricing, we assume that there only exists one monopoly ISP, which is usually not true in reality. An interesting question here is that when there exist competition in the data market, e.g., two ISPs in the market, how each ISP should choose schemes in time dependent pricing. When the level of competition change, e.g., the number of ISPs increases, how about the change of ISPs' choices about the schemes and the prices decided. The optimal decision of each ISP may not be the dynamic cap any more. When multiple ISPs exist in the market, the basic model and decision space will be also much different with the monopoly ISP. Besides the choices of prices and usage schemes, each ISP also needs to consider how much bandwidth they should buy, which decides its QoS provided. This makes the problem more complex and thus needs to be carefully studied and discussed for the future work.

- Second, most of current works about the sponsored data plan focus on the feasibility analysis, including the benefit of such plan for consumers, ISPs and CPs and the fairness analysis between rich and poor CPs. Some of them propose new schemes to have better performance, e.g, higher benefit for ISPs and higher capacity utilization. However, few of them design the sponsored data plan in reality. This includes the survey of the willingness to adopt sponsored data plan for end

users, the UI design so as to make the usage convenient, the usage data analysis of the sponsored data plan, etc. The real system design of sponsored data plan also needs to consider the security problem since a non-sponsored CP has strong incentives to masquerade as a sponsored CP.

- Third, when analyzing the proposed time dependent sponsoring, we assume that each user is a strategic user, which is sensitive to price subsidization. For example, when the price subsidization at next slot is slightly higher than the current slot, all the patient users arriving at current slot will migrate to the next slot. This is usually not true in reality. Slight change in price subsidization should also result in slight change in traffic migration. Moreover, the analysis of time dependent sponsoring also ignores the valuation degradation caused by delaying usage. Even with such simplifications, the model of time dependent sponsoring in this thesis is complex and hard to analyze. New models are required to simplify the analysis and solve the above problems.

- Fourth, we investigated the mobile data offloading from economic and technical views in the literature review, but did not study any detailed mobile data offloading problem in this thesis. As we discuss in literature review, there may exist much interaction among time dependent pricing, sponsored data plan and mobile data offloading. In the future, we will focus on the combination of sponsored data plan and mobile data offloading, which provides a new business model and may bring more benefits for all the stockholders, i.e., Internet service providers, content providers and consumers.

REFERENCES

[1] Citrix bytemobile, mobile analytics reports. http://www.bytemobile.com/news-events/mobile_analytics_report.html.

[2] A. Aijaz, H. Aghvami, and M. Amani. A survey on mobile data offloading: technical and business perspectives. *IEEE Wireless Communications*, 20(2):104–112, 2013.

[3] M. Andrews, G. Bruns, and H. Lee. Calculating the benefits of sponsored data for an individual content provider. In *Proc. of IEEE CISS*, 2014.

[4] M. Andrews, U. Ozen, M. I. Reiman, and Q. Wang. Economic models of sponsored content in wireless networks with uncertain demand. In *Proc. of Smart Data Pricing Workshop, IEEE*, 2013.

[5] Jason Ankeny. At&t ceo: Content providers asking for 'toll-free' data plans. http://www.fiercemobileit.com/story/att-ceo-content-providers-asking-toll-free-data-plans/2012-07-18.

[6] Apple. Compare iphone models. http://www.apple.com/iphone/compare/.

[7] M. Armstrong. Competition in two-sided markets. *RAND Journal of Economics*, 47(3):668–691, 2006.

[8] AT&T. Sponsored data from at&t. http://developer.att.com/apis/sponsored-data.

[9] Cathy Avgiris. Comcast to replace usage cap with improved data usage management approaches. http://corporate.comcast.com/comcast-voices/comcast-to-replace-usage-cap-with-improved-data-usage-management-approaches.

[10] Severin Borenstein. The long-run efficiency of real-time electricity pricing. *The Energy Journal*, 26(3):93–116, 2005.

[11] C. Borgs, O. Candogan, J. Chayes, I. Lobel, and H. Nazerzadeh. Optimal multiperiod pricing with service guarantees. *Management Science*, vol. 60(7):1792 – 1811, 2014.

[12] Stephen Boyd and Lieven Vandenberghe. Convex optimization. *Cambridge, U.K.: Cambridge Univ. Press*, 2004.

[13] E. Bulut and B. K. Szymanski. Wifi access point deployment for efficient mobile data offloading. In *ACM international workshop on Practical issues and applications in next generation wireless networks*, 2012.

[14] Sinead Carew. Users complain, at&t blames data tsunami, 2012. http://blogs.reuters.com/mediafile/2012/02/14/users-complain-att-blames-data-tsunami/.

[15] Chia-Husan Chang, Phone Lin, Junshan Zhang, and Jeu-Yih Jeng. Time dependent adaptive pricing for mobile internet access. In *Proc. of Smart Data Pricing Workshop, IEEE*, 2015.

[16] Brian X. Chen. Shared mobile data plans: Who benefits? http://bits.blogs.nytimes.com/2012/07/19/shared-data-plans-verizon-att/?_r=0.

[17] M. H. Cheung, R. Southwell, and J. Huang. Congestion-aware network selection and data offloading. In *Proc. of IEEE CISS*, 2014.

[18] Man Hon Cheung and Jianwei Huang. Dawn: Delay-aware wi-fi offloading and network selection. *IEEE Journal on Selected Areas in Communications*, 33(6):1214–1223, 2015.

[19] Cisco. Cisco visual networking index: Global mobile data traffic forecast update 2014c2019 white paper.

http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

[20] Cisco systems, cisco visual networking index: Forecast and methodology, 2011-2016. http://www.baltimoreaircoil.com.

[21] Citrix. Mobile analytics report, otcober 2012. http://www.citrix.com/.

[22] S. Dimatteo, P. Hui, B. Han, and V. O. K. Li. Cellular traffic offloading through wifi networks. In *IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)*, 2011.

[23] W. Dong, S. Rallapalli, R. Jana, L. Qiu, K. Ramakrishnan, L. Razoumov, Y. Zhang, and T. W. Cho. ideal: Incentivized dynamic cellular offloading via auctions. In *Proc. of IEEE INFOCOM*, 2013.

[24] Ina Fried. Fcc chairman says at&t sponsored data plans warrant watching. http://recode.net/2014/01/09/fcc-chairman-says-att-sponsored-data-plans-worth-monitoring/.

[25] Lin Gao, George Iosifidis, Jianwei Huang, Leandros Tassiulas, and Duozhe Li. Bargaining-based mobile data offloading. *IEEE Journal on Selected Areas in Communications*, 32(6):1114–1125, 2014.

[26] Phil Goldstein. At&t sponsored data partner syntonic wireless to launch 'toll-free' content store. http://www.fiercewireless.com/story/att-sponsored-data-partner-syntonic-wireless-launch-toll-free-content-store/2014-07-10.

[27] Phil Goldstein. Google joins with india's bharti airtel for toll-free wireless internet service. http://www.fiercemobileit.com/story/att-ceo-content-providers-asking-toll-free-data-plans/2012-07-18.

[28] Google. International broadband pricing study: Dataset for public use. http://policybythenumbers.blogspot.gr/2012/08 /international-broadband-pricing-study.html.

[29] S. Ha, S. Sen, C. J. Wang, Y. Im, and M. Chiang. Tube: Time-dependent pricing for mobile data. In *Proc. ACM SIGCOMM*, 2012.

[30] S. Ha, S. Sen, C. J. Wang, Y. Im, and M. Chiang. Tube: Time-dependent pricing for mobile data. In *Proc. of ACM SIGCOMM*, 2012.

[31] P. Hande, M. Chiang, R. Calderbank, and S. Rangan. Network pricing and rate allocation with content provider participation. In *Proc. of IEEE INFOCOM*, 2009.

[32] P. Hande, M. Chiang, R. Calderbank, and J. Zhang. Pricing under constraints in access networks:revenue maximization and congestion management. In *Proc. IEEE INFOCOM*, 2010.

[33] Benjamin E. Hermalin and Michael L. Katz. The economics of product-line restrictions with an application to the network neutrality debate. *Information Economics and Policy*, 19:215–248, 2007.

[34] L. Hu, C. Coletti, N. Huan, I. Z. Kovacs, B. Vejlgaard, R. Irmer, and N. Scully. Realistic indoor wi-fi and femto deployment study as the offloading solution to lte macro networks. In *IEEE Vehicular Technology Conference*, 2012.

[35] J. Huang, Q. Xu, B. Tiwana, Z. M. Mao, M. Zhang, and P. Bahl. Anatomizing application performance differences on smartphones. In *Proc. ACM MobiSys*, 2010.

[36] George Iosifidis, Lin Gao, Jianwei Huang, and Leandros Tassiulas. A double-auction mechanism for mobile data-offloading markets. *IEEE/ACM Transactions on Networking*, 23(5):1634–1647, 2015.

[37] L. Jiang, S. Parekh, and J. Walrand. Time-dependent network pricing and bandwidth trading. In *Proc. IEEE NOMS*, 2008.

[38] C. Joe-Wong, S. Ha, and M. Chiang. Sponsoring mobile data: An economic analysis of the impact on users and content providers. In *Proc. of IEEE INFOCOM*, 2015.

[39] Carlee Joe-Wong, Soumya Sen, and Sangtae Ha. Offering supplementary network technologies: Adoption behavior and offloading benefits. *IEEE/ACM Transactions on Networking*, 23(2):355–368, 2015.

[40] G. Kesidis, A. Das, and G. D. Veciana. On flat-rate and usage-based pricing for tiered commodity internet services. In *Proc. of IEEE CISS*, 2008.

[41] Peter Key. Comcast, level 3, netflix, the fcc: Busy week for neutrality debate. http://www.bizjournals.com/philadelphia/blogs/technology/2010/12/comcast-level-3-netflix-the-fcc.html.

[42] Dongmyung Lee, Jeonghoon Mo, Jean Walrand, and Jinwoo Park. A token pricing scheme for internet services. In *Proc. of ICQT*, 2011.

[43] J. Lee, Y. Yi, S. Chong, and Y. Jin. Economics of wifi offloading: Trading delay for cellular capacity. *IEEE/ACM Transactions on Wireless Communications*, 13(3):1540–1554, 2014.

[44] K. Lee, J. Lee, Y. Yi, I. Rhee, and S. Chong. Mobile data offloading:how much can wifi deliver? *IEEE/ACM Transactions on Networking*, 21(2):53–550, 2013.

[45] S. Li, K. Xu, Y. Liu, and J. Wu. Edge overlay multicast to support comparable multi-class services. *Journal of High Speed Networks*, 17(1):13–36, 2008.

[46] P. Loiseau, G. Schwartzy, J. Musacchio, and S. Amin. Incentive schemes for internet congestion management: Raffles versus time-of-day pricing. In *Annual Allerton Conf.*, 2011.

[47] Patrick Loiseau, Galina Schwartz, John Musacchio, Saurabh Amin, and S. Shankar Sastry. Congestion pricing using a raffle-based scheme. In *Proc. of IEEE Netgcoop*, 2011.

[48] Patrick Loiseau, Galina Schwartz, John Musacchio, Saurabh Amin, and S. Shankar Sastry. Incentive mechanisms for internet congestion management: Fixed-

budget rebate versus time-of-day pricing. *IEEE/ACM Transaction on Networking*, 22(2):647–661, 2014.

[49] Qian Ma, Ya-Feng Liu, and Jianwei Huang. Time and location aware mobile data pricing. In *IEEE International Conference on Communications (ICC)*, 2014.

[50] Qian Ma, Ya-Feng Liu, and Jianwei Huang. Time and location aware mobile data pricing. *IEEE Transactions on Mobile Computing*, 2015.

[51] Richard T. B. Ma. Subsidization competition: Vitalizing the neutral internet. In *Proc. of ACM CoNEXT*, 2014.

[52] Richard T. B. Ma, Dah Ming Chiu, John C. S. Lui, Vishal Misra, and Dan Rubenstein. Internet economics: The use of shapley value for isp settlement. In *Proc. of ACM CoNEXT*, 2007.

[53] Richard T. B. Ma and Vishal Misra. Congestion equilibrium for differentiated service classes. In *Proc. of Forty-Ninth Annual Allerton Conference*, 2011.

[54] Richard T. B. Ma and Vishal Misra. The public option: a non-regulatory alternative to network neutrality. In *Proc. of ACM CoNEXT*, 2011.

[55] Sue Marek. Verizon's shammo: Content providers see value in toll-free data model. http://www.fiercewireless.com/story/verizons-shammo-content-providers-see-value-toll-free-data-model/2013-05-22.

[56] Jeonghoon Mo and Jean Walrand. Fair end-to-end window-based congestion control. *IEEE/ACM Trans. Netw.*, 8(5):556–567, 2000.

[57] P. Nabipay, A. Odlyzko, and Z.-L. Zhang. Flat versus metered rates, bundling, and'bandwidth hogs'. In *Proc. ACM NetEcon*, 2009.

[58] P. Njoroge, A. E. Ozdaglar, N. E. Stier-Moses, and G. Y. Weintraub. Investment in two-sided markets and the net neutrality debate. *Review of Network Economics*, 12(4):355–402, 2013.

[59] A. Odlyzko, B. St. Arnaud, E. Stallman, and M. Weinberg. Know your limits: Considering the role of data caps and usage based billing in internet access service. Public Knowledge White Paper, May 2012, avaliable at http://www.publicknowledge.org/know-your-limits-considering-role-data-caps-and-us.

[60] Andrew Odlyzko. Paris metro pricing for the internet. In *Proc. of ACM EC*, 1999.

[61] Andrew Odlyzko. The volume and value of information. *International Journal of Communication*, vol. 6:920C935, 2012.

[62] Martin J. Osborne and Ariel Rubinstein. A course in game theory. *MIT press*, 1994.

[63] Ioannis Ch. Paschalidis and John N. Tsitsiklis. Congestion-dependent pricing of network service. *IEEE/ACM Transactions on Networking*, 8(2):171–184, 2000.

[64] F. Qian, K. S. Quan, J. Huang, J. Erman, A. Gerber, Z. M. Mao, S. Sen, and O. Spatscheck. Web caching on smartphones: Ideal vs. reality. In *Proc. ACM MobiSys*, 2012.

[65] H. Raj, S. Saroiu, A. Wolman, and J. Padhye. Splitting the bill for mobile data with simlets. In *Proc. ACM HotMobile*, 2013.

[66] Dan Rayburn. Will bandwidth caps strangle netflix? not likely. http://www.streamingmedia.com/Articles/Editorial/Featured-Articles/Will-Bandwidth-Caps-Strangle-Netflix-Not-Likely-76134.aspx.

[67] Filippo Rebecchi, Marcelo Dias De Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, and Marco Conti. Data offloading techniques in cellular networks: A survey. *IEEE Communications Surveys & Tutorials*, 17(2):580–603, 2014.

[68] N. Ristanovic, J.-Y. Le Boudec, A. Chaintreau, and V. Erramilli. Energy efficient offloading of 3g networks. In *IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS)*, 2011.

[69] Soumya Sen, Carlee Joe-Wong, Sangtae Ha, and Mung Chiang. A survey of smart data pricing: Past proposals, current plans, and future trends. *ACM Computing Surveys*, 46(15), 2013.

[70] N. Shetty, G. Schwartz, and J. Walrand. Internet qos and regulations. *IEEE/ACM Trans. Netw.*, 18(6):1725–1737, 2010.

[71] V. Siris and D. Kalyvas. Enhancing mobile data offloading with mobility prediction and prefetching. In *ACM international workshop on Mobility in the evolving internet architecture (MobiArch)*, 2012.

[72] Hyun soo Ahn, Mehmet Gumus, and Philip Kaminsky. Pricing and manufacturing decisions when demand is a function of prices in multiple periods. *Operations Research*, vol. 55(6):1039 – 1057, 2007.

[73] J. Tadrous, A. Eryilmaz, and H. E. Gamal. Pricing for demand shaping and proactive download in smart data networks. In *Proc. of Smart Data Pricing Workshop, IEEE*, 2013.

[74] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. V. Vazirani. How many tiers? pricing in the internet transit market. In *Proc. ACM SIGCOMM*, 2011.

[75] Verizon Wireless. The more everything plan. `http://www.verizonwireless.com/landingpages/more-everything/#how-it-works`.

[76] C. Wong, S. Ha, and M. Chiang. Time-dependent broadband pricing: Feasibility and benefits. In *Proc. IEEE ICDCS*, 2011.

[77] Yuan Wu, Hongseok Kim, Prashanth H. Hande, Mung Chiang, and Danny H.K. Tsang. Revenue sharing among isps in two-sided markets. In *Proc. of IEEE INFOCOM*, 2011.

[78] Chunsheng Xin and Min Song. Dynamic spectrum access as a service. In *Proc. IEEE INFOCOM*, 2012.

[79] K. Xu, Y. Zhong, and H. He. Can p2p technology benefit eyeball isps a cooperative profit. *IEEE Transaction on Parallel and Distributed Systems*, 25(11):1101–1111, 2014.

[80] Q. Xu, J. Huang, Z. Wang, F. Qian, A. Gerber, and Z. M. Mao. Cellular data network infrastructure characterization and implication on mobile content placement. In *Proc. ACM SIGMETRICS*, 2011.

[81] W. Yoon and B. Jang. Enhanced non-seamless offload for lte and wlan networks. *IEEE Communications Letters*, 17(10):1960–1963, 2013.

[82] M. Yuksel, K. K. Ramakrishnan, S. Kalyanaraman, J.D. Houle, and R. Sadhvani. Quantifying overprovisioning vs. class-of-service: Informing the net neutrality debate. In *Proc. of IEEE ICCCN*, 2010.

[83] L. Zhang and D. Wang. Sponsoring content: Motivation and pitfalls for content service providers. In *Proc. of Smart Data Pricing Workshop, IEEE*, 2014.

[84] L. Zhang, W. Wu, and D. Wang. Time-dependent pricing in wireless data networks: Flat-rate vs. usage-based schemes. In *Proc. of IEEE INFOCOM*, 2014.

[85] L. Zhang, W. Wu, and D. Wang. Sponsored data plan: A two-class service model in wireless data networks. In *Proc. of ACM SIGMETRICS*, 2015.

[86] X. Zhuo, W. Gao, G. Cao, and Y. Dai. Win-coupon: An incentive framework for 3g traffic offloading. In *Proc.of Network Protocols (ICNP)*, 2011.