공학박사학위논문

혼잡환경에서의 단순성의 대가:

정보통신산업에서의 가격정책과 수익에 관한 연구

Price of Simplicity under Congestion:

On the Revenue and Pricing Schemes in the Telecommunication Industry

2013년 2월

서울대학교 대학원

산업공학과

이 동 명

혼잡환경에서의 단순성의 대가:

정보통신산업에서의 가격정책과 수익에 관한 연구

Price of Simplicity under Congestion:

On the Revenue and Pricing Schemes in the Telecommunication Industry

지도교수 박 진 우

위 논문을 공학박사 학위논문으로 제출함.

2012년 10월

서울대학교 대학원

산업공학과

이 동 명

이동명의 공학박사 학위논문을 인준함.

2012년 12월

위원장	오 형식	(인)
부위원장	박 진우	(인)
위 원	박 종헌	(인)
위 원	오 정훈	(인)
위 원	박 재욱	(인)

ABSTRACT

Dongmyung Lee
Department of Industrial Engineering
The Graduate School
Seoul National University

In this dissertation, we consider pricing schemes and revenues in the telecommunication industry. Recently, broadband Internet subscribers are increasing exponentially. In 1998, domestic Internet users were 3.1 million and have increased to 30 million in 2003 among which the broadband users were 10 million.

Despite the rapid increase in Internet demand, the prevailing pricing scheme, flat-rate pricing, has not been changed, which is inevitably pulling down the provider's revenue to its limit. Moreover, under the situation called 'disparity in Internet usage' where a small fraction of users generate most of the traffic, the flat-rate pricing made the users with small usage subsidize the ones with massive consumption.

Due to the aforementioned problems, the network providers are recently considering to adopt usage-based pricing. However, this has brought a strong customer resistance and thus, most of the providers are keep using the flat pricing.

In light of this, an interesting study related to pricing schemes and revenue in the network was conducted. In 2008 *IEEE JSAC*, one article argued that the revenue loss from using a simple entry fee in lieu of revenue-maximizing

scheme is small. The authors defined this loss as the 'price of simplicity' and showed mathematically that this price is low in a various environment.

However, what they have overlooked is 'congestion,' which is one of the crucial element in modeling the network environment. In effect, current Internet is a huge network connected with numerous users. In this context, a user's net-utility from using the Internet consists of not only the pure utility of service and the price paid but also the congestion cost where the cost grows as the number of users increases.

Thus, in this thesis, we reexamine the price of simplicity under the environment that exists "congestion externality." In particular, we compare revenues obtained using a flat price and two-part tariff and analyze the effect of congestion on the revenue loss when using a simple entry fee in lieu of the two-part tariff. Previous study has shown that when there is no delay disutility the revenue loss is small, which leads to a low "price of simplicity." However, in this study, we show that in an extreme case where all users are identical, the price of simplicity is substantial. Then we consider a more practical scenario where users have different preferences, and show that even in this case, under congestion externality, the price can be extremely high.

Keywords : Pricing, Internet, Flat Price, Two-Part Tariff, Price of Simplicity, Congestion Externality.

Student number : 2007-20639

Contents

영문 초록 **i**

1 Introduction **1**

1.1 Motivation 2

1.2 Purpose of this Study 2

2 Internet Pricing **5**

2.1 Pricing Schemes in Practice 5

2.1.1 Flat-rate Pricing 5

2.1.2 Usage-based Pricing 6

2.2 Pricing Schemes proposed in the Literature 6

2.2.1 Congestion Pricing 6

2.2.2 Priority Pricing 6

2.2.3 Smart-Market Pricing 7

2.2.4 Effective Bandwidth Pricing 7

2.2.5 Time-of-Day Pricing 8

2.2.6 Paris Metro Pricing 8

2.2.7 Cumulus Pricing 9

2.2.8 Token Pricing 9

3 Paris Metro Pricing **11**

3.1 System model 13

3.1.1 Single Pricing 13

3.1.2 Differentiated Pricing 15

3.2 Impact of PMP on Revenue and User Subscription 17

3.2.1 Uniform Distribution 19
3.2.2 Non-uniform Distributions 20
3.3 Competition: Duopoly 25
3.3.1 Three prices 25
3.3.2 Four prices 27
3.3.3 Discussion 30
3.4 Summary 30

4 Token Pricing **32**
4.1 Token Pricing Scheme 34
4.1.1 Fixed Congestion 35
4.1.2 Variable Congestion 37
4.1.3 Continuous Time 40
4.2 Numerical Analysis 41
4.2.1 Discrete Epochs 41
4.2.2 Continuous Time 42
4.3 Flat vs Token 44
4.3.1 Capacity-Sharing Service 47
4.3.2 Latency-Based Services 47
4.4 Token Scheme with Single Class 49
4.4.1 Comparison with Flat and Token with single service class 52
4.5 Summary 52

5 Price of Simplicity **55**
5.1 Flat-rate and Volume-based Pricing 55
5.2 Model 59
5.2.1 Identical User Case 60
5.2.2 Heterogeneous Users 70

6 Discussion **89**

6.1 Implementation of PMP in the Internet 89

6.2 Assumption on θ 89

6.3 Implementation of Token pricing in the Internet 90

6.4 Assumption on the utility function $u(x_i; v)$ 90

7 Conclusions **92**

Appendix A Analysis for Heterogeneous Users with $d = 0$ **94**

References **96**

국문 초록 **107**

List of Tables

2.1 Summary of Internet Pricing Schemes 10

3.1 Summary of Notations for Paris Metro Pricing 12
3.2 Optimal revenues and prices for firms A and B (three prices) 27
3.3 User subscription for firms A and B (three prices) 27
3.4 Optimal revenues and prices for firms A and B (four prices) . 29
3.5 User subscription for firms A and B (four prices) 29

4.1 Summary of Notations for Token Pricing 33

5.1 Comparison of Flat and Volume-based Pricing 56
5.2 Summary of Related Work 60
5.3 Summary of Notations for Price of Simplicity 63
5.4 All Cases for Comparing R_F and R_T 69

List of Figures

1.1 Internet ecosystem 1

3.1 The two functions y_1 and y_2 for the illustration of solution $x^*_\pi(p)$ of (3.3.3) 14

3.2 The pdfs of θ representing each different network environment 19

3.3 Numerical results for the three non-uniform distribution cases (including that of uniform distribution): Revenues with respect to $\alpha_{\pi 2}$ for each user distribution 21

3.4 Numerical results for the three non-uniform distribution cases (including that of uniform distribution): Prices with respect to $\alpha_{\pi 2}$ for each user distribution 22

3.5 Numerical results for the three non-uniform distribution cases (including that of uniform distribution): User participation with respect to $\alpha_{\pi 2}$ for each user distribution 23

3.6 Maximum revenues and user subscription for each user distribution 24

4.1 The Markov chain $\{s_n, n \geqslant 0\}$ 38

4.2 Values of Token pricing scheme with respect to K: (a) $g_1(x_\tau, p_\tau)$, (b) $g_2(x_\tau, p_\tau)$ 43

4.3 Values of Token pricing scheme with respect to M: $g_3(x_\tau, q)$ 44

4.4 Network capacity for Flat and Token pricing scheme with two service classes 45

4.5 Comparison of Flat and Token pricing scheme with two service classes (A=3) 48

4.6 Token scheme with Single Class: A single day is divided into two time zones 50
4.7 Comparison of Flat and Token pricing scheme with single service class (A=3) 53

5.1 Illustration of the Optimality Conditions 66
5.2 Revenue and Price of Simplicity for Identical Users with $d(T;C) = T/C$ 71
5.3 Revenue and Price of Simplicity for Identical Users with $d(T;C) = 1/(C-T)$ 72
5.4 Price of Simplicity for Heterogeneous Users Case: (a) PoS w.r.t. α for $d = 0$, (b) PoS w.r.t. N where '(i)' and '(h)' refers to identical and heterogeneous users, respectively. 79
5.5 Indifferent User type (v_0) and Delay for Heterogeneous Users Case 80
5.6 Revenue and Price of Simplicity (identical vs. heterogeneous) with respect to ϵ 81
5.7 Price of Simplicity with Users having Different Tastes for Congestion: (a) $\epsilon \sim U[0.1, 1]$; (b) $\epsilon \sim U[1, 2]$. 83
5.8 Welfare (W) and User Surplus (U) for Heterogeneous Users Case 86
5.9 Comparison of user surplus between the two pricing schemes: Flat and Two-part Tariff ($d = T/C$). $MS_F(MS_T)$ denotes the market share when using flat price (two-part tariff), and $v_{F>T}$ ($v_{T>F}$) stands for the user types where a flat price (two-part tariff) gives higher surplus than a two-part tariff (flat price). 87

5.10 Change in net user surplus generated by the transition from Flat price to Two-part Tariff: (a) and (c) Fraction of users who experience net user surplus gain or loss from the transition of pricing scheme; (b) and (d) Total sum of surplus change for the users experiencing net user surplus gain or loss. 88

1. Introduction

In this dissertation, we consider pricing schemes and revenues of the Internet service provider. Normally, an Internet market consists of three parties involved in the content delivery. First, there is a content provider (CP) which creates content for users. Second, an ISP delivers these content from the CP to the end-users. Lastly, an end-user receives the content created by the CP which is delivered through the ISP. The pricing schemes and revenues considered in this dissertation are particularly concerned with the ISP-user relationship rather than ISP-CP (Fig. 1.1). Thus, the revenue of ISP is obtained not from the CPs but from the end-users.

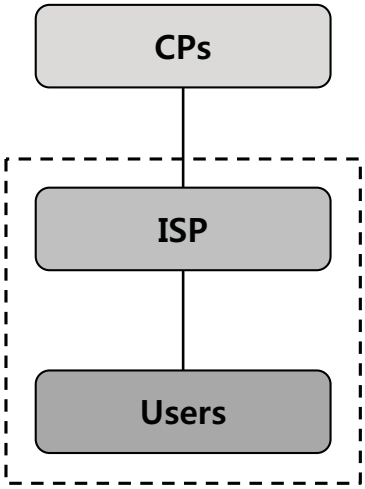


Figure 1.1: Internet ecosystem

1.1 Motivation

Internet service pricing has received considerable attention from economists and engineers. Due to the recent data traffic explosion, researchers have proposed several pricing mechanisms to allocate the scarce network resources efficiently [62, 39, 53]. However, there are several issues regarding the implementation of these sophisticated schemes. Associating a price with individual packets may lead to enormous accounting and billing overhead. In addition to the overwhelming complexity, they generally put too much burden on customers [66, 84]. Experience with the commercial side of the Internet illustrates that customers have a strong preference for simple schemes over complex mechanisms [77]. In an effort to address these issues, several researchers have suggested schemes that are simpler but still support some, if any, network resource efficiency [2, 27, 76]. Shenker et al.'s "Edge Pricing," locally computes charges based on simple *expected* values of congestion [27]. One example is QoS-sensitive time-of-day pricing that encourages users to time-shift their calls to later (or earlier) hours when rates are lower. In [76] Odlyzko proposed "Paris Metro Pricing (PMP)" for the Internet. The idea is to partition a network into serveral logically separate channels, each of which only differs in price, which makes the channels with higher prices attract less traffic, and thereby better service.

1.2 Purpose of this Study

Despite these efforts, current commercial ISPs adopt only a few pricing plans other than the ones described above. Of these schemes, flat pricing and volume-based pricing are the most predominant schemes in practice. According to [77], 'simplicity' is likely to be much more important on the Internet than in other communcation services and therefore one can expect the two

aforementioned pricing plans will dominate the market for the time being. However, researchers do not have a common opinion on which of the two pricing schemes is better. To summarize the debate on both pricing schemes, one can conclude that volume-based pricing is superior to flat pricing in terms of efficiency and revenue while flat pricing is superior in terms of simplicity, user acceptance, and thus practicality. In light of this issue, one study reported an interesting finding that flat pricing is not that inferior to other complex schemes in terms of revenue. [93] first defined a concept called the "*price of simplicity*," which is basically a lower bound on the ratio of flat fee revenue and maximum revenue. They showed that the loss associated with using a simple entry fee is small in various situations. In addition, they also compared with the Paris Metro Pricing and concluded that the gain from this type of price discrimination is again small. If this is the case, flat-rate pricing may not be such a bad choice for ISPs.

Unfortunately, previous studies have already found that for congested networks this may not be true. In effect, what [93] has overlooked is what economists refer to as "congestion externality," a crucial element in modeling the practical data network. Edell and Varaiya [21], in their INDEX experiment, concluded that a flat-rate is inefficient and volume-based pricing would be better for both ISPs and users. Kesidis et al. compared the flat and volume-based pricing schemes for tiered Internet services and claimed that usage-based pricing leads to a reduction in the aggregate overloads and higher user utilities versus flat pricing [55]. These studies are similar to ours in that they showed that flat pricing is inferior to a volumed-based scheme with regard to a certain aspect. However, the conclusion of [21] is based on an empirical study and [55] only stated that usage-based pricing reduces congestion, but did not mention the revenue difference compared to a flat price. Therefore, in this study, we aim to provide a more analytical result

on the existence and the extent of revenue loss created by using a simple flat entry fee.

- First, we consider an ideal case where all users have identical tastes in services. We compare the revenue and social welfare of two linear pricing schemes, flat price and two-part tariff, and show that the price of simplicity gets extremely high as the number of users increases.

- Second, we study a more realistic scenario with heterogeneous users. The price of simplicity, in this case, also increases as the network becomes congested, but the rate of increase is slower compared to the case in which users are identical. The reason is that providers can drop users of low willingness-to-pay to prevent the network from getting too congested.

- Last, we found that a two-part tariff is superior to a flat price in terms of user surplus and thus social welfare.

This thesis is organized as follows. Section 2 reviews some of the well-known Internet pricing plans in practice and those proposed in the literature. In chapters 3 and 4, we analyze the two simple yet efficient pricing schemes, Paris Metro Pricing (PMP) for the Internet proposed by Odlyzko [76] and our Token pricing scheme. Section 5.1 outlines the debate between flat-rate pricing and volume-based pricing and the related work. In section 5.2 we address the concept called the 'price of simplicity' by first considering an identical user case to derive analytical results. Next, we consider users of different preferences and determine several implications of the price of simplicity and social welfare. Finally, we summarize and conclude the dissertation in section 7.

2. Internet Pricing

Due to the popularity of mobile devices (especially, smartphones and tablets) and an enormous growth in quantity and quality of services (e.g., apps, videos, and clouds), the need for more sophisticated pricing schemes has grown substantially. This has brought about various pricing schemes proposed by researchers in various fields including economics, engineering, and management. Here, we summarize some of the well-known Internet pricing proposals in practice and in the literature.[1]

2.1 Pricing Schemes in Practice

2.1.1 Flat-rate Pricing

Flat rate is the access to the Internet at all hours and days of the year and for all (Internet) users at a fixed and usually cheap tariff. In the past, providers have charged a flat monthly fee for broadband access, irrespective of the actual time spent on the Internet or real usage [92]. Although many ISPs have introduced flat rates for Internet access in the past, most of them stopped providing the service after some time. The reason is that unlimited Internet services are not economically viable [5]. In other words, costs incurred from using flat-rate pricing exceeds the revenue. Thus, ISPs usually provide "flat up to a cap," a variation of flat rate that sets a maximum limit on the usage beyond which some form of penalty cost (proportional to the usage above the limit) is incurred for users [92].

[1] For an extensive survey of broadband pricing proposals refer to [92] and references therein

2.1.2 Usage-based Pricing

Usage-based pricing, as the name implies, is a pricing scheme in which a (total) payment of a service is based on its consumption or usage [98, 60]. Usage-based pricing is currently the dominant pricing plan around the world (e.g., U.S., Europe, and Asia) since it is regarded as a possible solution to (the limitations of) flat rate pricing. However, it particularly favors ISPs over Internet users, which is why it has to overcome the user acceptance issue in order to be a sustainable pricing scheme in practice.[2]

2.2 Pricing Schemes proposed in the Literature

2.2.1 Congestion Pricing

Congestion pricing has a dynamic nature such that the network announces prices based on current congestion levels and the consumer response to these real-time prices is fed back to the control loop to compute a new set of prices. Ganesh et al. [33] introduces a scheme in which the burden of rate allocation from the network is shifted to the end-systems. The proposed scheme which uses congestion prices to provide both feedback and incentives to end-systems also proposed a mechanism that enables users to achieve faster convergence to equilibrium.

2.2.2 Priority Pricing

In [12], Cocchi et al. analyzed a pricing scheme in a multiple service network with priority classes. Users may request different classes of service by modifying the bits in their packets. Through simulation, they demonstrate that it is feasible to set the prices in a way that every user is more satisfied with the combined cost and performance of a network with graduated prices. That

[2]See chapter 5.1 for a more detailed comparison between the two practical pricing schemes.

is, some users enjoy less performance with reduced price while some users experience higher quality service with some monetary penalty.

2.2.3 Smart-Market Pricing

In [63], the authors propose a scheme using an auction. The user associates a price with each packet, carried in the packet's header, denoting the willingness-to-pay of the user for transmission. This price, which acts as a 'bid,' can be modified by the user. The network obtains and sorts all the bids after which a threshold value is determined. All the packets whose bid exceeds this threshold gets transmitted. The threshold value is naturally determined by the network's capacity limit and represents the marginal congestion cost. Each transmitted packet is then charged the same amount, the marginal cost of congestion.

2.2.4 Effective Bandwidth Pricing

Effective bandwidth pricing [52, 51] is designed to induce the user to declare the true values for the mean and the peak cell rates of general traffic sources during call admission control (CAC). The user is charged according to a linear function placed tangent to the effective bandwidth curve of the source [29]. One of the important characteristics of this pricing scheme is that the network can deduce the anticipated load generated by a user. The scheme allows the network to infer the actual function from the user's declaration. The user is assumed that he wants to minimize the economic cost of the connection. After the network understands this anticipated load, the effective bandwidth-based CAC can work properly. All user should provide is the expected value of the traffic stream. For example, in the case of on-off sources with known peak rates, the user only needs to notify the expected mean rate which can be easily measured.

2.2.5 Time-of-Day Pricing

As the name implies, the rates in time-of-day (ToD) pricing vary according to the certain time (interval) of day at which the service is consumed. This scheme is most widely used for utilities. In the industrial electricity market, for instance, the price per kWh might be 3 cents during all hours, except from 8:00 am to 8:00 pm when the price might be 5 cents. Besides this simple form of ToD rates, more complicated variations can have a large number of time frames (i.e., intervals) with each having different rates. ToD pricing has also been used for Internet access. BSNL in India offers unlimited night time (2-8 am) downloads on monthly data plans of $10 and above. Also, European provider Orange has a "Dolphin Plan" for $23.58 per month that allows unlimited Internet access during a so-called "happy hour" that corresponds to consumers' morning hours (8-9 am), lunch time (12-1 pm), late afternoon break (4-5 pm), or late night (10-11 pm) [92]. Parris et al. [80] considered a variation of ToD pricing, which divides a day into peak and off-peak periods and incorporates the time elasticity of demand. They conclude that peak-load pricing alleviates peak utilization and reduces blocking probability of all traffic classes, and improves revenue by making a more even distribution of demand over peak and off-peak periods.

2.2.6 Paris Metro Pricing

Paris metro pricing (PMP) for the Internet was proposed by Odlyzko [76] as a simple solution to provide differentiated services. It is to partition a network into several logically separate subnetworks, each of which is identical in treating the packets (i.e., best effort) but prices are differentiated. This setting makes the subnetworks with higher prices attract less traffic, and thereby more attractive to users (in terms of congestion). Despite the fact that it does not provide a complete efficiency in network utilization, it is the

simplest (and thus practical to both providers and users) solution for product differentiation.[3]

2.2.7 Cumulus Pricing

The cumulus pricing scheme (CPS) is based on a contract between a user and the service provider. A customer has to specify his expected user requirements along with a flat rate to be paid for the service. Depending on the discrepancies between the specified (expected) and actual demand, if the excess demand is above the threshold the contract is renegotiated. In [82], the authors show that CPS balances the following three requirements for a suitable Internet pricing scheme in an integrated manner: network efficiency, user acceptance, and technical feasibility. Also, Hayel and Tuffin study such cumulus pricing scheme and use a metaheuristic approach to optimize the total network revenue in terms of the renegotiation threshold [41].

2.2.8 Token Pricing

Lee et al. [19] introduced a scheme called Token pricing in an attempt to propose both practical and efficient Internet pricing scheme. As in flat-rate pricing, the users face a fixed price. However, users consume tokens only when they want a higher quality of service (QoS) while the network is congested. This mechanism encourages users to consume in great amount when they have a high utility for the service. As a result, the social welfare increases. The benefits of such a pricing scheme in reducing peak network congestion in a real network is yet to be explored [92].[4]

[3]For a detailed analysis of PMP scheme see 3.

[4]For the whole analysis of Token pricing scheme see 4.

Table 2.1: Summary of Internet Pricing Schemes

	Pricing Schemes	References
Pricing Schemes in Practice	Flat-rate pricing	[93], [3], [31], [78]
		[67], [79]
	Usage-based pricing	[98], [40], [60]
Efficient Pricing Schemes	Congestion pricing	[33], [81], [49]
	Priority pricing	[12], [24], [38], [65]
	Smart-market pricing	[63], [62], [68], [69]
	Effective bandwidth pricing	[51], [52], [53], [13]
		[14], [28], [45], [46]
	Edge pricing	[27]
	Expected capacity pricing	[11], [10], [50], [15]
		[20]
	Responsive pricing	[72], [61], [70], [71]
		[44]
	Proportional fairness pricing	[35], [54], [16]
	Application-based pricing	[23]
	Auction-based pricing	[42], [91], [100], [94]
		[73], [57], [47]
	Reservation-based pricing	[80], [79], [17]
	Time-of-Day Pricing	[79]
Towards More Practical Pricing Schemes	Paris metro pricing	[76], [6]
	Cumulus pricing	[41], [82], [83]
	Token pricing	[19]

3. Paris Metro Pricing

The usage of the *Internet* was dominated by traditionally 'typical' data services (e.g., e-mail, web browsing, file transfer, etc.). These services do not require severe bandwidth overhead on the network since the traffic generated by those applications usually tolerate relatively large packet delays. However, as we clearly witness these days, new internet applications such as VoIP, IPTV, and many smart phone applications can be charaterized as delay-constrained (or delay-sensitive), and thus require higher requirements. In addition, by 2014, mobile data traffic will double every year through 2014, increasing 39 times between 2009 and 2014, according to Cisco forecast [9]. This increase of diverse quality of service (QoS) requirements gives the rationale to develop new methods capable of treating the delay-constrained and non-delay-constrained traffic differently. A possible solution is to give priority to the delay-constrained traffic in the queues of the network [27] but without an appropriate *pricing* scheme, any prioritization is useless [64].

In this section, we study the use of pricing mechanism called Paris Metro Pricing (PMP) proposed by Odlyzko [76]. Under the PMP scheme, the network is split into subnetworks. The tariff for each subnetwork is different, expecting a lower congestion for highly priced networks. This method does not offer any QoS guarantees, so that it is somewhat weak compared to several other approaches. However, due to its simplicity, it is, indeed, very attractive to many practitioners.

In this regard, a few papers have examined the PMP scheme for charging packet networks. In [85], the authors present a mathematical model in which all packets are generated by the same kind of application and all users have

Table 3.1: Summary of Notations for Paris Metro Pricing

Symbol	Meaning
N	total number of users
α_i	fraction of network i
θ	user type parameter
X	number of active users (single)
X_i	number of active users in network i (differentiated)
x_π	network congestion (single)
$x_{\pi i}$	network congestion in network i (differentiated)
p	network access price (single)
p_i	network i's access price (differentiated)
$R(\cdot)$	revenue function

the same valuation of QoS. Even though they have showed the existence and uniqueness of the stability using queueing theory, the assumptions seem too strong to have practical implications. Paper [26] is the closest work to ours in that, based on their proposed model, they tried to show whether a network service provider (a single-constrained monopolist) can be profitable using PMP strategy. However, there are several differences with our study. First, although they modeled the user's satisfaction as the utility function they only showed the surplus from the providers perspective. In addition, the results they showed were merely for uniform distribution. Most of the papers in which PMP is analyzed take the uniform distribution of users as one of the assumptions for their model. However, in [34] the authors pointed out that it is required to assess the importance of those assumptions. In this spirit, we adopt a basic model from [98] and further extend it so as to analyze the economic aspects of this pricing scheme and to address the limitations of the existing papers.

3.1 System model

3.1.1 Single Pricing

In this section, we develop a system model based on [98] with one network which will be used throughout the paper. Consider a communication system with a large population of N users each charaterized by a type θ that is an independent random variable distributed in $[0, 1]$ with $f(\theta)$ and $F(\theta)$ as its probability density function and cumulative distribution function respectively. A user of type θ finds the network connection acceptable if the number of users X using the network and the price p are such that

$$\frac{X}{N} \leqslant 1 - \theta \quad \text{and} \quad p \leqslant \theta. \tag{3.1}$$

In this expression, N is the capacity of the network. Due to our analysis purpose, note that the capacity in the model is different from the conventional one that uses *bps*. Also, we will later divide the network into two subnetworks, each with capacity $N/2$. The expression interprets that a user with a large value of θ is willing to pay quite a lot for the connection but he expects a low utilization (or congestion) for a high quality of service. Conversely, a user with a small value of θ does not want to pay much for his connection but is willing to tolerate high delays. For example, we can view the users with large θ as users of VoIP and those with small θ as web browsers.

Assume that the network connection price is $p \in (0, 1)$. If the number of users in the network at a certain time period is X, then a user of type θ connects if the inequalities above are satisfied, i.e., if $\theta \in [p, 1 - X/N]$. Since θ is distributed in $[0, 1]$, the probability that a random user connects is $[F(1 - X/N) - F(p)]^+$. Accordingly, the number X of users that connect is binomial with mean $N \times [F(1 - X/N) - F(p)]^+$, so that

$$\frac{X}{N} \approx \left[F(1 - \frac{X}{N}) - F(p) \right]^+ \tag{3.2}$$

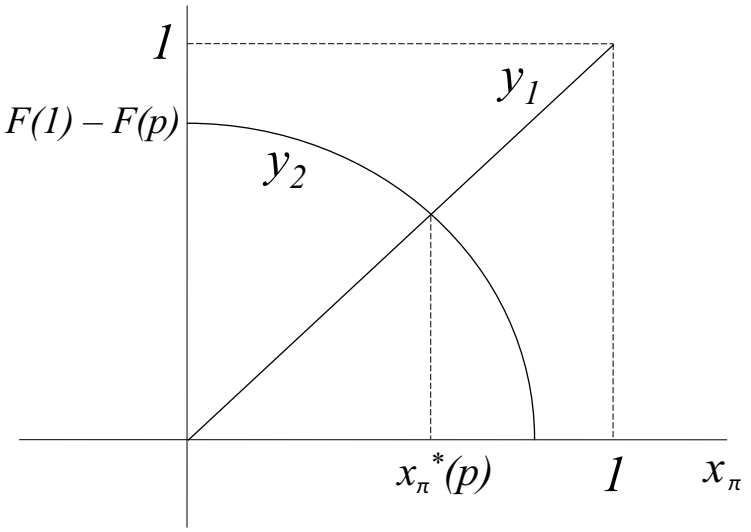


Figure 3.1: The two functions y_1 and y_2 for the illustration of solution $x_\pi^*(p)$ of (3.3)

by the law of large numbers, since N is large. By letting $x_\pi = X/N$, (1) can be expressed as

$$x_\pi = \left[F(1 - x_\pi) - F(p)\right]^+ . \tag{3.3}$$

Let the left-hand side and the right-hand side of (2) be $y_1(x_\pi)$ and $y_2(x_\pi)$ respectively. Then, for a fixed p, the solution of (3.3) can be illustrated as in Figure 3.1.

Since y_1 is increasing and y_2 is non-increasing, with both functions sharing the same domain of $[0, 1]$, a unique solution $x_\pi(p)$ exists. By taking the derivative of (2) with respect to p we have

$$\frac{dx}{dp} = -\frac{f(p)}{1 + f(1 - x_\pi)} \; < \; 0 \tag{3.4}$$

which shows that x_π is a decreasing function of p.

Now, we solve the provider revenue ($R(p)$) maximization problem. In other words, we find the price p that maximizes the product of the number

of users of the network and the price p, that is

$$p^* \quad = \quad \arg\max_p R(p) = p \cdot x_\pi(p). \tag{3.5}$$

Since $R(p)$ is a continuous function with $R(0) = 0$ and $R(1) = 0$, there exists at least one solution to the above maximization problem. Generally, it is nontrivial to find a closed form solution for $x_\pi(p)$, hence, for solving (3) a numerical method is performed. Nevertheless, to illustrate, we show a simple example when θ is uniformly distributed in $[0, 1]$. Equation (1), then, reduces to

$$\frac{X}{N} \approx \left(1 - \frac{X}{N} - p\right)^+. \tag{3.6}$$

Solving this expression we find that $x_\pi := X/N = (1-p)/2$. The operator can maximize his revenues by choosing the value of p that maximizes $px_\pi = p(1-p)/2$. The maximizing price is $p = 1/2$ and the corresponding value of px_π is $1/8$, which measures the revenue divided by N[1].

3.1.2 Differentiated Pricing

Consider now a situation where Paris Metro Pricing is applied. The network service provider divides the network into two subnetworks, each with different capacity. That is, there exist two networks: network 1 with price p_1 and capacity $N\alpha_{\pi 1}$ and network 2 with price p_2 and capacity $N\alpha_{\pi 2}$ (where $\alpha_{\pi 1} + \alpha_{\pi 2} = 1$). Without loss of generality, we assume $p_1 > p_2$. The users will select one of the two networks, based on the prices and utilizations. A user joins if there is an acceptable network and he chooses the cheapest network if both are acceptable. Moreover, if both networks are acceptable and, by

[1]For merely flat pricing, due to its simplicity, we were able to get results from curvilinear distributions (beta distribution). The pdf of the beta distribution is defined as $f(x, \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1} / \int_0^1 u^{\alpha-1}(1-u)^{\beta-1}\, du$ and the results (maximum revenue) for some instances are as follows: 0.1045 ($\alpha = 0.5$, $\beta = 0.5$), 0.1176 ($\alpha = 2$, $\beta = 5$), 0.1620 ($\alpha = 5$, $\beta = 2$), 0.1488 ($\alpha = 2$, $\beta = 2$), and 0.1610 ($\alpha = 3$, $\beta = 3$)

any chance, have the same price, a user will join the one with the smallest utilization because it offers a marginally better QoS.

If the number of users in the two networks are X_1 and X_2 respectively, then a user of type θ chooses network 2 if $X_2/(N\alpha_{\pi 2}) \leqslant 1-\theta$ and $p_2 \leqslant \theta$. The probability that θ falls between p_2 and $1 - X_2/(N\alpha_{\pi 2})$ is then

$$\left[F(1-\frac{X_2}{N\alpha_{\pi 2}}) - F(p_2)\right]^+. \tag{3.7}$$

Then, $x_{\pi 2} := X_2/N$ is given by

$$x_{\pi 2} = \left[F(1-\frac{x_{\pi 2}}{\alpha_{\pi 2}}) - F(p_2)\right]^+. \tag{3.8}$$

A user will opt for network 1 if $X_1/(N\alpha_{\pi 1}) \leqslant 1-\theta$, $p_1 \leqslant \theta$ and $X_2/(N\alpha_{\pi 2}) > 1-\theta$. Thus, we find that $x_{\pi 1} := X_1/N$ is such that

$$x_{\pi 1} = \left[F(1-\frac{x_{\pi 1}}{\alpha_{\pi 1}}) - F(\max\{p_1, 1-\frac{x_{\pi 2}}{\alpha_{\pi 2}}\})\right]^+. \tag{3.9}$$

To determine the prices p_1 and p_2 that maximize the revenue of the operator, one needs to maximize the total revenue R obtained by both two subnetworks over p_1 and p_2, mathematically it is equivalent to solving the following problem

$$R = \max_{p_1,p_2} \; p_1 x_{\pi 1} + p_2 x_{\pi 2}. \tag{3.10}$$

By solving the problem when the network capacity is divided exactly half (again with uniform distribution of θ), one can show that the maximum occurs for $p_1 = 8/13$ and $p_2 = 11/26$ and that the maximum revenue equals to $25/156$. Eventually, we may conclude that the service differentiation with Paris Metro Pricing increases the revenue from $1/8$ to $25/156$, or by 28.2%. This can be one good rationale to use PMP instead of not using it. In this regard, we show, from another perspective, that it is recommended to use PMP. The provider who is willing to use PMP will want to know how much additional capacity is needed (without using PMP) in order to have as much

as revenue obtained from using PMP (In this analysis we assume that the cost of increasing the network capacity is considerable).

Proposition 3.1.1 *PMP gives an increase in revenue by 28.2% which is equivalent to the amount when using only one network (and not using PMP) and increasing the network capacity by 78.6% given the original capacity N*

Proof: Let the increased capacity be $N(1+\beta)$. Considering the new increased capacity gives

$$\frac{X}{N} \approx \left(1 - \frac{X}{N(1+\beta)} - p\right)^+.$$

Solving for x, we have

$$x = \frac{1+\beta}{2+\beta}(1-p),$$

then solving the optimization problem (3.10), we have the optimal revenue with extended capacity as R_{ext}. By equating this with the optimal revenue obtained by PMP, R_{pmp} gives corresponding β as

$$\begin{aligned} R_{ext} = \frac{1}{4}\left(\frac{1+\beta}{2+\beta}\right) &= \frac{25}{156} = R_{pmp}, \\ \beta &= \frac{44}{56}. \end{aligned}$$

□

3.2 Impact of PMP on Revenue and User Subscription

Since we have seen the price that maximizes the provider's revenue and its corresponding revenue with θ following uniform distribution, we will now see how these values as well as the fraction of users joining the network change with more general cdfs of θ. The rationale behind this is that uniform distribution alone cannot represent various type of user population. For instance, there might be a community of users in which the majorities use real-time

services. On the other hand, one user group might consist of majority of people requiring non-delay-constrained services (e.g., email). Therefore, we need to represent these typical user distributions and incorporate them in the study of analyzing the impact of PMP.

In this regard, we define three additional probability distributions (f_1, f_3 and f_4) that capture three typical user distributions as follows (figure 2):

$$f_i(\theta) = \begin{cases} 2 - 2\theta & \text{if } \ i = 1 \\ 1 & \text{if } \ i = 2 \\ 2\theta & \text{if } \ i = 3 \end{cases}$$

where $\theta \in [0, 1]$ for all three distributions. Also,

$$f_4(\theta) = \begin{cases} 4\theta & \text{if } \ 0 \leqslant \theta \leqslant \frac{1}{2} \\ 4 - 4\theta & \text{if } \ \frac{1}{2} < \theta \leqslant 1 \end{cases}$$

The first pdf (f_1) comes from a general concave cdf of θ and stands for a network consisting of high population of users generating non-delay-constrained traffic . The second function comes from a general linear cdf of θ and exactly corresponds to the uniform distribution and represents a well balanced network of users. The third one (f_3) is a pdf that represents a convex cumulative distribution function indicating the networks where the majority of users require high QoS levels. The last one (f_4) is a pdf of a 'S-shaped' cdf and designates a network with majority of users requiring intermediate QoS levels. We will denote the network environment for each user population type as NT1, NT2, NT3, and NT4 henceforth.

To see whether PMP actually gives benefits to both providers and users we, first, have to check whether it is always better irrespective of network environment. Secondly, even though this is true, we also have to check whether the average surplus of using PMP over every network environment is better than the value when it is not used. This comes from the fact that the type

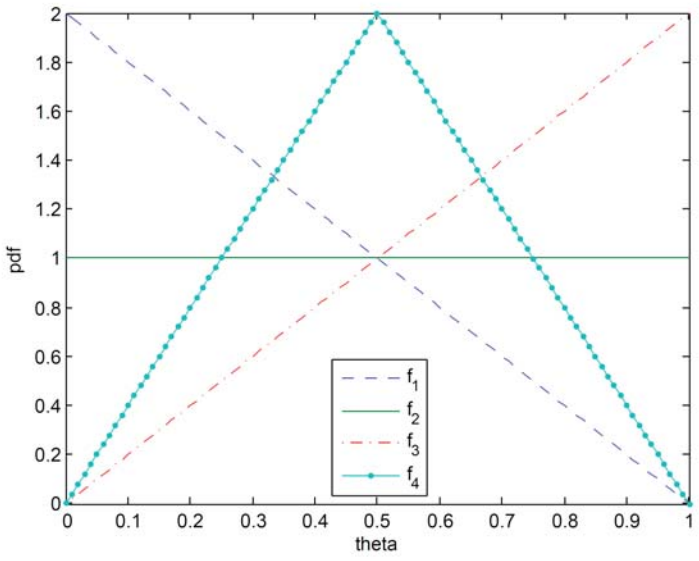


Figure 3.2: The pdfs of θ representing each different network environment

of user population will change over time since users consisting the network will require different service types at different time segment of interest. In light of this, we derive the maximum revenue and the corresponding user participation for each of the four user distributions. We will denote $\widehat{R}^1_{max}$, $\widehat{R}^2_{max}$, $\widehat{R}^3_{max}$ and $\widehat{R}^4_{max}$, and $\widehat{\alpha}^1_{\pi 2}$, $\widehat{\alpha}^2_{\pi 2}$, $\widehat{\alpha}^3_{\pi 2}$ and $\widehat{\alpha}^4_{\pi 2}$ as the maximum revenues and their corresponding $\alpha_{\pi 2}$ values achieved by using PMP scheme for the distributions f_1, f_2, f_3 and f_4

3.2.1 Uniform Distribution

For the uniform distribution, we could express $\widehat{R}^2_{max}$, p_1, p_2, $x_{\pi 1}$, and $x_{\pi 2}$ with respect to just one variable $\alpha_{\pi 2}$ after some manipulation. Each of these values are illustrated in Figure **??** and we find that the maximum revenue as well as the corresponding $\alpha_{\pi 2}$ is as follows:

$$\widehat{R}^2_{max} = 0.1608 \ , \quad \widehat{\alpha}^2_{\pi 2} = 0.43$$

In this example, the revenue increases by 28.6% compared to flat pricing. We could notice, from this analysis, that $\alpha_{\pi 2}$ which maximizes revenue is in fact slightly less than 0.5. This tells us that when using two subnetworks to apply PMP, dividing it exactly half may not be the best choice for the network service provider. Also, since the range of $\alpha_{\pi 2}$ that results in the revenue within 90% of the maximum value covers about 68% ($\geqslant 50\%$) of its whole interval, we could say that the revenue is not that sensitive to the ratio of dividing a network.

Figure **??**c shows how the network utilization changes as $\alpha_{\pi 2}$ increases. As you see from the figure, the total network utilization (denoted as subscription f2 in Fig. 3c) is at maximum when $\alpha_{\pi 2} = 0.43$. This is the exact $\alpha_{\pi 2}$ value that maximizes the total revenue. This implies that, when the network service provider maximizes his own profit, the network utilization is also maximized which is naturally true since the more users join the network the more revenue can be achieved (however, this does not imply that those two values have to be exactly the same, instead it means they will tend to have similar values). Moreover, it is easy to see from Figure 1c that whatever the value of $\alpha_{\pi 2}$ is, PMP strategy always results in higher utilization than not performing it.

3.2.2 Non-uniform Distributions

In effect, if the user population is not represented by a uniform distribution, it is difficult to derive the maximum revenue of the service provider when PMP is used. This is due to the fact that it involves a maximization problem over three variables of $\alpha_{\pi 2}$, p_1, and p_2. The results for the three non-uniform user distributions is illustrated in figures 3.3-3.5 and we find the maximum revenues and corresponding $\alpha_{\pi 2}$ values as follows (for the derivation of all the values for uniform and non-uniform distributions see [25]):

$$\widehat{R}^1_{max} = 0.1512, \quad \widehat{\alpha}^1_{\pi 2} = 0.45$$

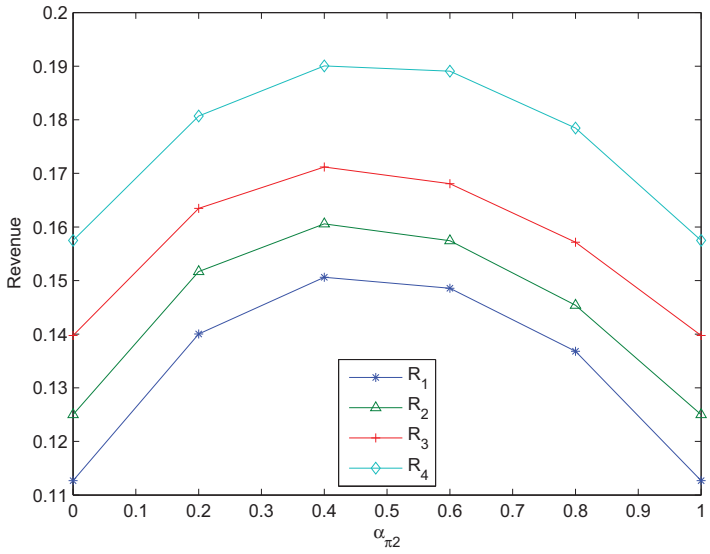


Figure 3.3: Numerical results for the three non-uniform distribution cases (including that of uniform distribution): Revenues with respect to $\alpha_{\pi 2}$ for each user distribution

$$\widehat{R}^3_{max} = 0.1713, \quad \widehat{\alpha}^3_{\pi 2} = 0.45$$

$$\widehat{R}^4_{max} = 0.1908, \quad \widehat{\alpha}^4_{\pi 2} = 0.50.$$

From the previous section we have seen that when the network consists of a well-balanced users, the capacity ratio between two subnetworks is not that sensitive as long as it is close enough to the optimal ratio. We, now, try to see whether this is also true for the rest of the three network types. The sensitivity of the total revenue with respect to $\alpha_{\pi 2}$ can be derived from the R-$\alpha_{\pi 2}$ plots for each of the three non-uniform distributions (Figure 3.3). That is, we would like to see how the revenue changes when the ratio of the capacities between two subnetwork is slightly altered from the optimal value ($\alpha^*_{\pi 2}$). From Figure 3.3 we could see that the range of $\alpha_{\pi 2}$ that produces the

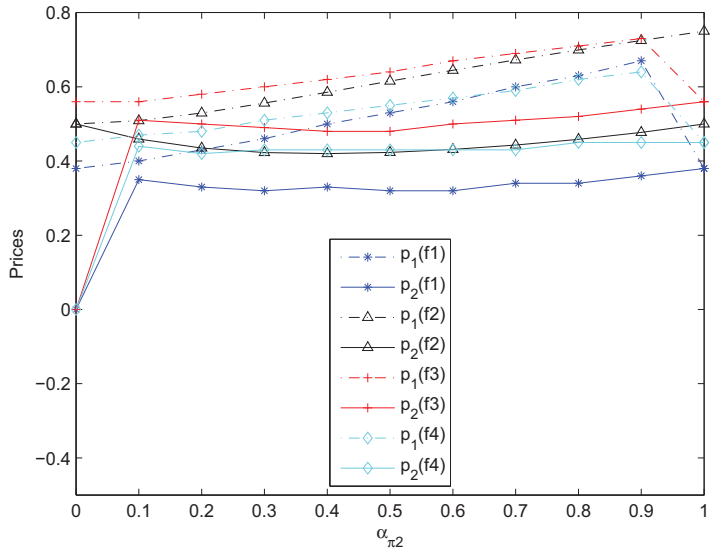


Figure 3.4: Numerical results for the three non-uniform distribution cases (including that of uniform distribution): Prices with respect to $\alpha_{\pi 2}$ for each user distribution

revenue within 90% of maximum value spans 65%, 74%, and 77% of its overall interval for NT1, NT3 and NT4 respectively. Thus, with the sensitivity of NT2 we could say that the revenue achieved in the environment NT1 is relatively sensitive to the ratios of its subnetwork capacity which should be considered for a provider when its servicing environment is of NT1 most of the time. However, in general (and on average), determining the proportion of the subnetworks is not that critical.

Comparing with the maximum revenues when PMP is not used, we eventually see that no matter what type of user population may be, using PMP always produces a higher revenue for the network service providers (Figure 3.6a). Similarly, PMP scheme is always favored by network users, regardless

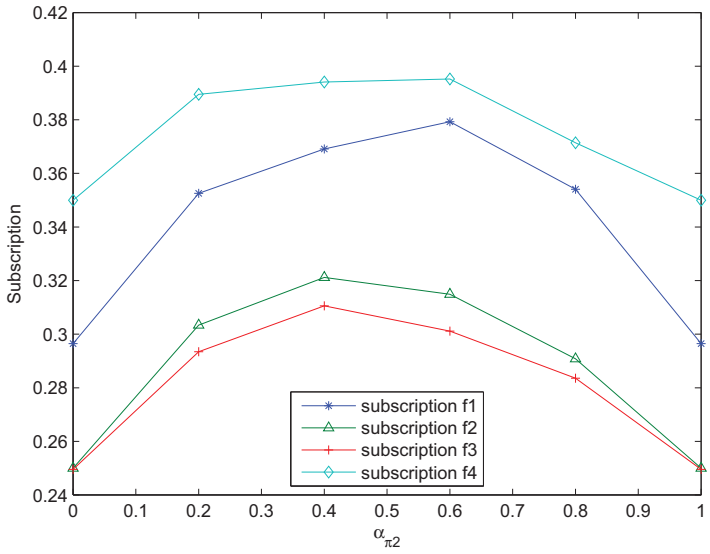
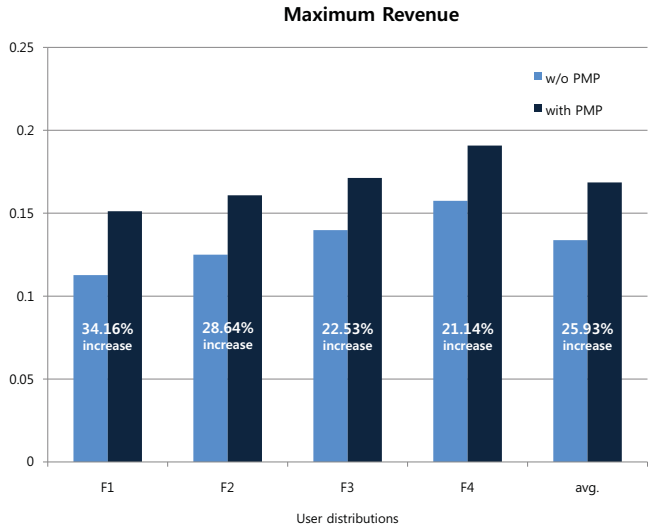


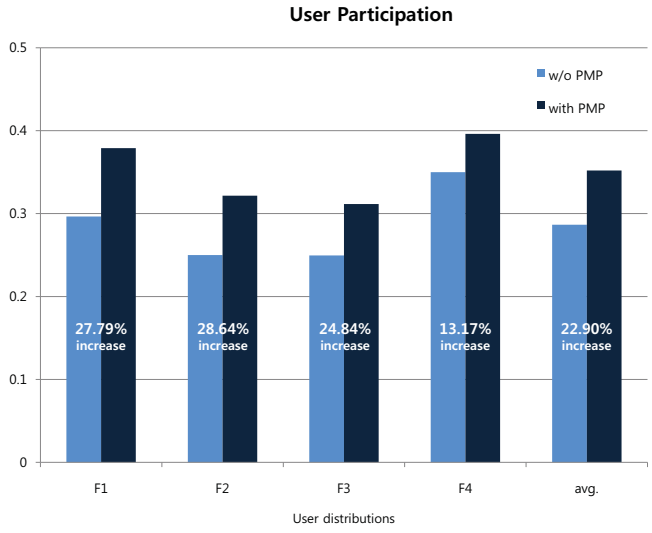
Figure 3.5: Numerical results for the three non-uniform distribution cases (including that of uniform distribution): User participation with respect to $\alpha_{\pi 2}$ for each user distribution

of the characteristics of user groups as shown in Figure 3.6b. The users' satisfaction is measured by the number of users who join the network (user participation) because, based on our system model, only the satisfied users receive the network services. In summary, the average values of 4 network environment between using PMP and not using PMP clearly show that PMP is, indeed, a superior pricing scheme in terms of both the revenue and network utilization (user participation).

Observation 3.2.1 *In a network where only one network service provider (a monopolist) provides internet service, it is always better for the provider to use Paris Metro Pricing scheme than to just provide a general service with flat pricing. In addition, PMP scheme is always favored by the users*



(a) Maximum revenue



(b) User subscription

Figure 3.6: Maximum revenues and user subscription for each user distribution

consisting the network.

Observation 3.2.2 *For a network service provider to achieve the maximum revenue, determining the fraction of the capacity of the subnetworks is less critical for all of the network environment.*

3.3 Competition: Duopoly

In previous sections, we have noticed that a single network service provider (a monopolist) will have an advantage for using PMP scheme. Then one might ask, "what happens when there are two network service providers? Does the Nash equilibrium even exist for the duopoly case?" In this section, we try to answer this question by modeling the situation using game theory.

3.3.1 Three prices

We first analyze the situation where there are two service providers with only one using PMP. That is, one ISP (firm A) provides the service with a single price p_a and the other ISP (firm B) uses PMP with the lower price p_{b1} and the higher price p_{b2}. Both the competing firms have equal capacity where firm B splits the two subnetworks exactly half. Then, with this situation, there are 5 possible cases as follows:

1. $p_a < p_{b1} < p_{b2}$.
2. $p_{b1} < p_a < p_{b2}$.
3. $p_{b1} < p_{b2} < p_a$.
4. $p_a = p_{b1} < p_{b2}$.
5. $p_{b1} < p_a = p_{b2}$.

Let R_A and R_B be the revenue (function) of the two firms respectively. From the similar derivation as in section 2 we can obtain the revenue functions of both firms for each of the 5 cases as follows:

1. $R_A = \frac{2}{3}(p_a - p_a^2)$,
   $R_B = p_{b1}\min(\frac{1-p_{b1}}{2}, \frac{1-p_a}{6}) + p_{b2}\min(\frac{1-p_{b2}}{2}, \frac{1}{2}\min(\frac{1-p_{b1}}{2}, \frac{1-p_a}{6}))$.

2. $R_A = p_a\frac{2}{3}\min(1-p_a, \frac{1-p_{b1}}{2})$,
   $R_B = \frac{p_{b1}(1-p_{b1})}{2} + p_{b2}\frac{1}{2}\min(1-p_{b2}, \frac{2}{3}\min(1-p_a, \frac{1-p_{b1}}{2}))$.

3. $R_A = p_a\frac{2}{3}\min(1-p_a, \frac{1}{2}\min(1-p_{b2}, \frac{1-p_{b1}}{2}))$,
   $R_B = \frac{p_{b1}(1-p_{b1})}{2} + p_{b2}\frac{1}{2}\min(1-p_{b2}, \frac{1-p_{b1}}{2})$.

4. $R_A = \frac{1}{2}(p_a - p_a^2)$,
   $R_B = \frac{p_{b1}(1-p_{b1})}{4} + p_{b2}\frac{1}{2}\min(1-p_{b2}, \frac{3(1-p_{b1})}{4})$.

5. $R_A = p_a\frac{1}{2}\min(1-p_a, \frac{1-p_{b1}}{2})$,
   $R_B = \frac{p_{b1}(1-p_{b1})}{2} + p_{b2}\frac{1}{4}\min(1-p_{b2}, \frac{1-p_{b1}}{2})$.

Let R_A^* and R_B^* be the optimal revenue for each firm and p_a^*, p_{b1}^*, and p_{b2}^* be the corresponding prices. Then the values for each of the 5 cases are summarized in table 3.2.

Based on the results, we try to find the Nash equilibrium price vector $(\bar{p}_a, \bar{p}_{b1}, \bar{p}_{b2})$ such that

$$R_A(\bar{p}_a, \bar{p}_{b1}, \bar{p}_{b2}) \geqslant R_A(p_a, \bar{p}_{b1}, \bar{p}_{b2}) \quad \forall p_a$$

and

$$R_B(\bar{p}_a, \bar{p}_{b1}, \bar{p}_{b2}) \geqslant R_B(\bar{p}_a, p_{b1}, p_{b2}) \quad \forall p_{b1}, p_{b2}$$

However, one can verify that there are no such vector and conclude that there exist no Nash equilibrium with three prices case.

Table 3.2: Optimal revenues and prices for firms A and B (three prices)

Case	p_a^*	p_{b1}^*	p_{b2}^*	R_A^*	R_B^*
1)	1/2	5/6	11/12	0.167	0.108
2)	3/4	1/2	5/6	0.125	0.194
3)	17/20	2/5	7/10	0.085	0.225
4)	1/2	1/2	5/8	0.125	0.180
5)	3/4	1/2	3/4	0.090	0.172

Table 3.3: User subscription for firms A and B (three prices)

Case	x_a^*	x_{b1}^*	x_{b2}^*
1)	1/3	1/12	1/24
2)	1/6	1/4	1/12
3)	1/10	3/10	3/20
4)	3/8	3/8	3/16
5)	3/16	1/4	3/16

3.3.2 Four prices

We now consider the situation where the two firms both go for PMP scheme. In other words, firm A provides a differentiated service with two prices p_{a1} and $p_{a2}(> p_{a1})$ and, similarly, firm B divides the network into two subnetworks with prices p_{b1} and $p_{b2}(> p_{b1})$. Now, due to the symmetry, we need only consider 6 cases which are

1. $p_{b1} < p_{b2} < p_{a1} < p_{a2}$.
2. $p_{b1} < p_{a1} < p_{b2} < p_{a2}$.
3. $p_{a1} < p_{b1} < p_{b2} < p_{a2}$.
4. $p_{b1} = p_{a1} < p_{b2} < p_{a2}$.

5. $p_{b1} < p_{a1} < p_{b2} = p_{a2}$.

6. $p_{b1} = p_{a1} < p_{b2} = p_{a2}$.

And one can find the revenue functions of both firms for the three cases as follows:

1. $R_A = p_{a2}\frac{1}{2}\min(1-p_{a2}, x_{a1}) + p_{a1}\min(\frac{1-p_{a1}}{2}, \frac{1-p_{b2}}{4})$,
   $R_B = p_{b2}\frac{1-p_{b2}}{2} + p_{b1}\frac{1-p_{b1}}{2}$.

2. $R_A = p_{a2}\frac{1}{2}\min(1-p_{a2}, x_{b2}) + p_{a1}\min(\frac{1-p_{a1}}{2}, \frac{x_4}{2})$,
   $R_B = p_{b2}\frac{1}{2}\min(1-p_{b2}, x_2) + p_{b1}\frac{1-p_{b1}}{2}$.

3. $R_A = p_{a2}\frac{1}{2}\min(1-p_{a2}, x_{b2}) + p_{a1}\frac{1-p_{a1}}{2}$,
   $R_B = p_{b2}\frac{1}{2}\min(1-p_{b2}, x_{b1}) + p_{b1}\frac{1}{2}\min(1-p_{b1}, x_{a1})$.

4. $R_A = p_{a2}\frac{1}{2}\min(1-p_{a2}, x_{b2}) + p_{a1}(\frac{1-p_{a1}}{3})$,
   $R_B = p_{b2}\frac{1}{2}\min(1-p_{b2}, x_{b1}) + p_{b1}\frac{1-p_{b1}}{3}$.

5. $R_A = p_{a2}\frac{1}{3}\min(1-p_{a2}, x_{a1}) + p_{a1}\min(\frac{1-p_{a1}}{2}, \frac{x_4}{2})$,
   $R_B = p_{b2}\frac{1}{3}\min(1-p_{b2}, x_2) + p_{b1}\frac{1-p_{b1}}{2}$.

6. $R_A = p_{a2}\frac{1}{3}\min(1-p_{a2}, x_{a1}) + p_{a1}(\frac{1-p_{a1}}{3})$,
   $R_B = p_{b2}\frac{1}{3}\min(1-p_{b2}, x_{b1}) + p_{b1}\frac{1-p_{b1}}{3}$.

where x_{a2}, x_{a1}, x_{b2}, and x_{b1} are the number of users in the networks in which the prices are p_{a2}, p_{a1}, p_{b2}, and p_{b1} divided by the capacity respectively and are as follows:

$$x_{b1} = \frac{1}{2}(1-p_{b1}),\ x_{b2} = \min(\frac{1-p_{b2}}{2}, \frac{x_{b1}}{2}).$$
$$x_{a1} = \min(\frac{1-p_{a1}}{2}, \frac{x_{b2}}{2}),\ x_{a2} = \min(\frac{1-p_{a2}}{2}, \frac{x_{a1}}{2}).$$

Table 3.4: Optimal revenues and prices for firms A and B (four prices)

Case	p^*_{a2}	p^*_{a1}	p^*_{b2}	p^*_{b1}	R^*_A	R^*_B
1)	0.93	0.86	0.7	0.4	0.0927	0.225
2)	0.93	0.75	0.88	0.5	0.1217	0.1778
3)	0.94	0.5	0.88	0.76	0.1532	0.144
4)	0.84	0.5	0.67	0.5	0.1505	0.1943
5)	0.86	0.7	0.86	0.4	0.1453	0.2007
6)	0.67	0.5	0.67	0.5	0.1578	0.1578

Table 3.5: User subscription for firms A and B (four prices)

Case	x^*_{a2}	x^*_{a1}	x^*_{b2}	x^*_{b1}
1)	7/200	7/100	3/20	3/10
2)	3/100	1/8	3/50	1/4
3)	3/100	1/4	3/50	3/25
4)	1/12	1/3	1/6	1/3
5)	1/10	3/20	1/10	3/10
6)	2/9	1/3	2/9	1/3

For each of the three cases, the optimal revenues (R^*_A, R^*_B) and corresponding prices (p^*_{a2}, p^*_{a1}, p^*_{b2}, p^*_{b1}) are shown in table 3.4.

Similar to the three prices case, we observe that there are no Nash equilibrium price vector for this case. Thus, in the duopoly model, we conclude that there are no equilibria based on the solution concept of Nash equilibrium when both firms use PMP[2]. Moreover, it is interesting to notice that even when both firms decide to provide services with single pricing there are no

[2]We have analyzed the cases with the concept of ϵ-equilibrium as well but the result did not change

Nash equilibria [98].

3.3.3 Discussion

It is also interesting to notice that the result is model-dependent. In a related literature, Gibbens et al. [34] showed that if there are more than one firm (in the market) then they do not differentiate their networks in equilibrium. That is, in a duopoly setting, the result which states that there is no equilibrium with PMP is same as ours but the fact that there is indeed an equilibrium with single prices is different.

The nonexistence of equilibria does not mean that there is no incentive to differentiate the prices in the real-world duopoly setting[3]. For instance, assume there are two providers each dividing the market share equally with single prices. If firm A's price (p_A) is lower than firm B's price (p_B), clearly, firm A will increase its profit by introducing another price, say p_{A2}, around p_B to take away some of firm B's market share. On the other hand, firm B will introduce a lower price, say p_{B2}, around p_A to maximize its profit. Therefore, in reality, we might observe the situation where there are 2 providers each with 2 prices.

3.4 Summary

In this section, we developed a model for Paris Metro Pricing strategy and demonstrated the profit incentive for a NSP to use PMP in a variety of scenarios. In particular, we analyzed the consumer behavior under PMP by allowing the model to define each users' condition which, when satisfied, they opt for joining the network. We evaluated the revenue of using PMP when there is a single provider (a monopolist) and determined the optimal fraction

[3]Here we assume that, in the real world, ISPs cannot change its level of service price freely over time due to political and social reasons.

of the two subnetworks to be divided in order to maximize the profit. We have seen that using PMP, indeed, increases both the revenue and subscription. Also, we have looked at a competition setting where two NSPs provide PMP. As it turned out, we noticed that in a duopoly case there exist no (pure) Nash equilibrium even when the duopolists go for single pricing.

4. Token Pricing

Researchers have proposed many pricing schemes for the Internet [62, 76, 48]. *Flat-rate pricing* is the most widely used scheme. The users pay a fixed monthly fee and get an unmetered access. The main advantages of this model are its low administrative and billing costs for the ISP as well as predictability for the users.

An *usage-based pricing* scheme charges users based on the amount of data they use [2]. Rates may vary depending on the time of day to encourage a smoother utilization of the available bandwidth resources [8]. However, such a scheme requires ISPs to monitor accurately each user's utilization and to implement a detailed billing scheme based on these measurements. Users face a variable bill and may get discouraged to use the network.

As more and more Internet services require a different level of quality of service (QoS), numerous techniques have been proposed to provide differentiated service levels. Some of these proposals use pricing mechanisms to improve the economic efficiency [62, 91]. However, many of these schemes are complicated and/or involve substantial costs in both development and operations [76].

In their recent analysis, Schwartz et. al. [90] show that the transition to multiple service classes is socially desirable but may increase the cost for some users for the same level of service. To limit this undesirable effect, they propose a regulation which sets an upper bound on the fraction of the network capacity reserved for high priority service. They show that this regulation can limit the distributional consequences while enabling providers to increase their revenue. However, this scheme uses differentiated pricing,

Table 4.1: Summary of Notations for Token Pricing

Symbol	Meaning
s	number of tokens a user is currently holding
$V(s)$	maximal total expected discounted value (fixed congestion)
p_τ	congestion in service 1 and steady-state probability (discrete)
$V(s, p_\tau)$	maximal total expected discounted value (variable congestion)
x_τ	pure service 1 utility
$g(x_\tau, p_\tau)$	service 1 value
β_τ	discount factor
$a_\tau(s)$	threshold value above which service 1 is consumed (discrete)
γ	threshold value (continuous)
K	number of tokens required for service 1 (discrete)
λ	application arrival rate (continuous)
M	number of tokens obtained per day (continuous)
$R_\tau(M, \lambda)$	value per token (continuous)
C	network capacity
α_τ	fraction of service 1 capacity
$p_D(p_N)$	day (night) time congestion (single class)
$g_D(g_N)$	day (night) time utility (single class)

which requires traffic monitoring and may face resistance from users and providers.

In this section, we analyze a pricing scheme called 'Token Pricing' that has the advantages of flat-rate pricing, yet promotes a more efficient usage of the network resources.

4.1 Token Pricing Scheme

When using *token pricing*, users pay a fixed monthly fee for Internet access. Each user receives a number of *tokens*. The provider offers a high-quality Service 1 and a normal-quality Service 2. For instance, the services could correspond to different "bearers" (GBR [1] or non-GBR) in an LTE network [22].

Service 2 requires no tokens but may become congested. On the other hand, Service 1 requires some tokens and, consequently, is likely to be less congested than Service 2, thus offering a better quality by a Paris Metro pricing effect [76]. Since the number of tokens is limited, users have an incentive to use Service 1 when they derive a sufficiently higher utility from that service. Consequently, we can expect the social welfare of the network to increase since valuable resources get used for more valued applications.

A secure agent in the user's browser can monitor the token usage so that the provider does not face much additional complexity. Also, the provider does not need additional billing mechanisms.

Below, we model the token pricing scheme and the behavior of users and we analyze the improvement in social welfare. For simplicity, the analysis proceeds in three steps. We first consider discrete time models and explores the optimal strategy for a user who maximizes his total discounted utility. The next section ignores the effect of users on the congestion of the network. And the following sections extends the analysis to include the congestion effect. Lastly we study a continuous time model for a user who maximizes his long-term average utility. That section builds on the results of the previous two.

[1] GBR: Guaranteed Bit Rate

4.1.1 Fixed Congestion

This section studies the behavior of a user facing token pricing. For simplicity, the model is in discrete time. Time is divided in epochs, such as ten-minute intervals for instance. During each epoch, the user accesses the Internet once for an application such as a file download, a voice or video call, an email session or web browsing. The added value for the user of using the high-quality Service 1 compared to using Service 2 depends on the application.

With the token pricing scheme, each user can use Service 1 only once every K epochs, on average. The user must choose when to use Service 1. The intuition is that the user will use Service 1 only when he benefits sufficiently from that valuable service. For instance, the user may choose Service 1 for an important video call instead of wasting his tokens for a file download that can be delayed. Consequently, for a given level of congestion of Service 1, one may expect the users to benefit more from that service than under a flat pricing scheme. The analysis confirms that intuition.

Every epoch n, each user gets 1 token and faces some application whose utility is $x_{\tau n}$ higher for Service 1 than for Service 2. The $x_{\tau n}$'s are independent and identically distributed random variables. It costs K tokens per epoch to use Service 1.

Assume that the user has s tokens. The maximal total expected discounted value $V(s)$ for the user of having s tokens satisfies the following dynamic programming equation:

$$V(s) = E[\max\{\beta_\tau V(s+1), x_{\tau n} + \beta_\tau V(s+1-K)\}], \tag{4.1}$$

where $V(s) = -\infty$ for $s < 0$ (so that the user cannot use tokens when he has fewer than K). In this expression $\beta_\tau \in (0,1)$ is the discount factor.

The optimal policy is then to use K tokens if and only if

$$x_\tau + \beta_\tau V(s+1-K) > \beta_\tau V(s+1), \tag{4.2}$$

i.e., if and only if

$$x_\tau > a_\tau(s) := \beta_\tau[V(s+1) - V(s+1-K)]. \tag{4.3}$$

Thus, as expected, the user chooses to use Service 1 only for applications that benefit sufficiently from that service. As the next result shows, the user is more likely to use Service 1 when he has accumulated many tokens, which is not surprising.

Theorem 4.1.1 *$a_\tau(s)$ is nonincreasing.*

Proof: Let

$$V_{n+1}(s) = E[\max\{x_\tau + \beta_\tau V_n(s+1-K), \beta_\tau V_n(s+1)\}],$$

with $V_0 = 0$. Then one can show that $V_n(s) \to V(s)$ for all s as $n \to \infty$.

We want to show that

$$\Delta_n(s) := V_n(s+1) - V_n(s+1-K)$$

is nonincreasing in s.

We do this by induction in n. Assume this is true for n. We show for $n+1$.

Note that

$$\begin{aligned} V_{n+1}(s) &= \beta V_n(s+1-K) \\ &\quad + E[\max\{x_\tau, \beta_\tau V_n(s+1) - \beta_\tau V_n(s+1-K)\}] \\ &= \beta_\tau V_n(s+1-K) + E[\max\{x_\tau, \beta_\tau \Delta_n(s)\}]. \end{aligned}$$

Thus,

$$\begin{aligned} \Delta_{n+1}(s) &= \beta_\tau V_n(s+2-K) + E[\max\{x_\tau, \beta_\tau \Delta_n(s+1)\}] \\ &\quad - \beta_\tau V_n(s+2-2K) - E[\max\{x_\tau, \beta_\tau \Delta_n(s+1-K)\}] \end{aligned}$$

$$= \beta_\tau \Delta_n(s{+}1{-}K){+}E[\max\{x_\tau, \beta_\tau \Delta_n(s{+}1)\}{-}\max\{x_\tau, \beta_\tau \Delta_n(s{+}1{-}K)\}].$$

Now, as s increases, the Δ_ns in this expression do not increase. Consider, for $a \leqslant b$,

$$\phi(a,b,x) := b + \max\{x,a\} - \max\{x,b\}$$
$$= \begin{cases} a, & \text{if } x < a \\ x, & \text{if } a \leqslant x \leqslant b \\ b, & \text{if } b < x. \end{cases}$$

Let $a' \leqslant a$ and $b' \leqslant b$. It is clear from the above that

$$\phi(a',b',x) \leqslant \phi(a,b,x),\ \forall x$$

Hence,

$$E[\phi(a',b',x)] \leqslant E[\phi(a,b,x)].$$

Consequently, for any $s < s'$, by letting

$$a = \beta_\tau \Delta_n(s+1),\ b = \beta_\tau \Delta_n(s+1-K),$$

$$a' = \beta_\tau \Delta_n(s'+1),\ b' = \beta_\tau \Delta_n(s'+1-K),$$

we conclude that

$$\Delta_{n+1}(s') \leqslant \Delta_{n+1}(s).$$

□

4.1.2 Variable Congestion

In this section, we consider that the value of Service 1 for an application characterized by x_τ as before is $g(x_\tau, p_\tau)$, where p_τ is the congestion level of Service 1. The interpretation is that, as it gets more congested, Service 1 becomes less valuable compared to Service 2. Thus, $g(x_\tau, p_\tau)$ is decreasing in p_τ and nondecreasing in x_τ.

The maximal total expected discounted value $V(s, p_\tau)$ for a user facing a Service 1 with congestion level p_τ when he has s tokens satisfies the following dynamic programming equations:

$$V(s, p_\tau) = E[\max\{\beta_\tau V(s+1, p_\tau), g(x_\tau, p_\tau) + \beta_\tau V(s+1-K, p_\tau)\}]. \quad (4.4)$$

The congestion level p_τ is the fraction of users who use Service 1. Thus, the value of p_τ is a function of the behavior of the users and we will consider that p_τ is the probability that a given user uses Service 1. The interpretation is that there are many users in the system and that each user faces the congestion that many other users generate for Service 1. By himself, one user has a negligible influence on the congestion level of Service 1. Thus, this model can be thought of as a mean-field limit of the system when the number of users increases.

To analyze this model numerically, we need an estimate of p_τ. This requires an estimate of the distribution of s. Now, s is a Markov chain that increases by 1 with probability $1 - p_\tau$ and decreases by $K - 1$ otherwise. The state-transition diagram of this Markov chain is shown in Figure 4.1.

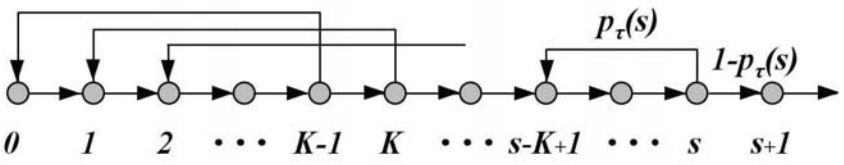


Figure 4.1: The Markov chain $\{s_n, n \geqslant 0\}$

The transition probability matrix of this Markov chain is such that, for

$s \geqslant K$,

$$p_\tau(s, p_\tau) = P(s, s-K+1, p_\tau) = 1 - P(s, s+1, p_\tau)$$
$$= P[\beta_\tau V(s+1, p_\tau) < g(x_\tau, p_\tau) + \beta_\tau V(s+1-K, p_\tau)] =: P[x_\tau > a_\tau(s, p_\tau)].$$

Also, for $s \leqslant K-1$,

$$P(s, s+1, p_\tau) = 1.$$

The following result is similar to Theorem 4.1.1.

Theorem 4.1.2 *In the case that $g(x_\tau, p_\tau)$ is nondecreasing in x_τ and nonincreasing in p_τ, $a_\tau(s, p_\tau)$ is nonincreasing in s.*

Proof: Let $g_p(x_\tau) = g(x_\tau, p_\tau)$. We assume the function g_p is nondecreasing in x_τ and nonincreasing in p_τ. For a fixed p_τ, if g_p is monotonically increasing

$$P[\beta_\tau V(s+1, p_\tau) < g(x_\tau, p_\tau) + \beta_\tau V(s+1-K, p_\tau)]$$
$$= P[g(x_\tau, p_\tau) > \beta_\tau V(s+1, p_\tau) - \beta_\tau V(s+1-K, p_\tau)]$$
$$= P[g_p(x_\tau) > \tau(s, p_\tau)] = P[x_\tau > g_p^{-1}(\tau(s, p_\tau))]$$
$$= P[x_\tau > a_\tau(s, p_\tau)].$$

where, $\tau(s, p_\tau) = \beta_\tau V(s+1, p_\tau) - \beta_\tau V(s+1-K, p_\tau)$.

Hence $a_\tau(s, p_\tau) = g_p^{-1}(\tau(s, p_\tau))$. Since $\tau(s, p_\tau)$ is nonincreasing in s for a fixed p_τ (Theorem 1) and g_p^{-1} is an increasing function, $a_\tau(s, p_\tau)$ is nonincreasing in s.

Next if g_p is nondecreasing but not monotonically increasing we have

$$P[g_p(x_\tau) > \tau(s, p_\tau)] = P[x_\tau > \max\{x_\tau | g_p(x_\tau) = \tau(s, p_\tau)\}].$$

Hence $a_\tau(s, p_\tau) = \max\{x_\tau | g_p(x_\tau) = \tau(s, p_\tau)\}$. Then we see that for $s_1 < s_2$

$$\tau(s_1, p_\tau) \geqslant \tau(s_2, p_\tau) \Leftrightarrow a_\tau(s_1, p_\tau) \geqslant a_\tau(s_2, p_\tau).$$

□

As we discussed earlier, p_τ is the probability that a user uses Service 1. If we knew p_τ, we could calculate the transition probabilities. These correspond to some stationary distribution for s. One could then determine p_τ from the following fixed point equation:

$$p_\tau = E[p_\tau(s, p_\tau)]. \tag{4.5}$$

The following theorem shows that the fixed point exists and is unique.

Theorem 4.1.3 *$E[p(s,p)]$ has a unique fixed point $p = \frac{1}{K}$.*

Proof: Let $\pi(s, p_\tau)$ be the stationary distribution of the Markov chain for a given congestion p_τ. Assume that the Markov chain starts in some state s_0 and let τ_n be the n-th time that the Markov chain returns to s_0. The number η_n of times that the user uses K tokens during $\{0, 1, \ldots, \tau_n\}$ must be such that $K\eta_n = \tau_n$. Thus, if $\eta_n(s)$ is the number of times during $\{0, 1, \ldots, \tau_n\}$ that the Markov chain is in state s and that the user uses tokens, then

$$\frac{1}{K} = \frac{\eta_n}{\tau_n} = \sum_s \frac{\eta_n(s)}{\tau_n} \to \sum_s \pi(s, p_\tau) p_\tau(s, p_\tau) = E[p_\tau(s, p_\tau)] \text{ as } n \to \infty.$$

where the limit follows from the ergodicity of the Markov chain. □

4.1.3 Continuous Time

In this section, we consider that a user gets M tokens per day and it costs 1 token for using Service 1. Applications arrive as a Poisson process with rate λ during the day where each has a utility that is higher by x_τ for using Service 1 instead of Service 2. We assume that x_τ is an exponentially distributed random variable with mean 1 and is independent across applications.

The user can hold his tokens and use them whenever he wants. To maximize the long-term average value of his tokens, the user adopts a stationary Markov policy that is of a threshold type. This follows from the fact that the

optimal policy for a total discounted cost is of that type and that the policy converges to the optimal long-term average cost policy as the discount factor β_τ goes to 1, as shown in [86].

As a result, the user uses one token if $g(x_\tau, q_1) > \gamma$ where

$$P[g(x_\tau, q_1) > \gamma] = \frac{M}{\lambda} = \exp\{-\widehat{\gamma}\}, \tag{4.6}$$

where $q_1 = M/c$ is the congestion level in Service 1, and $\widehat{\gamma} = g_{q_1}^{-1}(\gamma)$.

The value per token (i.e., the expected value of a single application in token scheme) is then

$$R_\tau(M, \lambda) := E[g(x_\tau, q_1) : g(x_\tau, q_1) > \gamma] = E[g(x_\tau, q_1) : X > \widehat{\gamma}]. \tag{4.7}$$

Then the total value generated by the token scheme is

$$MR_\tau(M, \lambda).$$

4.2 Numerical Analysis

In this section, we provide some numerical results to see how the value of the token scheme changes with respect to the variables K or M that determine the cost of Service 1.

4.2.1 Discrete Epochs

To see how the value differs with respect to the number of tokens required to use Service 1, we set up a numerical experiment with a two specific utility functions as follows:

$$g_1(x_\tau, p_\tau) = x_\tau - \eta p_\tau, \quad g_2(x_\tau, p_\tau) = \frac{x_\tau}{\eta(p_\tau + 1)} - \epsilon_\tau. \tag{4.8}$$

In these expressions, η denotes the number of users in a given network. In other words, since p_τ denotes the fraction of users using service 1, ηp_τ represents the actual congestion in the network that affects the user utility. Here,

$\epsilon_\tau > 0$ makes the utility negative when the congestion is severe. Without that term, the utility g_2 would always be positive, independently of the congestion level.

Figure 4.2(a), 4.2(b) shows the expected value of the token pricing scheme with 5 different congestion levels (η). We see that when the number of users in the network is low the overall value is high since the congestion disutility is low. Also there exists some $K = K^*$ that produces the highest value when the network is not under-utilized. However, when ones increases K beyond K^*, the value decreases since users get to use Service 1 for fewer applications.

4.2.2 Continuous Time

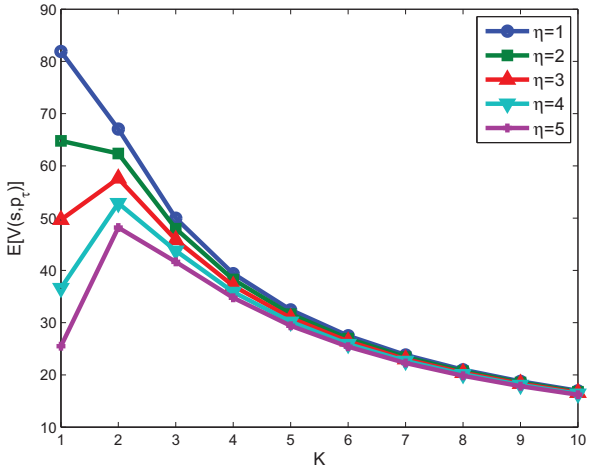
For the continuous time model we consider the following utility function $g_3(x, q)$:

$$g_3(x_\tau, q) = x_\tau - q. \tag{4.9}$$

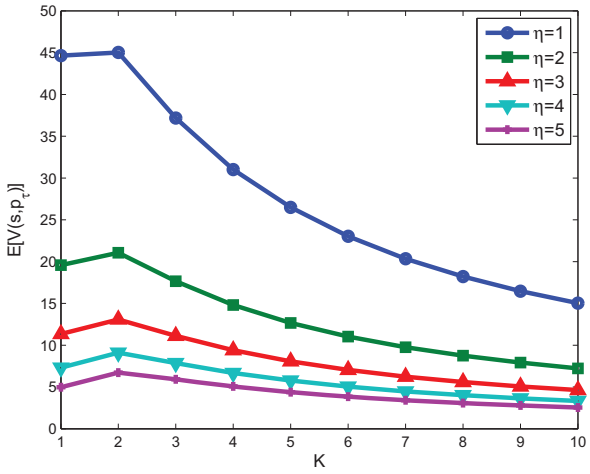
Then we have

$$R_\tau(M, \lambda) = 1 - \frac{M}{c} + \max\{0, \log \frac{\lambda}{M}\}. \tag{4.10}$$

Figure 4.3 illustrates the value $MR_\tau(M, \lambda)$ of the token scheme with respect to M for various demand values λ. We observe that there is an optimal number M of tokens that maximizes the user utility. In addition, we see that when user demand increases it is better to let users use premium service more (higher optimal M) as long as the total utility surplus produced by using multiple of files in service 1 is larger than the disutility due to congestion. Therefore, if the provider increases M larger than the optimal value given fixed demand (λ), the value tends to decrease as the disutility coming from congestion increases. Thus operators who adopt the token scheme should adjust the number of tokens.



(a)



(b)

Figure 4.2: Values of Token pricing scheme with respect to K: (a) $g_1(x_\tau, p_\tau)$, (b) $g_2(x_\tau, p_\tau)$

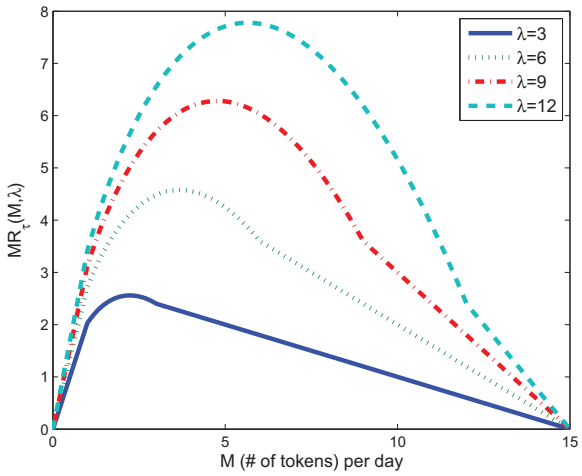


Figure 4.3: Values of Token pricing scheme with respect to M: $g_3(x_\tau, q)$

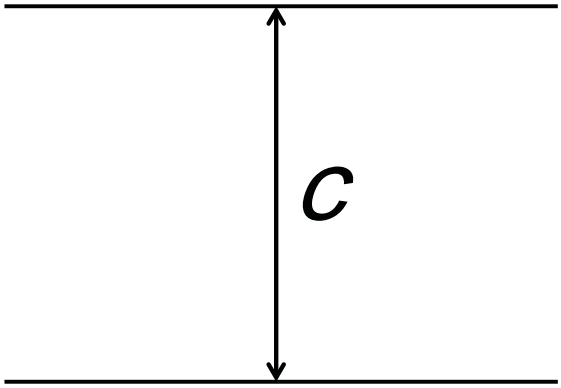
4.3 Flat vs Token

In the previous section, we have analyzed the proposed token pricing scheme and illustrated its characteristics. We observed that it is simple for both the provider and the users. In addition, we have seen that the token scheme increases user welfare by letting users accumulate tokens and use them only for relatively more valuable applications.

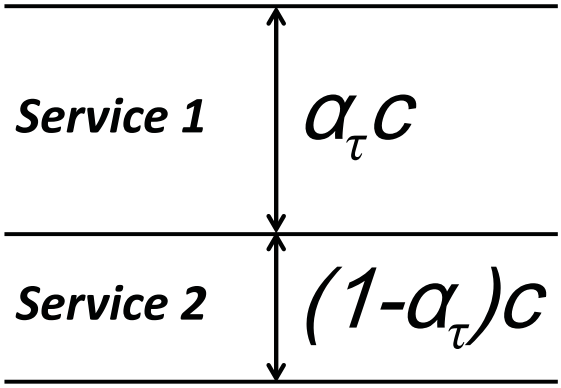
In this section we validate our proposed scheme by comparing it with flat-rate pricing. Each day, the user uses λ applications, on average, and each is of type x_τ, an exponentially distributed random variable with mean κ, independent across applications. In all cases, we use the same model for the increase $g(x_\tau, q)$ in utility of using Service 1 instead of Service 2 when the congestion level of Service 1 is q and the application is of type x_τ:

$$g(x_\tau, q) = A - x_\tau q. \tag{4.11}$$

In this expression, A denotes a positive benefit for receiving Internet service,



(a) Flat



(b) Token

Figure 4.4: Network capacity for Flat and Token pricing scheme with two service classes

which is independent of the pricing scheme.

First, consider flat-rate pricing. Assume that the congestion level of the flat-rate network is q_f. The total average daily value created by flat-pricing

is

$$W(\lambda) = \lambda E[g(x_\tau, q_f)] = \lambda(A - \kappa q_f), \tag{4.12}$$

since there are λ applications per day, on average.

For the token scheme, the operator chooses the portion α_τ of capacity to be allocated for Service 1 (respectively, $1 - \alpha_\tau$ for Service 2). The user decides whether to use Service 1 or Service 2 depending on the application type x_τ. Assume that the congestion level of Service i is q_i, for $i = 1, 2$, with $q_1 < q_2$. Then, as we saw in Section 4.1.3, the user uses Service 1 if

$$g(x_\tau, q_1) - g(x_\tau, q_2) = x_\tau(q_2 - q_1) > \gamma, \tag{4.13}$$

where the threshold γ is such that

$$P[x_\tau(q_2 - q_1) > \gamma] = \exp\{-\frac{\kappa\gamma}{q_2 - q_1}\} = \frac{M}{\lambda}. \tag{4.14}$$

Solving this equation allows to derive γ. We find $\gamma = -\frac{q_2 - q_1}{\kappa} \log\{\frac{M}{\lambda}\}$. Note that γ decreases with the fraction M/λ of applications that can use Service 1.

Since the value of the token scheme comes from two services, the total average value is given by

$$S(M, \alpha_\tau, \lambda) = \underbrace{M.E[g(x_\tau, q_1)|x_\tau(q_2 - q_1) > \gamma]}_{Service\ 1\ utility} + \underbrace{(\lambda - M)E[g(x_\tau, q_2)|x_\tau(q_2 - q_1) < \gamma]}_{Service\ 2\ utility}.$$

The provider using the token scheme chooses the number M of tokens per day and the capacity ratio α_τ that maximizes the value to users. Thus the total value $R_\tau(\lambda)$ of using the token scheme is as follows:

$$R(\lambda) = \max_{M, \alpha_\tau} S(M, \alpha_\tau, \lambda). \tag{4.15}$$

The value of any pricing scheme would depend upon the service type the users are using. The service types could be classified into two categories [6]: capacity-sharing service and latency-based service.

4.3.1 Capacity-Sharing Service

First, we consider a generic class of services, in which the total demand in a network is distributed among the processors with some fixed total capacity (per user) c. Thus, we define the congestion functions as:

$$q_f = \frac{\lambda}{c}, \quad q_1 = \frac{M}{\alpha_\tau c}, \quad q_2 = \frac{\lambda - M}{(1-\alpha_\tau)c}. \tag{4.16}$$

This type of congestion function is considered in prior work [34]. The interpretation is that if N applications share a capacity C, then each application faces a congestion disutility N/C. Accordingly, the value of the token scheme is

$$\begin{aligned} S_{CS}(M, \alpha_\tau, \lambda) \quad &= ME[g(\theta, q_1)|\Pi] + (\lambda - M)E[g(\theta, q_2)|\Pi^c] \\ &= M\left[A - \frac{\lambda\kappa}{\alpha_\tau c}\left(\frac{M}{\lambda}\right)^\kappa \left(1 + \log\left(\frac{\lambda}{M}\right)^\kappa\right)\right] \\ &+(\lambda - M)\left[A - \frac{\lambda\kappa}{(1-\alpha_\tau)c}\left\{1 - \left(\frac{M}{\lambda}\right)^\kappa \left(1 + \log\left(\frac{\lambda}{M}\right)^\kappa\right)\right\}\right]. \end{aligned}$$

where Π is the event that a user uses Service 1 as in (2).

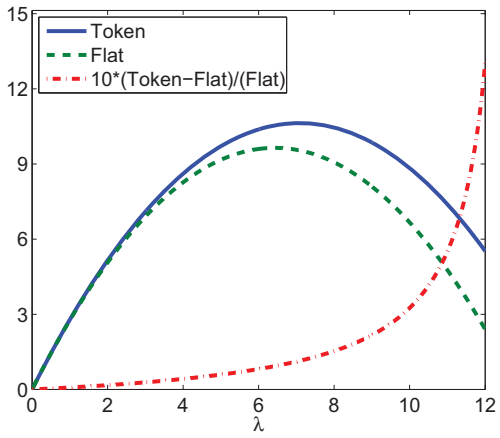
4.3.2 Latency-Based Services

In other services, latency may be a major concern. A simple model to capture the latency is by using an M/M/1 queue model [6]. Using this model, we define a congestion function for latency-based services as follows:

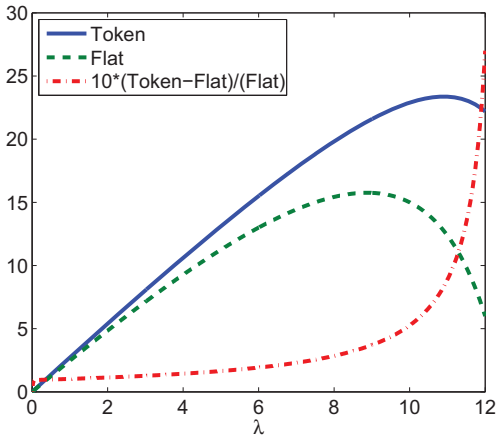
$$\text{Total expected delay} = \frac{1}{\mu - \lambda}$$

where λ is the arrival rate and μ is the service rate. This function captures the fact that when the network utilization gets close to the whole capacity, the delay goes to infinity. Then, assuming the service rate is the available capacity, the congestion functions for each of the services (Flat and Token) are

$$q_f = \frac{1}{c - \lambda}, \quad q_1 = \frac{1}{\alpha_\tau c - \lambda/M}, \quad q_2 = \frac{1}{(1-\alpha_\tau)c - \lambda(1 - 1/M)}. \tag{4.17}$$



(a) Capacity Sharing Service



(b) Latency Based Service

Figure 4.5: Comparison of Flat and Token pricing scheme with two service classes (A=3)

This type of congestion has been considered in prior work [85] as well. Hence the value of Token scheme in this type of service is

$$S_{LB}(M, \alpha_\tau, \lambda) = ME[g(\theta, q_1)|\Pi] + (\lambda - M)E[g(\theta, q_2)|\Pi^c]$$
$$= M\left[A - q_1\frac{\lambda\kappa}{M}\left(\frac{M}{\lambda}\right)^\kappa\left(1 + \log\left(\frac{\lambda}{M}\right)^\kappa\right)\right] \quad +$$
$$(\lambda - M)\left[A - \frac{q_2\lambda\kappa}{\lambda - M}\left\{1 - \left(\frac{M}{\lambda}\right)^\kappa\left(1 + \log\left(\frac{\lambda}{M}\right)^\kappa\right)\right\}\right]$$

Figure 4.7(a) and 4.7(b) show the numerical results for capacity sharing service and latency based service respectively. From the comparison, we see that, irrespective of service types, the token scheme always produces higher revenue than flat-rate pricing. Moreover, we observe that the surplus created by the token scheme should become more advantageous for future wireless Internet tariffing since, in the future, we expect (1) a drastic increase in wireless data traffic and (2) a much higher portion of delay-sensitive multimedia services.

4.4 Token Scheme with Single Class

In this section, we consider the Token scheme with single class instead of two service classes. We divide one day into two different time zones: peak time and off-peak time. We will use the notation 'Day' for peak time, and 'Night' for off-peak time since usually the network is more congested in day time. (However, this is just for the explanatory purpose and thus the peak/off-peak time could be some time segments during the day.)

Users pay a (fixed) monthly fee and are given some amount of tokens as before. For simplicity, we only consider the demand in Day time since these traffic contributes to the network overload while the network at Night is under-utilized. Unlike the previous case, using 'tokens,' here, means using the service right now (i.e., Day time) instead of at Night. Since the number of tokens is limited, users will use the tokens only when the utility gain of using that service at Day instead of at Night is sufficiently large, which will,

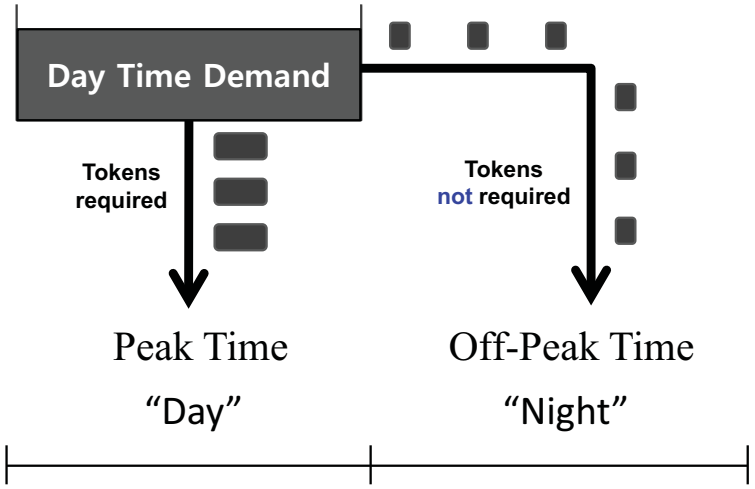


Figure 4.6: Token scheme with Single Class: A single day is divided into two time zones

possibly, lead to a load balancing of the network traffic between peak and off-peak hours.

Consider a case where user's decision is to use tokens and enjoy the service now (for their day time demand) and, if not, delay the service to Night time. Since he uses the service either at Day or Night, his decision to use tokens is based on the utility x_τ, the day time congestion p_D, and the night time congestion p_N, where $p_D + p_N = 1$. Here the utility x_τ is not an 'added' utility of using the service at Day, instead, it denotes the 'absolute' utility of the services that is supposed to use at Day time when Token scheme is not used (i.e., day time demand) and it is an independent and identically distributed random variable. As for the two-service classes case, a user gets one token per day and it requires K tokens to use the service at Day time. Thus the maximal total expected discounted value $V(s, p_D)$ for the user of

having s tokens facing the congestions in both time period (i.e., p_D and p_N) satisfies the following dynamic programming equations:

$$V(s, p_D) = E[\max\{g_N(x_\tau, p_N)+\beta_\tau V(s+1, p_D), g_D(x_\tau, p_D)+\beta_\tau V(s+1-K, p_D)\}].$$

$g_D(x_\tau, p)$ denotes the perceived Day time utility function where the service type of x_τ is used at Day time facing the congestion p_τ. Likewise, $g_N(x_\tau, p_\tau)$ is the perceived night time utility function. To represent the higher utility for using at Day time we have

$$\frac{\partial g_D(x_\tau, p_\tau)}{\partial x_\tau} > \frac{\partial g_N(x_\tau, p_\tau)}{\partial x_\tau}. \tag{4.18}$$

Also, as in the two-class case, the perceived utility functions $g_D(x_\tau, p_D)$ and $g_N(x_\tau, p_N)$ is decreasing in congestion p_D and p_N respectively, i.e.,

$$\frac{\partial g_D(x_\tau, p_D)}{\partial p_D} < 0, \quad \frac{\partial g_N(x_\tau, p_N)}{\partial p_N} < 0. \tag{4.19}$$

The optimal policy is to use K tokens if and only if,

$$g_N(x_\tau, p_N) + \beta_\tau V(s+1, p_D) < g_D(x_\tau, p_D) + \beta_\tau V(s+1-K, p_D)$$

$$\Leftrightarrow \beta_\tau\{V(s+1, p_D) - V(s+1-K, p_D)\} < g_D(x_\tau, p_D) - g_N(x_\tau, p_N)$$

$$\Leftrightarrow \beta_\tau\{V(s+1, p_D) - V(s+1-K, p_D)\} < \widehat{g}(x_\tau, p_D) \quad \Leftrightarrow \quad x_\tau > \widehat{a_\tau}(s, p_D),$$

where we define the function $\widehat{g}(x_\tau, p_D) := g_D(x_\tau, p_D) - g_N(x_\tau, p_N)$ and $\widehat{a_\tau}(s, p_D) = \widehat{g}_{p_D}^{-1}(\beta_\tau\{V(s+1, p_D) - V(s+1-K, p_D)\})$.

Now the optimal policy depends on two variables x_τ and p_D instead of three (x_τ, p_D, p_N), since $p_D + p_N = 1$. Also, $\widehat{g}(x_\tau, p_D)$ is increasing in x_τ and decreasing in p_D since $\frac{\partial \widehat{g}(x_\tau, p_D)}{\partial x_\tau} = \frac{\partial g_D(x_\tau, p_\tau)}{\partial p_D} - \frac{\partial g_N(x_\tau, p_\tau)}{\partial x_\tau} > 0$ from (4.18) and $\frac{\partial \widehat{g}(x_\tau, p_D)}{\partial p_D} = \frac{\partial g_D(x_\tau, p_\tau)}{\partial p_D} + \frac{\partial g_N(x_\tau, p_\tau)}{\partial p_N} < 0$ from (4.19).

Hence, the optimal policy is mathematically equivalent to the two-class service case with variable congestion (as in Sec. 2.2) except that the notation has changed from the congestion in Service 1, p_τ, to the congestion in Day time, p_D and from the utility function $g(x_\tau, p_\tau)$ to $\widehat{g}_(x_\tau, p_D)$. Thus, Theorems 2 and 3 hold for this case as well.

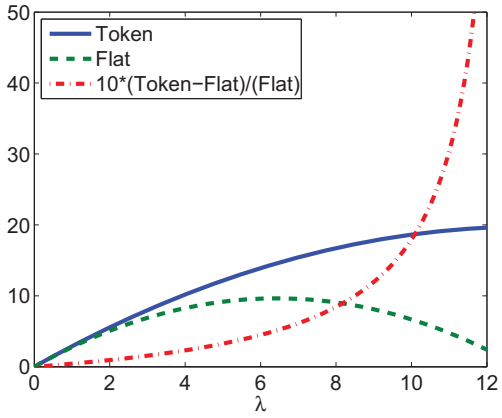
4.4.1 Comparison with Flat and Token with single service class

This section validates the Token Scheme with single service class by comparing with Flat pricing as well. Each day, the user has λ applications to use at peak time (i.e., Day time demand), on average, and each is of type x_τ, an exponentially distributed random variable with mean κ, independent across applications. Also it requires 1 token to use the service at Day time. For this scheme, using tokens means using the service at Day so each user will use M applications, on average, at Day time and the rest $\lambda - M$ at Night. The only difference with the two service classes case is that the capacity is not divided in any way since there is merely a single service. Thus the network capacity is c for both Day and Night.

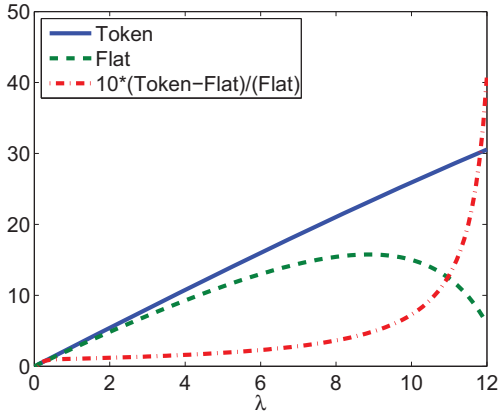
For Flat pricing, as there is no cost for using the service at Day, all the Day time demand is used at peak time. This means that all the Day time demand (λ) is used at peak time and none at night. From Fig. 4.7 we see that, similar to the two service classes case, Token sheme is superior to Flat scheme for both capacity sharing service and latency based service.

4.5 Summary

The design of tariff schemes for packet-based network services should not only take into account various aspects to make it technologically feasible but also consider social issues from the perspective of users to be actually used in practice. Researchers have proposed numerous pricing schemes to provide QoS since, in the future, we expect a rapid increase in the applications that require high QoS. Nevertheless, most of the schemes are too complicated and, above all, unfavored by users. In this spirit, we proposed a novel scheme called, Token pricing scheme, in a way that is simple and predictive to be well appreciated by users, and also efficient to provide QoS requirements. We have shown for both single service and two service classes that, comparing with the



(a) Capacity Sharing Service



(b) Latency Based Service

Figure 4.7: Comparison of Flat and Token pricing scheme with single service class (A=3)

current dominant Flat-rate pricing, our proposed scheme gives higher utility to the users, and especially significant increase of utility surplus for Latency

based services, with relatively low cost of implementation.

5. Price of Simplicity

5.1 Flat-rate and Volume-based Pricing

The telecommunication industry has been continuously debating the benefits of the two practical pricing plans: flat-rate (a.k.a. *all-you-can-eat*) pricing and volume-based (a.k.a. *pay-as-you-go*) pricing. For broadband services, the industry has been dominated by flat-rate prices for its predictability and user acceptance. However, together with the skyrocketing interest in applications and smartphones, this has led to a mobile data traffic explosion where about 3 percent of users generate about 40 percent of data traffic [36]. Given the situation, NSPs have tried to address this disparity in data usage to ensure that everyone is able to share the benefits delivered by the services. As a feasible economic solution, providers have introduced a volume-based pricing due to several advantages: (1) It is efficient in terms of resource supply and utilization ; (2) the price is more closely linked to the cost to deliver the service (cost allocation); and (3) it provides opportunities to develop new services. Nevertheless, it is not as widely used as flat-rate pricing because of the most important drawback: *customers' strong resistance.* Despite customers' preference, network providers are considering to charge their services with volume-based pricing for their coming LTE (Long Term Evolution) network [30, 96]. One of the possible reason behind this movement is that the backbone traffic is growing dramatically but the revenue is not. Thus, it is of critical matter for the provider whether current flat-rate pricing is sustainable in the long run in terms of revenue.

This dissertation contributes on the network literature that analyzes the two pricing schemes: flat-rate and usage based pricing. In [21], the authors

Table 5.1: Comparison of Flat and Volume-based Pricing

	Flat-rate Pricing	Volume-based Pricing
Resource Efficiency	×	△
Supply Efficiency	×	△
Cost Allocation	×	◯
Individual QoS	×	×
New Service	×	△
Predictability	◯	△
User Acceptance	◯	×

conducted a real-life market and technology trial named the Internet Demand Experiment (INDEX) in an attempt to determine how much users value different qualities of service of Internet access. The subject group of INDEX consists of 70+ customers including students, faculty and staff of the University of California at Berkeley. By providing them a sequence of service plans, each participant selects a particular pricing model. The core findings are as follows: (1) ISPs offering flat-rate service create large social inefficiencies in the form of waste, inter-subscriber subsidies, and tiered service. (2) However, usage-based pricing with differentiated service quality provides gains to providers and most consumers, and promotes the diffusion of broadband access. Kesidis et al. [55] also compares the two pricing schemes using queueing theory. The difference between this work and ours comes from the modeling of a utility function. The user utility in [55] is modeled as follows

$$U_n(\underline{\rho}) \equiv \begin{cases} \rho_n & \text{if } \sigma_k \rho_k \leqslant \Lambda \\ \rho_n \exp\left(\frac{-\sigma_k \rho_k}{\Lambda \beta_n}\right) & \text{else.} \end{cases} \tag{5.1}$$

where $\underline{\rho}$ denotes the vector of (user n) traffic rates, ρ_n, and $\beta_n \geqslant 0$ is a user-

dependent parameter, and Λ is an effective limit on the offered load. The utility is merely a function of supported load without any price parameter to represent a user's disutility. Moreover, this setting does not allow to define a proper revenue function of an ISP. Thus, although they compare the two pricing schemes and conclude that usage-based pricing is better than flat-rate in terms of degree of overload, the result does not cover the economic relationship between the provider and users.

MacKie-Mason and Varian [62] describes the economic theory of pricing congestible resource. They explore the implications of flat pricing and usage-based pricing for capacity expansion in a diverse environments. The major characteristic of [62] that differs from ours is that they considered ISP revenue maximization in terms of both price and capacity. This setting where the provider has two variables for decision making can be interpreted as a case where the provider is considering to enter the market or a situation where the management makes a long-term decision. In their work capacity is a variable while it is a fixed (given) constant in our model. The rationale behind this is based on the fact that prices are set more frequently than the capacity. Capacity change (e.g., expansion) incurs reasonably more cost than to make price changes and thus it would be natural to make those decisions in sequence rather than simultaneously if dealt both at the same time. Instead our model considers a short-term decision of ISPs where the capacity is given and does not change for sufficiently long period of time.

Among related work the closest one is [93]. In effect our research is motivated by theirs. In [93], the authors have shown that the loss of revenue from using simple entry fee is small in a range of environments. In particular, the utility function (of user i) they used is given by $u_i(x) = \alpha_i u(x)$, where α_i is user i's valuation, and u is a concave, nondecreasing function. They also introduced "Price of Simplicity," a revenue loss metric defined by

$\inf_{\alpha>0} \frac{R_{MUP}}{R_S}$ which is basically the worst-case ratio of the revenue under flat price (R_{MUP}) to the maximum revenue that the provider can obtain under complete information (R_S). In particular, for a given vector $\alpha = \{\alpha_1, ..., \alpha_N\}$ with $\alpha_1 \geqslant \alpha_2 \geqslant \ldots \geqslant \alpha_N$, the maximum revenue that the service provider can obtain by posting a single entry price, denoted by $R_{MUP}(\alpha, u))$ is given by

$$R_{MUP}(\alpha, u) = \max_{i \in \mathcal{N}} \left\{ i\alpha_i u\left(\frac{C}{i}\right) \right\}. \tag{5.2}$$

where C is the bandwidth of the link. Also, $R_S(\alpha, u)$ can be obtained by solving the following optimization problem.

$$\begin{aligned} \text{maximize}_{x \geqslant 0} \quad & \textstyle\sum_{i \in \mathcal{N}} \alpha_i u(x_i) \\ \text{subject to} \quad & \textstyle\sum_{i \in \mathcal{N}} x_i \leqslant C. \end{aligned}$$

They provide several examples supporting their claim. For the utility function $u(x) = \log(1 + x)$, the expression for the PoS is given by

$$\frac{\log(2)}{\sum_{i=1}^{14} \frac{\log(2)}{i\log\left(1+\frac{1}{i}\right)} \log\left(\frac{15i\log\left(1+\frac{1}{i}\right)}{\sum_{j=1}^{14} \frac{\log(2)}{j\log\left(1+\frac{1}{j}\right)}} \right)}. \tag{5.3}$$

Evaluating this yields the value of 87.8012% which represents a low price of simplicity. Also, for the utility function $u(x) = 1 - e^{-x}$, the PoS value is 84.4756%, again showing a low price of simplicity (according to their definition). Although the results are based on a rigorous analysis, they have overlooked a typical feature inherent in data network: *congestion*. It is clear that the utility function above encompasses only the *utility* of enjoying the service and no *disutility* coming from the delay one experiences. One can expect that this delay disutility might be negligible in an underutilized network but can be substantial in case of severe congestion. As the data traffic

is expected to grow dramatically there is no doubt that the delay disutility would be a crucial factor in terms of user's *perceived* utility. Therefore, in this paper, we aim to reinvestigate the price of simplicity in a network under *congestion*. We first redefine the term 'price of simplicity' as follows:

Definition 5.1.1 [Price of Simplicity] *The price of simplicity (PoS) is defined as one minus the ratio of the revenue under flat price (R_F) to the revenue under two-part tariff (R_T); i.e.,*

$$\text{PoS} = 1 - \frac{R_F}{R_T}.$$

Note that, unlike [93], we compare the revenue difference between flat price and two-part tariff as they are the most prevailing pricing plans in practice. By definition, high PoS means the revenue loss induced by using flat price is large. We would like to specify in which context the PoS is high and what exacerbates this phenomenon.

5.2 Model

In this section, we examine the concept called the *price of simplicity*. In order to analyze this concept we look through the relationship between the ISP and end users. An ISP will choose either flat pricing or two-part tariff for their service towards users.[1] We will first consider an ideal case where all users are identical in their preferences. Then we examine a more practical scenario where users are heterogeneous in their service preference and their congestion sensitivity.

[1]For the ISP-CP relationship, a "network neutrality debate" is a crucial issue that is spread throughout most of the world. For the literatures on network neutrality refer to [75], [89], [4], [37], [56], [7], [99], [88], [95], [87], [74], [18], [97], [32], [43], [59]

Table 5.2: Summary of Related Work

Reference	Features	Limitations
[62]	Considered two-part tariff and flat price in a congestible network. Capacity investment is also incorporated in the model to derive the optimum capacity level. Showed that for flat price the equilibrium usage and capacity is larger than for two-part tariff. Finally, concluded that equilibrium congestion of flat price is higher.	Considered only one class of delay function $\frac{T}{C}$, total use divided by capacity. Analysis is based on fixed user base.
[39]	Similar model to [62]. Nonlinear pricing is used. Assumed that consumer preferences are private information unknown to the firm. Concluded that the level of overall congestion is too low due to incomplete information.	Also used only one class of delay as in [62] with a fixed user base.
[93]	Defined a concept called "Price of Simplicity," a lower bound for the ratio between the revenue from flat pricing and maximum revenue. Concluded that loss of revenue from using simple entry fee is small. Also compared with a nonlinear pricing scheme based on Paris Metro Pricing and showed the gain from the price discrimination is small.	Delay disutility is not modeled. No analysis regarding social welfare.
[21]	Performed INDEX trial which involved actual users in UC Berkeley. Found that flat-rate creates large social inefficiency while usage-based price provides gains to providers and consumers.	Non-analytical. Population is limited.
[55]	Studied flat-rate versus usage-based pricing under overload conditions. Showed that in overload scenarios usage-based pricing reduces congestion and increases user's perceived utilization.	No analysis with regards to revenue and user's willingness-to-pay.

5.2.1 Identical User Case

We consider a monopoly network service provider (NSP) of capacity C that provides network connection services to users of size N. We consider a representative user i whose perceived utility for transmitting a traffic of volume

x is $u(x;v)$, where u is an increasing concave function and v is a user type parameter. The NSP charges $p(x)$ for transmitting traffic volume x, which can be a flat or a two-part tariff. For the flat charge, $p(x) = f_\phi$ while $p(x) = a + \beta x$ for the two-part tariff.[2] The total traffic T generated by N users, is less than or equal to the capacity C or

$$T := \sum_{i=1}^{N} x_i \leqslant C, \tag{5.4}$$

where x_i is the traffic volume of user i. With an increasing traffic volume T, users experience network delays, which in turn cause disutility $d(T;C)$ to users. The delay disutility function is a nondecreasing function of T such that $d(T;C) \to \infty$ as $T \to C$.

User's Problem

The perceived net-utility of user i including delay disutility is expressed by

$$\tilde{u}(x_i) = u(x_i;v) - p(x_i) - d(T;C). \tag{5.5}$$

User i determines his or her data rate x_i^* to maximize their net-utility given by (5.5), or

$$(\mathbf{P_{user}}) \quad x_i^* = \arg \max_{0 \leqslant x_i \leqslant \bar{x}} [u(x_i;v) - p(x_i) - d(T;C)], \tag{5.6}$$

where $\bar{x}$ is the maximum amount of traffic that a user can generate due to limitations on the access line or on time. When the net-utility value is negative $\tilde{u}(x_i^*) < 0$, user i does not subscribe to the service and the data rate $x_i^* = 0$. Therefore, a necessary condition for the NSP to make its revenue positive is $p(x_i^*) \leqslant u(x_i^*;v) - d(T^*;C)$, where $T^* = \sum_i x_i^*$. Given tariff $p(x)$,

[2]Throughout this section we use subscripts $_F$ and $_T$ to denote flat price and two-part tariff, respectively.

the first order condition of the user problem is

$$u'(x_i^*; v) = p'(x_i^*) + d'(T; C) + \mu; \tag{5.7}$$

$$\mu(x_i^* - \bar{x}) = 0, \tag{5.8}$$

where μ is a nonnegative Lagrangian variable for inequality $x \leqslant \bar{x}$. Note that $T = x_i^* + \sum_{j \neq i} x_j^*$.

When $\mu > 0$, the traffic volume x_i^* is equal to the maximum amount $\bar{x}$. This corresponds to an unsaturated network with low traffic volume. The strict positive μ implies $u'(x_i^*; v) > p'(x_i^*) + d'(T; C)$. When $\mu = 0$, the traffic volume is determined such that $u'(x_i^*; v) = p'(x_i^*) + d'(T; C)$. This corresponds to a saturated network with congestion.

Provider's Problem

The problem of the NSP can be formulated as follows:

($\mathbf{P_{NSP}}$)

$$\text{maximize} \qquad R := \textstyle\sum_i p(x_i) \tag{5.9}$$

$$\text{subject to} \qquad \textstyle\sum_i x_i \leqslant C, \tag{5.10}$$

$$p(x_i) \leqslant u(x_i; v) - d(T; C), \forall i \tag{5.11}$$

$$x_i \leqslant \bar{x}, \forall i \tag{5.12}$$

Inequality (5.10) is the network capacity constraint and inequality (5.11) means a nonnegative net-utility for all users. The last inequality (5.12) is the upper bound of a user's traffic volume. The problem can be simplified from the observation that $p(x_i) = u(x_i; v) - d(T; C)$ should be satisfied to be optimal. Otherwise, the provider can increase the charge of user i by a small value to increase its revenue. Hence, we have a modified problem

Table 5.3: Summary of Notations for Price of Simplicity

Symbol	Meaning
C	network capacity
N	total number of users
T	total traffic
x	user data rate
$\bar{x}$	maximum user data rate
v	user type parameter
$u(x;v)$	perceived user utility
$\tilde{u}(x;v)$	user's net-utility
$p(x)$	user payment for using data rate volume of x
f_ϕ	fixed price (flat)
a	fixed payment (volume)
β	marginal price (volume)
$d(T;C)$	delay disutility
R	provider's revenue
α	elasticity parameter
ϵ	congestion sensitivity

$(\mathbf{P'_{NSP}})$

$$\text{maximize} \quad R = \textstyle\sum_i [u(x_i;v) - d(T;C)] \tag{5.13}$$

$$\text{subject to} \quad \textstyle\sum_i x_i \leqslant C, \tag{5.14}$$

$$x_i \leqslant \bar{x}, \forall i \tag{5.15}$$

Proposition 5.2.1 *For an increasing concave utility function u and an increasing convex delay function d, the revenue maximizing traffic volume x_i^**

of user i is given by $x_i^* = \min[x_i', \bar{x}]$, where

$$x_i' = \left\{ x_i | u'(x_i; v) = N d'(N x_i; C) \right\}.$$

Each user pays the same amount $p(x_i^*) = [u(x_i^*) - d(N x_i^*; C)]$ for the service.

Proof: Since all users are identical, eq. (5.13) can be expressed as $N[u(x_i; v) - d(T; C)]$. Thus the first order condition is $u'(x_i'; v) - N d'(T; C) = 0$. Also, from eq. (5.15), we have $x_i^* = \min[x_i', \bar{x}]$. □

Unsaturated Network ($\mu > 0$)

When the network is not saturated or $\mu > 0$, the user's data rate $x_i^* = \bar{x}$ by the complementary slackness condition (5.8). The total traffic in the network is $T = N \cdot \bar{x}$, and the user's net-utility is $u(\bar{x}; v) - d(N\bar{x}; C) - p(x)$.

Under the flat pricing scheme, the charge is independent of traffic volume or $p(x) = f$, for $f \geqslant 0$. The first order condition (5.7) reduces to

$$u'(x_i^*; v) = d'(T; C) + \mu; \tag{5.16}$$

$$\mu(x_i^* - \bar{x}) = 0. \tag{5.17}$$

As the users subscribe to the service when their net-utility is nonnegative, we have

$$f_\phi \leqslant u(\bar{x}) - d(N\bar{x}; C). \tag{5.18}$$

To maximize the revenue R, the inequality (5.18) should be satisfied with equality or $f_\phi = u(\bar{x}) - d(N\bar{x}; C)$. If the delay disutility $d(N\bar{x}; C)$ is negligible, the provider's revenue R is

$$R = N[u(\bar{x}; v) - d(N\bar{x}; C)] \approx N u(\bar{x}; v).$$

For the two-part tariff, the optimal revenue can also be achieved as the revenue of the two-part tariff is greater than or equal to that of flat pricing.

Note that flat pricing is a special case of the two-part tariff when $\beta = 0$. For β in the range of $[0, u'(\bar{x}) - d'(N\bar{x}; c)]$, the optimal revenue can be achieved. The upper bound $u'(\bar{x}) - d'(N\bar{x}; c)$ follows from inequality (5.7). The optimal fixed value a^* is determined so that the net-utility value is zero or

$$a^* = u(\bar{x}) - d(N\bar{x}; C) - \beta^*\bar{x}. \tag{5.19}$$

Proposition 5.2.2 *When the network is not congested and all users are identical, both the flat charge and two-part tariff achieve the optimal revenue* $N[u(\bar{x}) - d(N\bar{x}; C)]$.

Proof: For the two-part tariff, the provider's problem is $\max_{a,\beta} N[a + \bar{x}\beta]$ *s.t.* $a \leqslant u(\bar{x}; v) - d(N\bar{x}) - \beta\bar{x}$. From eq. (5.19), the revenue reduces to $N[u(\bar{x}; v) - d(N\bar{x}; C)] \approx Nu(\bar{x}; v)$, which is equivalent to the flat price revenue. □

Saturated Network ($\mu = 0$)

When the number of users in the network is large enough, the network is prone to congestion. In this case, the delay disutility plays an important role in determining the data rate of users. Under flat pricing, a user's data rate x_F^* is determined such that

$$u'(x_F^*) = d'(Nx_F^*; C), \tag{5.20}$$

which is the first order condition. Similarly, the first order condition for the user's traffic volume x_T^* under the two-part tariff is given by

$$u'(x_T^*) = \beta + d'(Nx_T^*; C). \tag{5.21}$$

Figure 5.1 illustrates the optimal conditions (5.20) and (5.21). In the figure, x^* is the traffic volume that maximizes the revenue of the NSP such that $u'(x^*) = Nd'(Nx^*; C)$ (Proposition 1). The traffic volume x_F^* of flat

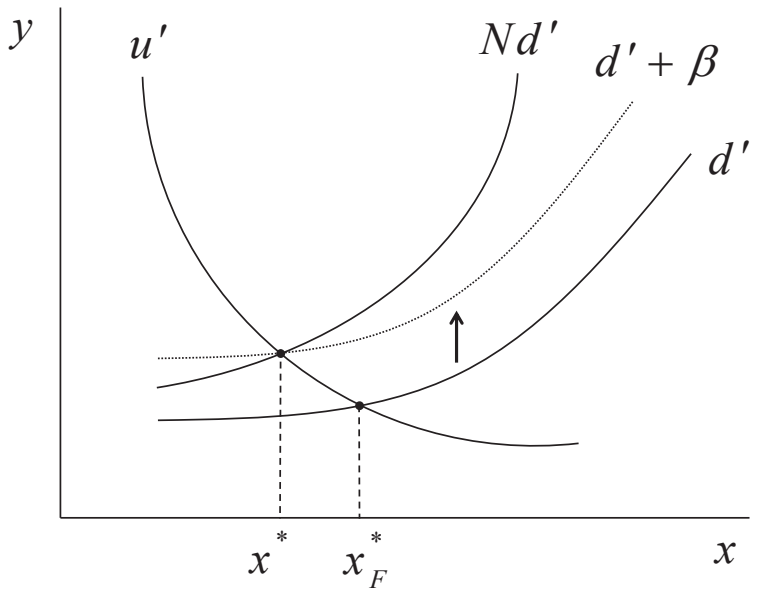


Figure 5.1: Illustration of the Optimality Conditions

pricing is higher than x^* as is shown in the figure. The traffic volume under flat pricing is greater than the revenue maximizing traffic volume.

However, the two-part tariff or volume-based charging scheme is different. By setting a proper unit price β, the total traffic volume can be controlled. If β is the difference between Nd' and d' or

$$\beta^* = (N-1)d'(Nx^*; C), \tag{5.22}$$

the two-part tariff achieves the revenue maximizing traffic volume x^*. The following proposition says that when network congestion is high, the revenue under flat pricing can be very small.

Proposition 5.2.3 *Assume that all users are identical and that the NSP charges users a flat price. For a given capacity, C, the total traffic $T^*(N)$ is nondecreasing and the optimal flat price* $\lim_{N\to\infty} f^*_\phi(N) = 0$.

Proof: Let $x^*(N)$ be a solution to equation (5.20) when the number of users is N and $T^*(N) := Nx^*(N)$. The $x^*(N)$ is unique because $u'(x)$ is strictly decreasing while $d'(T;C)$ is nondecreasing. Note that the delay disutility function $d(T(N);C)$ is convex in T, and $x(N)$ decreases in N. Similarly, it is evident that $T(N)$ increases in N and approaches C. Hence, the delay disutility $d(T(N);C)$ increases while $x(N)$ approaches zero. Therefore the net-utility of a user approaches zero when the number of users is large enough and only the small fraction of users will subscribe to the system. □

Proposition 5.2.3 shows that flat pricing is vulnerable to congestion as users would like to transmit as much data as possible once they subscribe to the service. The delay disutility is so high and users do not get any utility out of the system due to congestion.

Under the two-part tariff of $p(x) = a + \beta \cdot x$, if the NSP sets β according to (5.22), and a such that $u(x^*;v) - d(Nx^*;C) = \beta x^* + a$, then the network system can operate at a revenue maximizing equilibrium. The following proposition says that the difference in revenue from flat pricing and that of the two-part tariff can be quite large under severe congestion.

Proposition 5.2.4 *Assume that all users are identical. For a given capacity C,*

$$PoS \to 1 \quad as \quad N \to N_F$$

where N_F is the number of user at which the revenue of flat price becomes 0 i.e. $R_F(N_F) = 0$.

Proof: From proposition 5.2.3, we know that $\lim_{N\to N_F} f^*_\phi(N) = 0$. Since the revenue of flat pricing is Nf^*_ϕ, we see that $\lim_{N\to N_F} R_F = \lim_{N\to N_F} \frac{R_F}{R_T} = 0$. Now, let N_T be the number of user at which the revenue of two-part tariff becomes 0. Since two-part tariff is a generalization of flat price we have

$R_T(N) \geqslant R_F(N)\ \forall N$. From this we know that $N_T \geqslant N_F$ hence $\lim_{N \to N_F} \frac{R_F}{R_T}$ is well defined in $\mathbf{R}$. □

Examples

In this section, we provide examples for two typical delay functions, $d(T;C) = \frac{T}{C}$ and $d(T;C) = \frac{1}{C-T}$.[3] In our analysis, $\epsilon \in (0,1]$ is multiplied to the delay function to represent a user's sensitivity to congestion (i.e., $\epsilon d(T;C)$). For both cases we use the utility function $u(x;v) = vx^\alpha$, where v is some constant.[4] Note that with this utility function the price elasticity of demand is $\frac{1}{1-\alpha}$.[5]

Example 1 We first consider the delay function as $d(T;C) = \frac{T}{C}$, which is defined to be total traffic divided by capacity.[6] For this type of delay function we have the following corollary.

Corollary 5.2.5 *For identical users with a utility function $u(x;v) = x^\alpha$ and a delay function $d(T;C) = \frac{T}{C}$, the price of simplicity is monotonically increasing in N.*

Proof: To consider all cases, we first enumerate each case in Table 5.4. Next we provide the proofs of each of the four cases as follows:

(1-1): This case is when $N < C\left(\frac{\alpha}{\epsilon}\right)^{\frac{1}{\alpha}}$. If both T_F and T_T is equal to C the revenue is also equivalent, i.e., $R_F/R_T = 1$.

[3]For $d = 0$ (unsaturated), we have seen that the revenue of the two pricing schemes is equivalent. Thus, $\frac{R_F}{R_T} = 1$.

[4]We assume $v = 1$ without loss of generality.

[5]The price elasticity of demand is defined as the negative ratio of the percent change in demand to the percent change in price given by $E_d = -\frac{\Delta D/D}{\Delta p/p}$.

[6]The reason we consider this kind of delay is that (1) it was used in many other related studies [62, 6, 34] and (2) it allows us to derive explicit solutions.

Table 5.4: All Cases for Comparing R_F and R_T

(1) $T_F(N=1)=C$	(1-1) $T_F=C$ & $T_T=C$	(1-2) $T_F=C$ & $T_T<C$
(2) $T_F(N=1)<C$	(2-1) $T_F<C$ & $T_T<C$	(2-2) $T_F=C$ & $T_T<C$

(1-2): For this case, the revenue ratio is given by

$$\frac{R_F}{R_T}=\frac{1}{\left(\frac{\alpha}{\epsilon}\right)^{\frac{\alpha}{1-\alpha}}(1-\alpha)}\left\{\left(\frac{N}{C}\right)^{\frac{\alpha^2}{1-\alpha}}-\epsilon\left(\frac{N}{C}\right)^{\frac{\alpha}{1-\alpha}}\right\}.$$

Then if we differentiate this revenue ratio with respect to N it satisfies

$$\frac{d(R_F/R_T)}{dN}=\underbrace{s(\alpha)\cdot\frac{1}{C}\left(\frac{N}{C}\right)^{\frac{2\alpha-1}{1-\alpha}}}_{+}\underbrace{\left\{\alpha\left(\frac{C}{N}\right)^{\alpha}-\epsilon\right\}}_{-}\leqslant 0,$$

where $s(\alpha)=\frac{1}{\left(\frac{\alpha}{\epsilon}\right)^{\frac{\alpha}{1-\alpha}}(1-\alpha)}\cdot\frac{\alpha}{1-\alpha}$. The inequality $\alpha\left(\frac{C}{N}\right)^{\alpha}<\epsilon$ holds since this case is for $T_T<C$ and this can be written as follows:

$$N\geqslant C\left(\frac{\alpha}{\epsilon}\right)^{\frac{1}{\alpha}}\ \Leftrightarrow\ \alpha\left(\frac{C}{N}\right)^{\alpha}\leqslant\epsilon.$$

(2-1): For this case, the revenue ratio can be expressed as

$$\begin{aligned}\frac{R_F}{R_T}&=\frac{N\left(\frac{C\alpha}{\epsilon}\right)^{\frac{\alpha}{1-\alpha}}(1-\alpha N)}{N^{\frac{1-2\alpha}{1-\alpha}}\left(\frac{\alpha C}{\epsilon}\right)^{\frac{\alpha}{1-\alpha}}(1-\alpha)}\\&=N^{\frac{\alpha}{1-\alpha}}\cdot\frac{1-\alpha N}{1-\alpha}.\end{aligned}$$

Once again, if we take the derivative then

$$\frac{d(R_F/R_T)}{dN}=\underbrace{\frac{\alpha}{1-\alpha}\cdot N^{\frac{\alpha}{1-\alpha}}\cdot\frac{1}{N(1-\alpha)}}_{+}\underbrace{(1-N)}_{-}<0,$$

where the inequality holds since $N\geqslant 1$.

(2-2): This case is equivalent to (1-2). □

From figure 5.2 we see that when the network is unsaturated (small N), both revenue is identical as proposition 5.2.2. However, when the network is

saturated (large N) the revenue function is (1) concave increasing function if $\alpha < 1/2$ (Fig. 5.2(a)); (2) constant if $\alpha = 1/2$; and (3) convex decreasing if $\alpha > 1/2$ (Fig. 5.2(b)), all of which is larger than the flat price. The interpretation is that when the price elasticity is high ($\alpha > 1/2$) users are very sensitive to the price whether it is a service price or a congestion price, thus when the number of users increases the price perceived by the users also increases, which makes the total traffic decrease rapidly, and thus the revenue decreases. On the other hand, if users are less elastic ($\alpha < 1/2$), the congestion price will have relatively little effect on the user utility thus, as there are more users in the network, the decreasing rate of total traffic is lower than the rate of user increase, which leads to an increase in revenue.

Example 2 We next consider the delay function of $d(T;C) = \frac{1}{C-T}$, which basically represents a delay in $M/M/1$ queue. Let x_F denote the (optimal) user rate of the flat price and x_T that of the two-part tariff. Then the optimal user rate of each pricing scheme satisfies the respective first order conditions as follows:

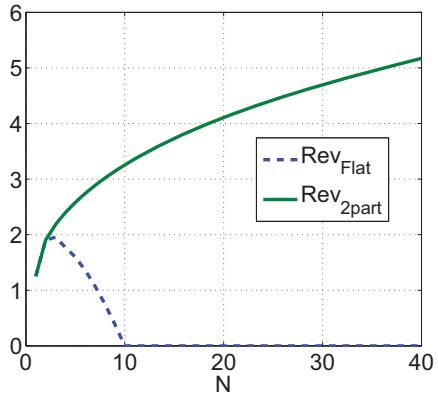
$$\alpha x_F^{\alpha-1} = \epsilon \frac{1}{(C - Nx_F)^2}, \quad \alpha x_T^{\alpha-1} = \epsilon \frac{N}{(C - Nx_T)^2}. \tag{5.23}$$

From the numerical experiment, we can observe that the PoS is increasing in N using the same parameter as in the previous example (Figure 5.3(c)).

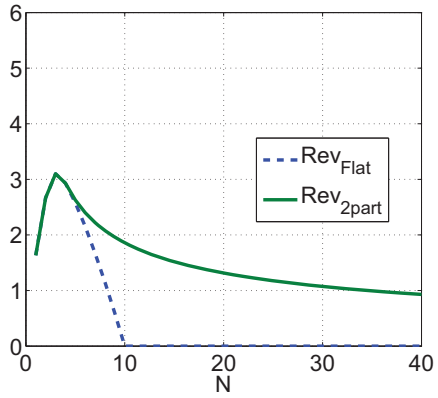
5.2.2 Heterogeneous Users

In this section, we consider a more realistic scenario in which users are heterogeneous and study how this heterogeneity affects the price of simplicity under congestion. Here we will consider the following two specific cases where consumer type is represented by a one-dimensional parameter.

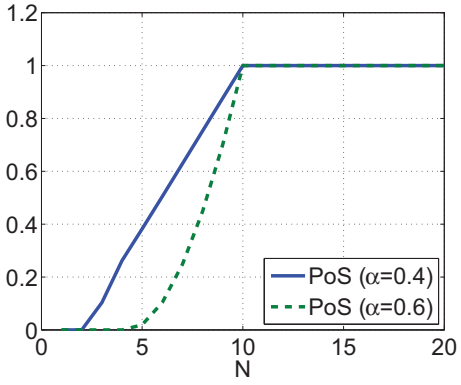
- Heterogeneity in terms of consumption preference (v)
- Heterogeneity in terms of congestion preference (ϵ)



(a) Revenue ($\alpha = 0.4$)

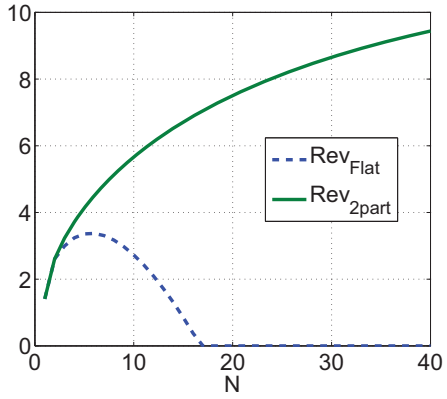


(b) Revenue ($\alpha = 0.6$)

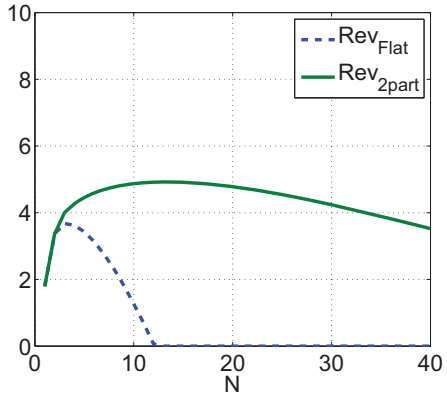


(c) Price of Simplicity

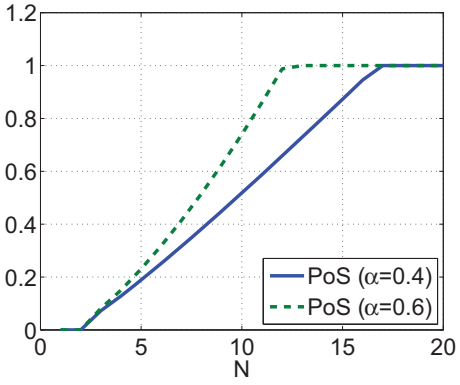
Figure 5.2: Revenue and Price of Simplicity for Identical Users with $d(T;C) = T/C$



(a) Revenue ($\alpha = 0.4$)



(b) Revenue ($\alpha = 0.6$)



(c) Price of Simplicity

Figure 5.3: Revenue and Price of Simplicity for Identical Users with $d(T;C) = 1/(C-T)$

Different Tastes for Consumption

Suppose the users differ only in their preferences for the network services. Unlike identical users, we assume that the perceived utility of a type-v user that generates a data volume x is given by

$$u(x; v) = vx^{\alpha}, \qquad 0 < \alpha < 1. \tag{5.24}$$

where we assume that v is distributed according to distribution function $G(v)$ in $[0, 1]$ with density $g(v) > 0$. In this expression, $\alpha \in (0, 1]$ is a parameter that corresponds to the *isoelastic* utility functions.[7] If the value of v is large, the user is willing to use more bandwidth since its perceived utility value is high.

The *net-utility* of a type-v user is defined as the perceived utility minus the payment $p(x)$ and delay disutility $d(T; C)$ to the service provider for the traffic volume x, where T is the total traffic and C is the capacity. The consumer of type v maximizes his or her net-utility and generates the traffic volume $x(v)$, where the first order condtion is

$$u'(x(v)) - p'(x(v)) - d'(T; C) = 0. \tag{5.25}$$

Note that a marginal user of type v_0 exists such that the net-utility is zero. Thus users in $[v_0, 1]$ join the service while $[0, v_0)$ do not. The total traffic T generated by all active users is less than or equal to the capacity C, i.e.,

$$T = \int_{v_0}^{1} x(v) N dv \leqslant C.$$

Given $x(v)$, the NSP aims to maximize its revenue R, i.e.,

$$\text{maximize } \left[R = \int_{v_0}^{1} p(x(v)) N dv \right].$$

[7] When $\alpha \to 1$, the utility function becomes linear, while it approaches a step function as $\alpha \to 0$. Note that all of the users have the same parameter α and have utilities that are proportional to one another.

For *flat price*, the provider's problem can be expressed as

$$\begin{aligned}
&\max_{v_0}\left[R=\int_{v_0}^{1} f_\phi N dv\right] \\
&=\max_{v_0} f_\phi(1-v_0)N \\
&=\max_{v_0}\ [u(x(v_0))-d(T;C)](1-v_0)N.
\end{aligned}$$

For *two-part tariff*, from the first order condition (5.25) we have $u'(x(\beta,v)) = \beta + d'(T;C)$, and since the net-utility of a marginal user of type v_0 is zero, we have

$$a = u(x(v_0)) - \beta x(v_0) - d(T;C), \tag{5.26}$$

where a is a function of β and v_0. Then the NSP problem reduces to

$$\begin{aligned}
&\max_{v_0,\beta}\ \left[R=\int_{v_0}^{1}[a+\beta x]N dv\right] \\
&=\max_{v_0}\ \ aN(1-v_0)+\beta T.
\end{aligned}$$

Different Tastes for Congestion

In addition to the heterogeneity of consumption preference v, we also incorporated the heterogeneity of *congestion* preference [39] so as to explore the price of simplicity more thoroughly. The net utility function is now given by

$$u(x;v) - p(x) - \epsilon d(T;C). \tag{5.27}$$

We now fix v as constant and let the congestion sensitivity parameter $\epsilon \in [\underline{\epsilon}, \overline{\epsilon}]$ be a random variable with distribution $H(\epsilon)$ and density $h(\epsilon)$. The analysis is similar to that of consumption preference except that, here, the marginal user type is denoted by ϵ_0 and the consumers in $[\underline{\epsilon}, \epsilon_0]$ join the service while $[\epsilon_0, \overline{\epsilon}]$ do not.

Analytical Results

The NSP problems above do not admit explicit solutions with general class of utility functions and most of the special cases. Nevertheless, here, we provide some analytical results for a particular utility function with $v \sim U[0,1]$, $\alpha = 1/2$ and $d(T;C) = T/C$ since it is the only case that admits a closed-form solution.

Lemma 5.2.6 *The revenue with flat price, R_F for heterogeneous user case with $v \sim U[0,1]$, $\alpha = 1/2$ and $d(T;C) = T/C$ is quasi-concave and unimodal in its domain* $\mathbf{dom}R_F = \{v_0 | v_0 \in [0,1]\}$

Proof: From the analysis of heterogeneous users, the revenue function with flat price is given by

$$R_F(v_0) = \frac{1}{12\epsilon} NC(1-v_0)\{6v_0^2 - N(1-v_0^3)\}$$

Differentiating with respect to v_0 we have

$$R'_F(v_0) = \frac{dR_F}{dv_0} = \frac{1}{12\epsilon} NC\{-4Nv_0^3 + 3(N-6)v_0^2 + 12v_0 + N\}$$

By applying the first order condition, we have found that there is merely one locally minimum (or maximum) point. Also,

$$R'_F(0) = \frac{1}{12\epsilon} CN^2 > 0 \text{ and } R'_F(1) = -\frac{1}{2\epsilon} CN < 0$$

Thus, the revenue function R_F first monotonically increases and then makes its peak at one of the points in $(0,1)$ then monotonically decreases. However, since $R''_F(0) = \frac{1}{\epsilon} CN > 0$ and $R''_F(1) = -\frac{1}{2\epsilon} CN(N+4) < 0$, the function is not concave/convex nor quasi-linear but quasi-concave. □

Proposition 5.2.7 *In the case of heterogeneous users with $v \sim U[0,1]$, $\alpha = 1/2$ and $d(T;C) = T/C$, when the provider uses flat price, eventually, there*

will be no single user type who subscribe to network service as the number of users increases, i.e., for a given capacity C,

$$\lim_{N\to\infty} v_{0F}^*(N) = 1$$

where v_{0F}^* denotes the marginal user type in flat pricing scheme.

Proof: From lemma 1 we know that R_F has a unique maximum value in its domain. By applying the first order condition, the optimal point v_{0F}^* that maximizes revenue is given by

$$v_{0F}^* = \frac{3N-18}{12N} + \left\{ \frac{(3N-18)^2}{144N^2} + \frac{1}{N} \right\} \cdot \frac{1}{\sqrt[3]{\xi(N)}} + \sqrt[3]{\xi(N)}$$

where

$$\xi(N) = \Delta(N) + \sqrt{\Delta(N)^2 - \left\{ \frac{(3N-18)^2}{144N^2} + \frac{1}{N} \right\}^3}$$

and

$$\Delta(N) = \frac{3N-18}{8N^2} + \frac{(3N-18)^3}{1728N^3} + \frac{1}{8}.$$

Thus, we have

$$\lim_{N\to\infty} v_{0F}^*(N) = \frac{1}{4} + \frac{1}{16} \cdot \frac{1}{\sqrt[3]{\lim_{N\to\infty} \xi(N)}} + \sqrt[3]{\lim_{N\to\infty} \xi(N)} = 1$$

□

From the two results above we drew the following proposition which states that the price of simplicity increases to its maximum as the number of users grows for heterogeneous users case.

Proposition 5.2.8 *In the case of heterogeneous users with $v \sim U[0,1]$, $\alpha = 1/2$ and $d(T;C) = T/C$, for a given capacity C,*

$$PoS \to 1 \quad as \quad N \to \infty$$

Proof: From lemma 2, we have

$$\lim_{N\to\infty} R_F^* = \lim_{N\to\infty} \frac{1}{12\epsilon} NC(1-v_{0F}^*)\{6v_{0F}^{*2} - N(1-v_{0F}^{*3})\} = 0$$

and since $R_F^* \leqslant R_T^* \ \forall N$, the following holds

$$\lim_{N\to\infty} \text{PoS} = \lim_{N\to\infty} 1 - \frac{R_F^*}{R_T^*} = 1$$

where R_F^* and R_T^* are the optimal revenue obtained by using flat price and two-part tariff respectively, given N. □

Although it is of specific case, $\alpha = 1/2$ is a typical example of concave utility functions and the delay function $d(T;C) = T/C$ is considered in many other studies for capturing user's disutility of congestion [62, 39, 6, 34]. Thus, we believe the analytical result obtained above gives a useful insight on the price of simplicity for heterogeneous user case together with our numerical results for more general utility and delay functions.

Numerical Results

In this section we analyze the price of simplicity for heterogeneous users based on a numerical experiment. Since we used $u(x;v) = vx^\alpha$, the price elasticity of demand is $\frac{1}{1-\alpha}$. Based on the previous literatures' findings we set $\alpha = 1/3$ as it is the elasticity estimate in data traffic.[8]

Price of Simplicity for Users with Heterogeneous Consumption Preference For the users with different tastes in consumption, we assume that the user type v is uniformly distributed in $[0, 1]$.

[8]Traditional elasticity estimate for voice traffic is approximately 1.05, while France Telecom obtained a value as high as 1.337 [1]. Lanning et al. provide an elasticity estimate for data traffic for various equipment that ranges between 1.3 and 1.7 [58]. This corresponds to the parameter α for data traffic ranging from 0.23 to 0.41, approximately.

Observation 5.2.9 *The price of simplicity is low when the network is not congested.*

Observation 5.2.9 says that flat pricing can be as good as the two-part tariff when the network is not congested. In Figure 5.4(a) it is apparent that for every utility function (represented as α) the upper bound of PoS is 0.22.[9]

Observation 5.2.10 *The price of simplicity for heterogeneous users increases in N when a delay disutility exists and remains constant when the distutility is negligible.*

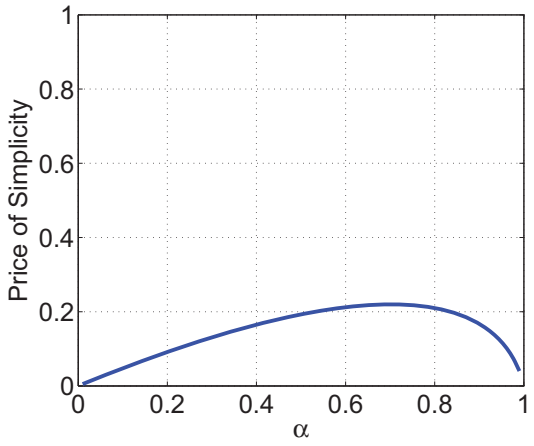
From Fig. 5.4(b), if congestion externality exists then the revenue difference between the two-part and flat price increases with N. Second, we observe that without congestion externality the price of simplicity in a data network ($\alpha = 1/3$) is 0.14.[10]

Also, the price of simplicity for heterogeneous users with $d = \frac{T}{C}$ is higher compared to when $d = \frac{1}{C-T}$. The reason is that for $d = \frac{1}{C-T}$, the delay difference of flat and two-part tariff is lower than for $d = \frac{T}{C}$ (Figs. 5.5(c) and 5.5(d)). This is due to the shape of the function where for very high traffic near the capacity (i.e., $T \approx C$), the slope of the former delay function is very steep while for intermediate traffic levels the slope of the former is steeper. The implication is that for "capacity sharing services" the revenue loss is more severe than that for "latency-based services."

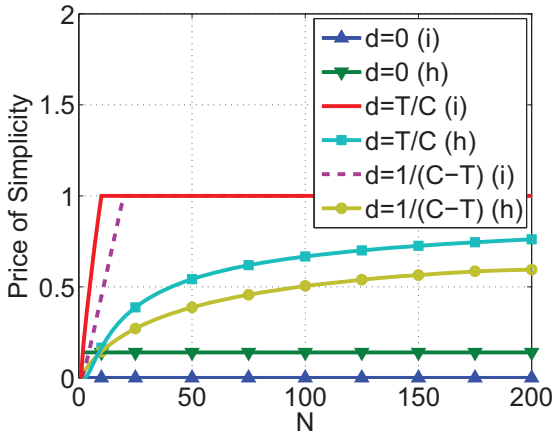
Observation 5.2.11 *The price of simplicity for heterogeneous users is lower compared to the identical users when a delay disutility exists.*

[9] This observation is consistent with the results in [93].

[10] In [93], the authors have used a utility function $\log(1+x)$, the shape of which is very similar to $x^{1/3}$ we used in this example, and produced a price of simplicity of 0.13. Also, the reason the value is constant over N can be found in the revenue function. $c = \frac{C}{N}$, which is the only variable related to N, cancels out when deriving $\frac{R_F}{R_T}$, thus the revenue ratio becomes independent in N.



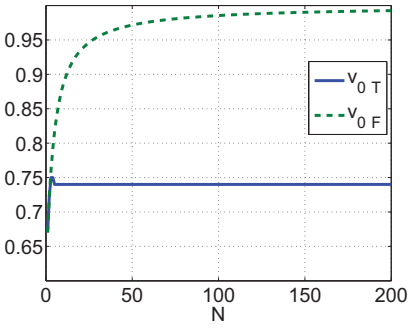
(a)



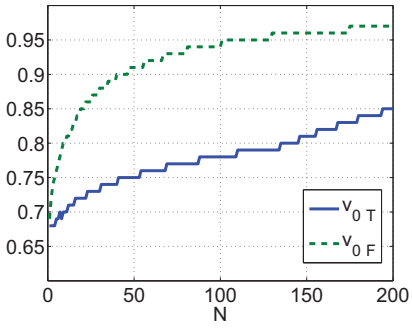
(b)

Figure 5.4: Price of Simplicity for Heterogeneous Users Case: (a) PoS w.r.t. α for $d = 0$, (b) PoS w.r.t. N where '(i)' and '(h)' refers to identical and heterogeneous users, respectively.

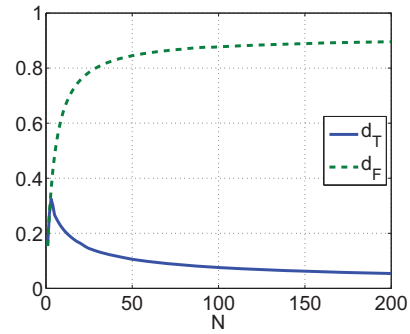
Note that the revenue of the flat price does not decrease all the way to zero as in the indentical user case (Fig. 5.4(b)). To see why, recall that the flat price is determined as $f_\phi = u(x^*; v) - d(T; C)$. Note that the delay



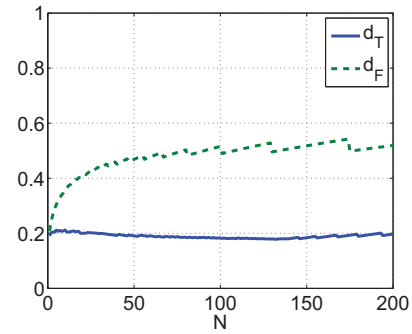
(a) $d(T;C) = T/C$



(b) $d(T;C) = 1/(C-T)$



(c) $d(T;C) = T/C$

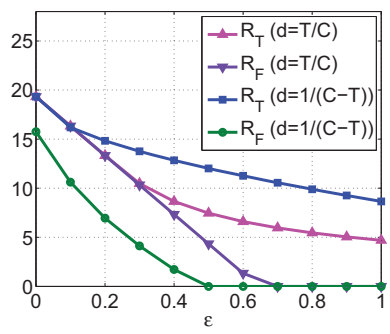


(d) $d(T;C) = 1/(C-T)$

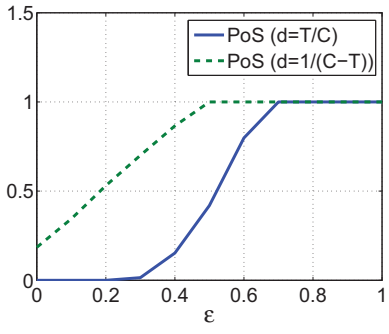
Figure 5.5: Indifferent User type (v_0) and Delay for Heterogeneous Users Case

function is increasing in T. Thus, if the total traffic increases, then delay increases and eventually the flat price decreases to 0. For the identical user case, from proposition 3, the revenue decreases to 0 as the number of users increases. However, for the heterogeneous case, it turns out that the revenue-maximizing NSP has a choice to drop users when the traffic gets too high, i.e., the provider balances the traffic by providing the service only to the users with a high willingness-to-pay.[11] Thus the heterogeneity of users allows the

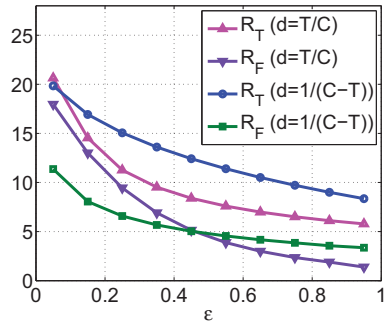
[11] Figs. 5.5(a) and 5.5(b) show that when there are many users in the network the provider sets v_0 high enough to prevent the delay from becoming so long so that the revenue does



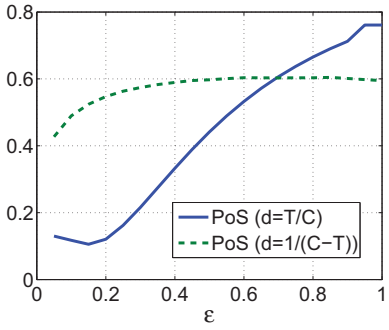
(a) Revenue (ident.)



(b) PoS (ident.)



(c) Revenue (hetero.)



(d) PoS (hetero.)

Figure 5.6: Revenue and Price of Simplicity (identical vs. heterogeneous) with respect to ϵ

opportunity for the provider adopting a flat price to maintain at least some amount of revenue even in a congested network.

Price of Simplicity for Users with Heterogeneous Congestion Preference

Observation 5.2.12 *The price of simplicity shows an increasing tendency with respect to the congestion sensitivity parameter constant ϵ. Also, even when the users are heterogeneous in terms of their preference towards con-*

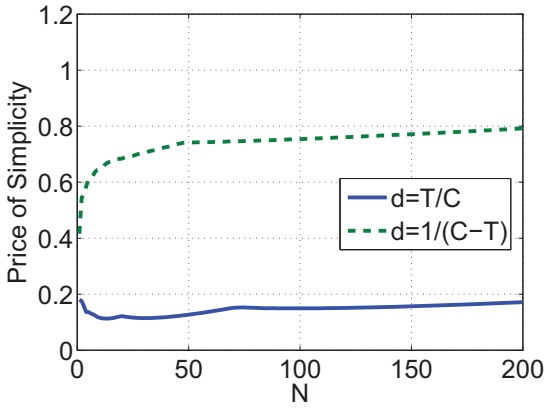
not decrease significantly to 0

gestion, the PoS shows an increasing tendency with respect to the number of users.

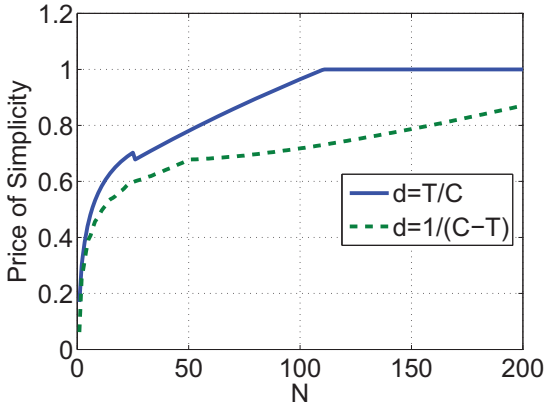
Figs. 5.6(a) and 5.6(c) show that the revenue decreases with ϵ for both the identical and heterogeneous user case. This is quite intuitive as for more congestion-sensitive users the net-utility will decrease, and so will the revenue. Figures 5.6(b) and 5.6(d) illustrate the fact that the price of simplicity has a positive slope, on average, which is similar to the previous case with respect to the number of users. Also, the PoS is lower for the heterogeneous case than that for the identical case. This is due to the fact that ϵ can be viewed as a slope of the delay function.

Figure 5.7 shows the price of simplicity with users having different tastes for congestion. We observe that, as for the users having different tastes for consumption, the PoS increases in N, and for $N = 200$, it reaches a value greater than 0.8 for both the delay functions considered.[12] The results suggest that whether consumers are differentiated by their preference towards consumption or congestion, the revenue loss experienced by the NSP generated by using flat price in lieu of two-part tariff is substantial under severe congestion.

[12]However, particularly for the delay function $d(T;C) = T/C$, when the user population consists of people with relatively low congestion sensitivity (Fig. 5.7(a)), the PoS value is small, which can be explained as follows. When most of the users are insensitive to congestion, then no matter how many users are in the system, users consume vast amount of traffic under flat pricing regime with $d(T;C) = T/C$. In this case, we have seen that the NSP will have no choice but to introduce δ to make the traffic slightly less than the capacity C. Note that the role of δ is similar to what β does in two-part tariff in that it tries to constrain the level of traffic. Thus, the flat revenue R_F for this special case is not exactly the flat revenue *per se* but more closer to a two-part tariff revenue R_T, which is why the value of PoS derived from the original definition is low.



(a)



(b)

Figure 5.7: Price of Simplicity with Users having Different Tastes for Congestion: (a) $\epsilon \sim U[0.1, 1]$; (b) $\epsilon \sim U[1, 2]$.

Social Welfare Previous results are concerned with the revenue of a network provider. However, to have a comprehensive analysis of the two pricing schemes, we should examine the welfare of users and the whole system. This

could be done by observing the *social welfare* which can be computed by

$$W = \int_{v_0}^{1} [vx(v)^{\alpha} - p(x) - d(T;C) + p(x)]Ndv \tag{5.28}$$

$$= \int_{v_0}^{1} [vx(v)^{\alpha} - d(T;C)]Ndv. \tag{5.29}$$

and we have the following two important observations.

Observation 5.2.13 *The two-part tariff is better than the flat price in terms of both social welfare and user surplus.*

With regard to welfare, Figure 5.8(a) shows that, if there is no congestion externality then the welfare and user surplus is fairly high compared to the case with externality. However, this is somewhat misleading, since obviously if there is a number of users generating a great number of traffic then users inevitably experience delay disutility. In a more realistic model, the welfare and user surplus is found to be lower for the flat price. Moreover, the user surplus is almost zero for the flat price since the delay disutility is extremely high. This is because the total traffic $T(N)$ for flat price is increasing in N which causes substantial delay disutility to all the active users for large N, which implies the total traffic cannot be constrained by using flat price. For the two-part tariff, on the other hand, since it has the ability to reach the social optimum by controlling β, the welfare is high and so is the user surplus.

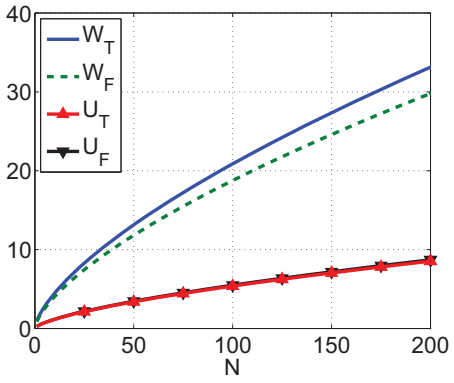
Also, The user surplus is positive for heterogeneous user case. For the identical user case, the provider extracts all of the user surplus to maximize the revenue leading to 0 user surplus. However, for heterogeneous user case, in Figure 5.8 we see that the user surplus, though small, is positive. The reason for this is that as in the identical user case, a fixed price f_{ϕ} is used to extract the user surplus. However, in the heterogeneous case, this is set

where the net utility of indifferent user v_0 is zero. Thus the users for which $v > v_0$ will have a positive surplus.

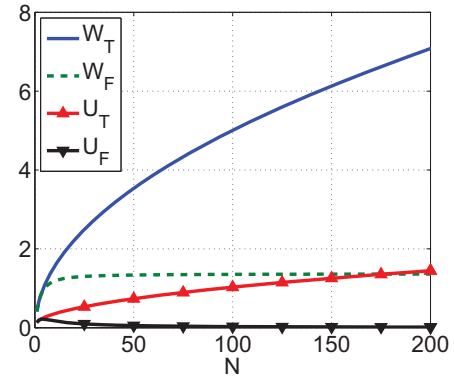
Observation 5.2.14 *The migration from flat price to two-part tariff may be undesirable under low congestion but becomes clearly desirable under severe congestion.*

To see how the transition or migration of the pricing schemes from flat price to two-part tariff affects the change in user surplus, we derive and plot the user surplus for the two pricing schemes in terms of every user type v (Fig. 5.9). Also, we plot the fraction of users that experience net user surplus change (i.e., gain or loss) as well as the total surplus increment or decrement for those users, generated by the transition (Fig. 5.10). First, we observe that when there are small number of users, high v users favor flat price over two-part tariff while low v users favor two-part over flat. The reason is that, the users of relatively high v (i.e., users with high consumption preference) consumes large amount of data and, under light congestion, the utility for consumption is higher than the disutility for delay. In this case, when the NSP uses flat price then the disutility for payment is smaller than that for two-part tariff. However, for users of relatively low v, two-part tariff is better since they consume small amount of data and hence the payment is less with two-part tariff than that with flat.

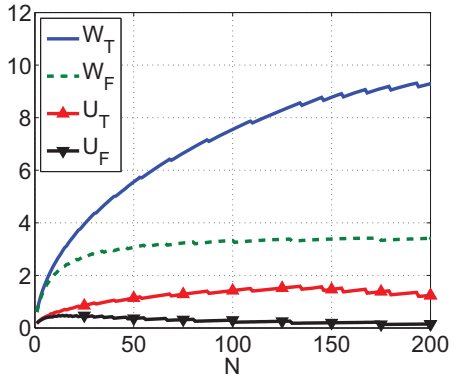
However, as the number of users grows, the delay disutility plays a critical role in the users' perceived net utility. We have seen that the delay with flat price increases while that with two-part tariff decreases or stay fairly constant (in small value) when N increases. Thus, even for the users of high consumption preference, the flat price is no longer better and eventually for high N, two-part tariff dominates flat price in terms of not only provider revenue but also user welfare (i.e., society as a whole).



(a) $d(T;C) = 0$

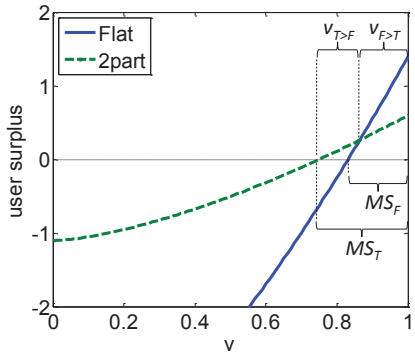


(b) $d(T;C) = T/C$

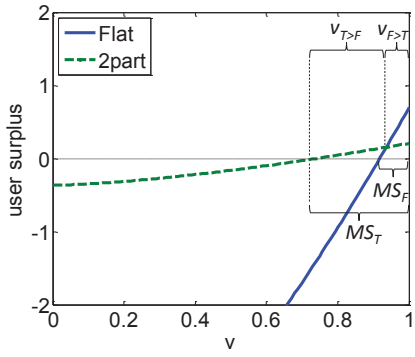


(c) $d(T;C) = 1/(C-T)$

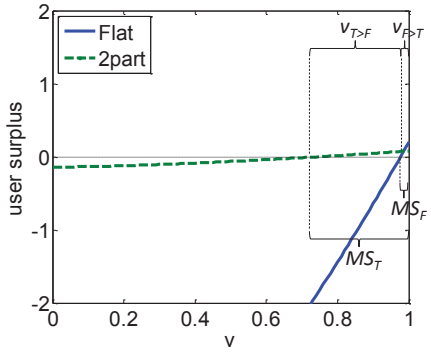
Figure 5.8: Welfare (W) and User Surplus (U) for Heterogeneous Users Case



(a) $\frac{N}{C} = 0.2$

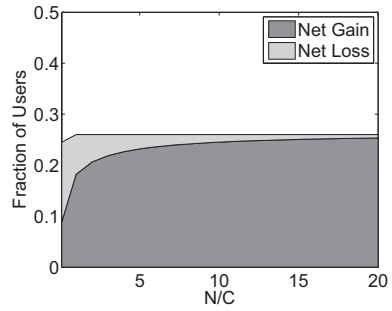


(b) $\frac{N}{C} = 1$

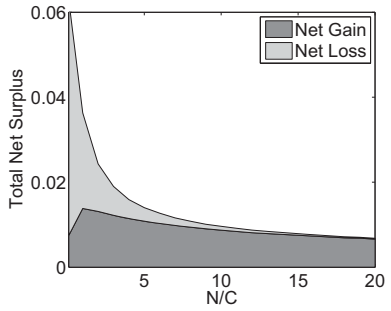


(c) $\frac{N}{C} = 10$

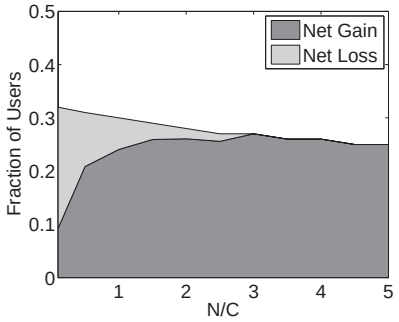
Figure 5.9: Comparison of user surplus between the two pricing schemes: Flat and Two-part Tariff ($d = T/C$). $MS_F(MS_T)$ denotes the market share when using flat price (two-part tariff), and $v_{F>T}$ ($v_{T>F}$) stands for the user types where a flat price (two-part tariff) gives higher surplus than a two-part tariff (flat price).



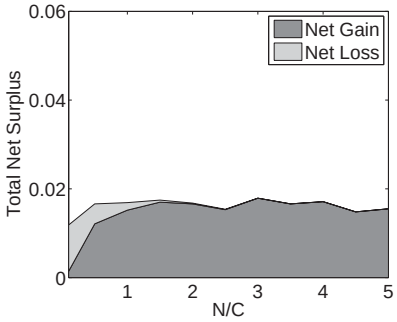
(a) Fraction of Users for $d = \frac{T}{C}$



(b) Total Net Surplus for $d = \frac{T}{C}$



(c) Fraction of Users for $d = \frac{1}{C-T}$



(d) Total Net Surplus for $d = \frac{1}{C-T}$

Figure 5.10: Change in net user surplus generated by the transition from Flat price to Two-part Tariff: (a) and (c) Fraction of users who experience net user surplus gain or loss from the transition of pricing scheme; (b) and (d) Total sum of surplus change for the users experiencing net user surplus gain or loss.

6. Discussion

In this section we discuss several practical aspects of our approach and assumptions.

6.1 Implementation of PMP in the Internet

The paris metro pricing as the name implies is certainly contrived for a metro. The question is, then, whether it can be applicable in the communication network. Technically, PMP would be easy to introduce [76]. PMP can be easily implemented under both IPv4 and IPv6 protocols. The IPv4 packets already have a 3-bit priority field that is unused. Since the number of subnetworks (i.e., classes) in PMP is not likely to exceed more than 4, this is more than sufficient. Interoperability would be easy, as all packets that do not contain any bits indicating class of service could be sent on the lowest cost (and lowest priority) channel [76]. Inside the network, changes would only have to be done in the router software. It would require to maintain logically separate queues or to give appropriate priority to packets from different channels. The current IETF standard provides all the technical tools for PMP implementation.

6.2 Assumption on θ

In chapter 3, we modelled two user characteristics with a single variable θ. This model is based on [98], and the author of [98] and we think that this assumption is reasonable. The interpretation is that a user with a large value of θ is willing to pay quite a lot for the service but he expects a low utilization for a high quality of service. Conversely, a user with a small value of θ does

not want to pay much for his connection but is likely to tolerate high delays. For example, users with large θ can be regarded as users of VoIP and those with small θ as e-mail and web browsers. Thus as long as there is no user with high willingness to pay that is okay with high delay, and a user with low willingness to pay that has strict delay tolerance, the model seems rational, since these users account for relatively low population compared to the users we consider in the model.

6.3 Implementation of Token pricing in the Internet

In chapter 4, we proposed a pricing scheme called Token pricing. The technical implementation of Token pricing scheme is not difficult since the basic framework of this scheme is identical to that of PMP. In order for the users to use this scheme, they first have to manage the tokens they receive every unit time interval. Every time a user faces an application to use, they should decide whether to use tokens and connect to the premium service or just save them and opt for the normal service. This entails keeping track of how many tokens a user currently holds and how long should the user wait for the new replenishment of tokens. This can be a tedious task for users and thus should be managed through a program (i.e., software) on behalf of the user which would be given by the provider adopting the Token pricing.

6.4 Assumption on the utility function $u(x_i; v)$

In chapter 5.2 we analyzed the price of simplicity by using the concave utility function $u(x_i; v)$ where v is the user type parameter representing willingness to pay. Like most of the consumer products, Internet service can be characterized by the concave utility function as it exhibits the "law of diminishing marginal utility." For instance, a user's utility increases fast when he uses a small number of Internet application services initially. However, if there are

many applications running already, from then on, the user's utility increases relatively slower for enjoying additional apps than the initial period. Thus, the assumption on the concavity of the utility function seems reasonable.

In addition, our utility function allows for the heterogeneity in user preferences. Even though the concavity may be reasonable, each user might have different rate of the change in their utility. This is modelled by the parameter v, which we represent as user's willingness to pay. By reflecting users' different characteristic in their valuation in the utility function we were able to analyze the price of simplicity more realistically. Also, we considered the user's difference in congestion sensitivity by ϵ, because each user may have different tolerance level when it comes to service delay.

The notion of maximum possible traffic volume $\bar{x}$ is introduced so as to model another realistic user behavior. When the price is flat and when we assume that a user can afford the service, a user should consume infinite traffic volume (i.e., $x_i = \infty$) if there is no upper bound on the usage. However, this is not true in the real world. Even if the service is affordable and the payment is irrespective of the service usage, there is a certain limitation on time, access line, or innate human capacity limit. Thus, $\bar{x}$ represents the maximum traffic volume a user can use without any price constraint.

7. Conclusions

In this dissertation, we studied pricing schemes and revenues of Internet service providers. Internet pricing is becoming more and more important as a means for efficient network management. Researchers from various fields (e.g., economics, engineering, and mathematics) proposed numerous pricing schemes so as to give efficiency in the current lightly organized Internet. However, these schemes are mostly complex in nature which made it unlikely for both ISPs and end users to accept them.

In light of this, simple yet efficient pricing schemes have been proposed from late 1990s. Paris metro pricing (PMP) divides the network into two subnetworks. The only difference between these two subnetworks is their service prices. This scheme allows users to decide on which network to select, and from our analysis it provides gains to both ISPs in revenue and users in subscription. Token pricing scheme is similar to this PMP in a sense that it provides users the incentive to use the network efficiently by dividing the network into two (logical) subnetworks (either physically or in time). A user facing this token pricing scheme pays a certain amount for a given interval (possibly, a month) as in flat pricing. In addition, an ISP gives a user a certain number of tokens. Users are to use this token only when the service is of importance and otherwise they will not use them. The mechanism of token pricing is similar to PMP such that requiring tokens makes the corresponding network less congested. From our analysis, this token pricing scheme gave a higher user value than a normal flat pricing and this surplus value increases with congestion.

In reality, however, the pricing plans commercial ISPs are using are flat-

rate pricing and usage-based pricing. There are several papers analyzing these two pricing schemes in practice and, among them, one article argued that the price of simplicity (PoS) defined by the revenue loss of an ISP from using a flat-rate instead of revenue-maximizing scheme is small. However, this work has overlooked the congestion externality in their model. Thus, we compared the revenues of two practical pricing schemes, flat-rate pricing and two-part tariff, in an attempt to reinvestigate the revenue loss created by using a simple entry fee under congestion. Our main findings are:

1. When there is no congestion, the PoS is low

2. When there is congestion, the PoS is high. Moreover, PoS gradually increases as the number of users increases.

From our numerical study, we also find that the two-part tariff is superior to the flat price in terms of social welfare and user surplus in an overloaded network. Given the huge expected increase in global data (IP) traffic, two-part tariff is likely to be desirable in the future data network.

Future work should include an examination of the effect of competition between providers on the price of simplicity. Also, it would be interesting to analyze an environment where providers can increase their capacity every unit period and determine how this dynamic setting affects the revenue loss of a simple flat entry.

A. Analysis for Heterogeneous Users with $d = 0$

In a *two-part tariff*, a user is charged a price $a + \beta x$ with usage x. That is, a user pays a fixed price a to join the network plus a price βx proportional to usage. The user problem in which a user of type v chooses x to maximize his or her net-utility can be formulated as follows:

$$(\mathbf{P_{user}}) \quad \text{maximize}_x \; u(v;x) - [a + \beta x] = vx^{\alpha} - a - \beta x.$$

The maximizing value of x is $x(v) = \left(\frac{\alpha v}{\beta}\right)^{\frac{1}{(1-\alpha)}}$, $0 \leqslant \alpha \leqslant 1$. The resulting utility is then $u(v; x(v))$ given by

$$u(v;x(v)) = k_1(\alpha)\left(\frac{v}{\beta^{\alpha}}\right)^{\frac{1}{(1-\alpha)}} - a,$$

where $k_1(\alpha) = \alpha^{\frac{\alpha}{1-\alpha}}(1-\alpha)$. This utility is nonnegative if $v \geqslant v_0$ where $a = k_1(\alpha)\left(\frac{v_0}{\beta^{\alpha}}\right)^{\frac{1}{(1-\alpha)}}$. Since v is uniformly distributed, the total traffic in the network is then

$$T = Nk_2(\alpha)\left(\frac{1}{\beta}\right)^{\frac{1}{(1-\alpha)}}\left(1 - v_0^{\frac{2-\alpha}{1-\alpha}}\right),$$

where $k_2(\alpha) = \alpha^{\frac{1}{1-\alpha}}\left(\frac{1-\alpha}{2-\alpha}\right)$. The total revenue of the provider is then R where

$$R \;=\; aN(1-v_0) + Nk_2(\alpha)\left(\frac{1}{\beta}\right)^{\frac{\alpha}{1-\alpha}}\left(1 - v_0^{\frac{2-\alpha}{1-\alpha}}\right).$$

In this expression, $aN(1-v_0)$ is the fixed part of the price where $N(1-v_0)$ users subscribe and pay a flat price a to join; the term βT is the contribution of the usage-based price.

Assuming that the network has a total capacity C, the provider aims to maximize R subject to $T \leqslant C$. That is, the provider solves the following problem:

$$(\mathbf{P_{NSP}}) \qquad \max_{a,\beta} \quad r = a(1 - v_0) + \beta t$$
$$\text{s.t.} \quad t \leqslant c,$$

where $c = C/N$, $t = T/N$, and $r = R/N$. We normalized the problem $(\mathbf{P_{NSP}})$ by a factor of N. By using a and t derived above we find

$$r = \frac{k_1(\alpha)v_0^{\frac{1}{1-\alpha}}(1 - v_0) + k_2(\alpha)(1 - v_0^{\frac{2-\alpha}{1-\alpha}})}{\beta^{\frac{\alpha}{1-\alpha}}}.$$

We can expect the full capacity to be used at the optimal price, so that $k_2(\alpha)\left(1 - v_0^{\frac{2-\alpha}{1-\alpha}}\right)\beta^{\frac{1}{\alpha-1}} = c$, i.e.,

$$\beta = \left(k_2(\alpha)\left(1 - v_0^{\frac{2-\alpha}{1-\alpha}}\right)c^{-1}\right)^{(1-\alpha)}. \tag{A.1}$$

Substituting this expression for β into r, we find

$$r = \left(\frac{c}{k_2(\alpha)}\right)^{\alpha} \frac{k_1(\alpha)v_0^{\frac{1}{1-\alpha}}(1 - v_0) + k_2(\alpha)(1 - v_0^{\frac{2-\alpha}{1-\alpha}})}{\left(1 - v_0^{\frac{2-\alpha}{1-\alpha}}\right)^{\alpha}}.$$

This expression is increasing in c, which justifies our assumption that the full network capacity is used at the optimal price. Note that r is a function of a single variable v_0 and is unimodular with a unique maximizer v_0^*.

References

[1] M. Aldebert, M. Ivaldi, and C. Roucolle. Telecommunication demand and pricing structure. In *In: Proc. International Conference on Telecommunications Systems: Modeling and Analysis*, 1999.

[2] J. Altman and K. Chu. How to charge for network services - flat-rate or volume-based? *Computer Networks*, 36(5):519–531, 2006.

[3] L. Anania and R.J. Solomon. Flat: The minimalist price. In L.W. McKnight and J.P. Bailey, editors, *Internet Economics*, pages 91–118, Cambridge, Massachusetts, 1997. MIT Press.

[4] G.S. Becker. Net neutrality and consumer welfare. *Journal of Competition Law and Economics*, 6(3):497–519, 2010.

[5] M. Bourreau, C. Gacon, and M. Columbo. The economics of internet flat rates. *Communications & Strategies*, pages 131–152, 2001.

[6] C.-K. Chau, Q.W. Wang, and D.-M. Chiu. On the viability of paris metro pricing for communication and service networks. In *Proc. of the IEEE INFOCOM*, 2010.

[7] J.P. Choi and B.C. Kim. Net neutrality and investment incentives. *Rand Journal of Economics*, 41(3):446–471, 2010.

[8] R.H. Chowdhury. Internet pricng. In *TKK T-110.5190 Seminar on Internetworking*, 2006.

[9] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2009-2014. 2010-2.

[10] D.D. Clark. Combining sender and receiver payments in the internet. In *Telecommunication Research Policy Conf.*, 1996.

[11] D.D. Clark. Internet cost allocation and pricing. In L.W. McKnight and J.P. Bailey, editors, *Internet Economics*, pages 216–52, Cambridge, Massachusetts, 1997. MIT Press.

[12] R. Cocchi, D. Estrin, S. Shenker, and L. Zhang. A study of priority pricing in multiple service class networks. *ACM SIGCOMM Computer Communication Review*, 21(4):123–130, 1991.

[13] C. Coucoubetis, F. Kelly, and R. Weber. Measurement-based usage charges in communication networks. In *Statistical Laboratory Research Report 1997-19, University of Cambridge, to appear in Operations Research*, 1997.

[14] C. Coucoubetis and V.A. Siris. An evaluation of pricing schemes that are based on effective usage. In *Proc. IEEE Int'l Conf. Communications (ICC'98)*, Atlanta, Georgia, USA, 1998.

[15] C. Coucoubetis, V.A. Siris, and G.D. Stamoulis. Integration of pricing and flow control for available bit rate services in atm networks. In *Proc. IEEE Globecom '96*, London, UK, 1996.

[16] C. Courcoubetis, G.D. Stamoulis, C. Manolakis, and F. Kelly. An intelligent agent for optimizing qos-for-money in priced abr connections. In *Tech. rep. Institute of Computer Science Foundation for Research and Technology Hellas and Statistical Laboratory, University of Cambridge. Preprint*, 1998.

[17] L. Delgrossi and D. Ferrari. Charging schemes for reservation-based networks. *Telecommunication Systems*, 11(1):127–137, 1999.

[18] A. Dhamdhere and C. Dovrolis. Can isps be profitable without violating network neutrality? In *Proceedings of the 3rd international workshop on Economics of networked systems*, pages 13–18. ACM, 2008.

[19] D.Lee, J. Mo, J. Walrand, and J. Park. A token pricing scheme for internet services. In *Proc. of Seventh ICQT*. LNCS, 2011.

[20] R. Edell, N. McKeown, and P. Varaiya. Billing users and pricing for tcp. *IEEE JSAC*, 13(7):1–14, 1995.

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

[22] H. Ekstrom. Qos constrol in the 3gpp evolved packet system. *IEEE Communications Magazine*, 47:76–83, 2009.

[23] Ericsson. Differentiated mobile broadband. White Paper. `www.ericsson.com/res/docs/whitepapers/differentiated_mobile_broadband.p%df`, 2011.

[24] A. Gupta et al. Priority pricing of integrated services networks. In L.W. McKnight and J.P. Bailey, editors, *Internet Economics*, pages 323–52, Cambridge, Massachusetts, 1997. MIT Press.

[25] D. Lee et al. Analysis of paris metro pricing for wireless internet services. In *Proceedings of the International Conference on Information Networking*, 2011.

[26] R. Jain et al. Analysis of paris metro pricing strategy for qos with a single service provider. *LNCS*, 2092:44–58, 2001.

[27] S. Shenker et al. Pricing in computer networks: Reshaping the research agenda. *ACM Computer Communication Review*, pages 19–43, 1996.

[28] V.A. Siris et al. Usage-based charging using effective bandwidths: Studies and reality. In *16th Int'l. Teletraffic Congress (ITC-16)*, Edinburgh, UK, 1999.

[29] M. Falkner, M. Devetsikiotis, and I. Lambadaris. An overview of pricing concepts for broadband ip networks. *IEEE Communications Surveys*, pages 2–13, 2000.

[30] Wireless Federation. Usage-based pricing for LTE network eyed by Verizon. `www.WirelessFederation.com/news/`.

[31] P.C. Fishburn and A.M. Odlyzko. Dynamic Behavior of Differential Pricing and Quality of Service Options for the Internet. `www.research.att.com/~amo`, 1998.

[32] R. Frieden. A primer on network neutrality. *Intereconomics*, 43(1):4–15, 2008.

[33] A. Ganesh, K. laevens, and R. Steinberg. Congestion pricing and user adaptation. In *Proceedings of IEEE INFOCOM*, pages 959–965. IEEE, 2001.

[34] R. Gibbens, R. Mason, and R. Steinberg. Internet service classes under competition. *IEEE Journal on Selected Areas in Communications*, 18(2):2490–2498, 2000.

[35] R.J. Gibbens and F.P. Kelly. Resource Pricing and the Evolution of Congestion Control. `www.statslab.cam.ac.uk/~frank/evol.html`, 1998.

[36] P. Goldstein. Is Usage-based Pricing Inevitable? `http://www.fiercewireless.com/`.

[37] H. Guo, H.K. Cheng, and S. Bandyopadhyay. Net neutrality, broadband market coverage, and innovation at hte edge. *Decision Sciences*, 43(1):147–172, 2012.

[38] A. Gupta, D.O. Stahl, and A.B. Whinston. Pricing of services on the internet. In *IMPACT: How ICCC Research Affects Public Policy and Business Markets, A Volume in Honor of G. Kozmetsky, F. Phillips and W.W. Cooper*, Cambridge, Massachusetts, 1995. MIT Press.

[39] J. Han. Monopoly pricing of congestible resources with incomplete information. *Journal of Economic Research*, 12:243–270, 2007.

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

[41] Y. Hayel and B. Tuffin. A mathematical analysis of the cumulus pricing scheme. *Computer Networks*, 47:907–921, 2005.

[42] J. Hayer. Transportation auction: a new service concept. *M.S. thesis, University of Alberta. TR-93-05*, 1993.

[43] C. Hogendorn. Broadband internet: net neutrality versus open access. *International Economics and Economic Policy*, 4(2):185–208, 2007.

[44] M.L. Honig and K. Steiglitz. Usage-based pricing of packet data generated by a heterogeneous user population. In *Proc. IEEE Infocom*, Boston, MA, 1995.

[45] H. Jiang and S. Jordan. Connection establishment in high-speed networks. *IEEE JSAC*, 9(7):1150–61, 1995.

[46] H. Jiang and S. Jordan. The role of price in the connection establishment process. *European Trans. Telecommunications and Related Technologies*, 6(4), 1995.

[47] L. Jiang, S. Parekh, and J. Walrand. Time-dependent network pricing and bandwidth trading. In *IEEE Network Operations and Management Symposium Workshops*, pages 193–200. IEEE, 2008.

[48] C. Joe-Wong, S. Ha, and M. Chiang. Time-dependent broadband pricing. *Information Theory and Applications Workshop*, 2011.

[49] C. Joe-Wong, S. Ha, and M. Chiang. Time-dependent broadband pricing: Feasibility and benefits. In *31st International Conference on Distributed Computing Systems*, pages 288–298. IEEE, 2011.

[50] S. Kalyanaraman and T. Ravichandran. Dynamic capacity contracting: A framework for pricing the differentiated services internet. In *1st Int'l. Conf. Information and Computation Economics, Submitted*, 1998.

[51] F. Kelly. On tariffs, policing, and admission control for multiservice networks. *Operations Research Letters*, 15, 1994.

[52] F. Kelly. Tariffs and effective bandwidths in multiservice networks. In *Int'l Teletraffic Conf. ITC '14*, pages 401–10, 1994.

[53] F. Kelly. Charging and rate control for elastic traffic. *European Transactions on Telecommunications*, 8(1):33–37, 1997.

[54] F. Kelly, A.K. Maulloo, and D.K.H. Tan. Rate control for communication networks: Shadow prices, proportional fairness, stability. *Journal of the Operational Research Society*, 49:237–52, 1998.

[55] G. Kesidis, A. Das, and G. de Veciana. On the flat-rate and usage-based pricing for tiered commodity internet services. In *CISS 2008*, 2008.

[56] J. Kramer and L. Wiewiorra. Network neutrality and congestion sensitive content providers: Implications for content variety, broadband investment, and regulation. *Information Systems Research*, pages 1–19, 2012.

[57] K.R. Lam, D.M. Chiu, and J.C.S. Lui. On the access pricing and network scaling issues of wireless mesh networks. *IEEE Transactions on Computing*, 56:140–146, 2007.

[58] S.G. Lanning, D. Mitra, Q. Wang, and M. Wright. Optimal planning for optical transport networks. *Philosophical Transactions of The Royal Society*, 358(1773):2183–2196, 2000.

[59] R. Lee. Subsidizing creativity through network design: Zero-pricing and net neutrality. *Journal of Economic Perspectives*, 23(3):61–76, 2009.

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

[61] J. MacKie-Mason, L. Murphy, and J. Murphy. Responsive pricing in the internet. In L.W. McKnight and J.P. Bailey, editors, *Internet Economics*, pages 279–303, Cambridge, Massachusetts, 1997. MIT Press.

[62] J. Mackie-Mason and H. Varian. Pricing congestible network resources. *IEEE Journal on Selected Areas in Communications*, 13(7), 1995.

[63] J.K. Mackie-Mason and H.R. Varian. Pricing the internet. In *Int'l Conf. Telecommunication Systems Modelling, Nashville, TN, USA*, pages 378–93, 1994.

[64] M. Mandjes. Pricing strategies under heterogeneous service requirements. *Computer Networks*, 42:231–249, 2003.

[65] P. Marbach. Analysis of a static pricing scheme for priority services. *IEEE/ACM Transactions on Networking*, 12:312–325, 2004.

[66] L.W. McKnight and J. Boroumand. Pricing internet services: after flat rate. *Telecommunications Policy*, 24(6-7):565–590, 2000.

[67] L.W. McKnight and B. Stiller. Charging and accounting for internet services. In *ACM Sigcomm*, 1998.

[68] J. Murphy and L. Murphy. Bandwidth allocation by pricing in atm networks. *IFIP Transactions C:Communications Systems*, C-24:333–351, 1994.

[69] J. Murphy, L. Murphy, and E.C. Posner. Distributed pricing for embedded atm networks. In *Proceedings of the International Teletraffic Conference*, pages 1053–1063. Elsevier, 1994.

[70] L. Murphy and J. Murphy. Feedback and pricing in atm networks. In *Proc. IFIP TC6 3rd Wksp. Perf. Modelling and Evaluation of ATM Networks*, Ilkley, England, 1995.

[71] L. Murphy and J. Murphy. Pricing for atm network efficiency. In *Proc. 3rd Int'l. Conf. Telecommunication Systems Modelling and Analysis*, Nashville, USA, 1995.

[72] L. Murphy, J. Murphy, and J. MacKie-Mason. Feedback and efficiency in atm networks. In *Proc. IEEE ICC*, Ilkley, England, 1995.

[73] J. Musacchio and J. Walrand. Wifi access point pricing as a dynamic game. *IEEE/ACM Transactions on Networking*, 14:289–301, 2006.

[74] J. Musacchio, J. Walrand, and G. Schwartz. Network neutrality and provider investment incentives. In *Asilomar Conference on Signals, Systems and Computers*, pages 1437–1444. IEEE, 2007.

[75] E. Null. The difficulty with regulating network neutrality. *Cardozo Arts and Entertainment Law Journal*, 29:459–493, 2011.

[76] A. Odlyzko. Paris metro pricing for the internet. In *Proc. ACM Conference on Electronic Commerce*, pages 140–147, 1999.

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

[78] A.M. Odlyzko. The Economics of the Internet: Utility, Utilization, Pricing, and Quality of Service. `www.research.att.com/~amo`, 1998.

[79] C. Parris and D. Ferrari. A resource-based pricing policy for real-time channels in a packet-switching network. In *Int'l Comp. Science Institute, Berkeley. Technical Report tr-92-018*, 1992.

[80] C. Parris, S. Keshav, and D. Ferrari. A framework for the study of pricing in integrated networks. In *tech. rep., TR-92-016*. Tenet Group, ICSI, UC Berkeley, 1992.

[81] I.C. Paschalidis and J.N. Tsitsikilis. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking*, 8:171–184, 1998.

[82] P. Reichl, D. Hausheer, and B. Stiller. The cumulus pricing model as an adaptive framework for feasible, efficient, and user-friendly tariffing of internet services. *Computer Networks*, 43:3–24, 2003.

[83] P. Reichl and B. Stiller. Edge pricing in space and time: theoretical and practical aspects of the cumulus pricing scheme. In *Proceedings of 17th International Teletraffic Congress*, Salvador da Bahia, Brasil, 2001.

[84] J.W. Roberts. Internet traffic, qos and pricing. In *Proc. of the IEEE*, pages 1389–1399, 2004.

[85] D. Ros and B. Tuffin. A mathematical model of the paris metro pricing scheme for charging packet networks. *Computer Networks*, 46:73–85, 2004.

[86] S. Ross. In *An Introduction to Stochastic Dynamic Programming*. Academic Press, 1983.

[87] C. Sandvig. Network neutrality is the new common carriage. *Info*, 9(2):136–147, 2007.

[88] B.V. Schewick. Towards and economic framework for network neutrality regulation. *Journal of Telecommunications and High Technology Law*, 5:329–391, 2007.

[89] F. Schuett. Network neutrality: a survey of the economic literature. *Review of Network Economics*, 9(2), 2010.

[90] G. Schwartz, N. Shetty, and J. Walrand. Impact of qos on internet user welfare. *LNCS*, 5385:716–723, 2008.

[91] N. Semret, R.R.-F. Liao, A.T. Campbell, and A.A. Lazar. Pricing, provisioning and peering: dynamic markets for differentiated internet services and implications for network interconnections. *IEEE Journal on Selected Areas in Communications*, 18:249–2513, 2000.

[92] S. Sen, C. Joe-Wong, S. Ha, and M. Chiang. Pricing data: Past proposals, current plans, and future trends. arXiv:1201.4197 [cs.NI].

[93] S. Shakkottai, R. Srikant, A. Ozdaglar, and D. Acemoglu. Price of simplicity. *IEEE Journal on Selected Areas in Communications*, 26(7):208–216, 2008.

[94] H. Shen and T. Basar. Optimal nonlinear pricing for a monopolistic network service provider with complete and incomplete information. *IEEE JSAC*, 25:1216–1223, 2007.

[95] J.G. Sidak. A consumer-welfare approach to network neutrality regulation of the Internet. *Journal of Competition Law and Economics*, 2(3):349–474, 2006.

[96] Sprint. Clearwire Reach Deal on WiMax, LTE Services. `www.lteportal.com/`.

[97] S. Wallsten and S. Hausladen. Net neutrality, unbundling and their effects on international investment in next-generation networks. *Review of Network Economics*, 8(1):90–112, 2009.

[98] J. Walrand. Economic models of commnication networks. In *Performance Modeling and Engineering (Z. Liu and C.H. Xia, eds.)*, pages 57–90, New York, 2008. Springer Publising Company.

[99] T. Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 2:141–179, 2003.

[100] H. Yaiche, R.R. Mazumdar, and C. Rosenberg. A game theoretic framework for bandwidth allocation and pricing in broadband networks. *IEEE/ACM Transactions on Networking*, 8:667–678, 2000.

초 록

서울대학교 대학원
산업공학과
이 동 명

본 연구는 통신산업에서의 가격정책과 이에 따른 수익에 관한것이다. 최근 초고속 인터넷 가입자는 기하급수적으로 증가하고 있다. 1998년 310만명이던 국내 인터넷 사용자는 2003년 약 3000만 명으로 증가하였고 이 중 초고속 인터넷 가입자는 1040만명으로 늘어났으며, 이후 증가추세는 계속 되고 있다.

그러나 인터넷 수요의 증가에 반해 인터넷 요금제는 정액제를 유지하고 있으며, 이는 인터넷 사업자의 수익을 한계점에 이르게 하고 있다. 또한, 인터넷 이용량이 많은 일부 사용자가 전체 트래픽의 대부분을 차지하는 이른바 트래픽 불균형(Disparity in Internet usage) 문제가 있는 상황에서 현재의 정액제는 사용량이 상대적으로 적은 이용자가 대량 트래픽 이용자의 비용을 대신 지불할 수 밖에 없는 상황을 만들었다.

위의 문제점들로 인하여 최근 사업자들 사이에서는 인터넷 서비스에 종량제를 도입하자는 움직임이 대두되고 있다. 그러나, 사용자들은 이에 완강히 반대하고 있으며 이로 인하여 인터넷 요금제는 대부분의 사업자가 정액제를 계속 유지하고 일부에서만 종량제를 실시하고 있는 실정이다.

이러한 현 상황에서 요금제와 수익에 관한 흥미로운 연구결과가 발표되었다. 2008년 통신분야의 권위있는 저널 *IEEE JSAC*에 발표한 한 논문에서 사업자가 정액제를 사용할 경우 수익을 최대화 시키는 요금제를 사용했을 때에 비해 수익손실이 상대적으로 크지 않다는 결론을 내렸다. 저자들은 이

러한 손실을 단순성의 대가(Price of simplicity)로 정의하고 수학적인 분석을 통해 여러 환경에서 이러한 대가가 낮음을 보였다.

하지만, 위의 논문은 실제 네트워크 환경에서 중요한 요소인 '혼잡'을 간과하고 있다. 실제로 인터넷은 다수의 사용자가 서로 연결되어있는 거대한 네트워크로서 인터넷을 사용하면서 얻는 순효용(Net-utility)을 구성하는 것은 사용하였을 때 얻는 효용과 지불가격 이외에도 혼잡비용이 추가적으로 존재하고 있는 것이 특징이다. 인터넷에서는 사용자가 많아질수록 혼잡도가 심해지며 이는 개인의 순효용에 대해 서비스 가격 이외에 추가적인 비용으로 영향을 미친다.

따라서, 본 연구에서는 이러한 혼잡 외부성(Congestion externality)이 존재하는 상황에서 위의 단순성의 대가(Price of simplicity)를 다시 분석한다. 특히, 실제로 사업자들이 채택하고있는 정액요금제(Flat price)와 이부요금제(Two-part tariff)를 사용했을 때 얻는 수익을 비교하고, 혼잡(Congestion)이 이부요금제 대신 정액요금제를 사용하였을 때 발생하는 수익손실에 미치는 영향을 분석한다. 기존연구에 의하면 지연 비효용(Delay disutility)이 없을 경우 이 수익손실이 적고 이는 낮은 단순성의 대가(Price of simplicity)로 귀결된다. 하지만, 본 연구에서는 먼저 사용자(User)들이 동일할 경우 단순성의 대가가 매우 크다는 것을 보인다. 또한, 사용자들의 서비스에 대한 선호도가 각기 다른, 좀 더 실제적인 환경에서도 혼잡환경 하에서는 위의 단순성의 대가는 매우 클 수 있다는 것을 보인다.

주요어 : 가격정책, 인터넷, 정액요금제, 이부요금제, 단순성의 대가, 혼잡 외부효과.

학번 : 2007-20639

감사의 글

관악에서 보낸 6년의 긴 시간을 이제 와서 돌이켜 보니 제 인생의 많은 부분에서 배움과 성장을 했던 시간이 아니었나 반추해 보게 됩니다. 지난 세월을 마무리하고 다시금 새로운 시작을 준비하는 이 시점에서, 본 논문을 쓰는 과정에 있어서나 다른 삶의 문제에 있어서나 혼자 힘만으로 이룰 수 있는 것은 극히 적었을 것이며 앞으로도 그러할 것임을 고백합니다. 매일 철저히 제 인생을 이끌어 주시는 하나님의 은혜에 감사드리며, 그동안 힘이 되어주신 분들께 감사의 말씀을 드리고자 합니다.

우선, 대학원 과정을 처음 시작할 때부터 마칠 때까지 언제나 깊은 사랑으로 배움의 길을 열어 주셨으며, 산업공학의 의의와 학문에 대한 진지한 열정과 인생에서의 지혜를 가르쳐주신 박진우 교수님께 깊은 존경과 감사를 드립니다. 연구의 전반적인 과정을 비롯하여 네트웍 분야에 대해 가르쳐주시고 일깨워주셨으며 연구의 엄격함과 인생의 진지함도 아울러 가르쳐주신 모정훈 교수님께 진심으로 감사를 드립니다. 또한, 부족한 점을 일깨워 주시고 따뜻한 관심으로 논문을 지도해주신 오형식 교수님, 박종헌 교수님과, 바쁘신 와중에도 기꺼이 논문 심사를 맡아 상세한 지적과 조언을 해주신 연세대학교 박재옥 교수님께 깊은 감사를 드립니다. 산업공학에 입문한 이후 계속해서 보살펴 주시고 가르침을 주셨던 서울대학교 산업공학과 모든 교수님들께 이 자리를 빌어 고개 숙여 감사를 드립니다.

세계 최고의 대학에서 수준 높은 연구환경을 경험하게 해주시고 같이 연구할 기회를 주셨을 뿐만 아니라 아름다운 버클리 캠퍼스를 소개시켜주신 UC Berkeley Jean Walrand 교수님께도 감사의 마음을 전합니다.

자동화연구실을 매개로 하여 많은 시간들을 함께해주신 선후배님들에게 감사를 드립니다. 먼저, 자동화연구실을 처음 소개시켜주고 대학원을 망설이던 내게 지원을 독려해준 강현이에게 고마움을 전합니다. 연구실의 대선배로서 논문 연구와 대학원 생활에 대한 충고 및 격려를 아끼지 않으셨던 정한일 선배님, 장태우 선배님께 깊은 감사를 드립니다. 연구실에 처음 들어왔을 때 산업공학 및 생산관리에 관한 책들을 권해주신 형곤이형, 연구실 생활 동안 여러 가르침을 주신 해중이형, 종경이형, 재현이형, 홍범이형, 목민이형, 지연누나, 현일이형, 명란누나, 진선이형, 장원이형과 부담없이 생활할 수 있도록 배려해 주신 건남이형, 정섭누나, 성호형, 영호형, 영균이형, 기헌이, 준영이형에게 감사드립니다. 많은 도움을 받았으나 선배로서 제대로 챙겨주지 못한 진우, 치호, 영우, 정호, 소연, Louis, 경휘, 영지누나, 성훈, 정훈이형, 준호형, 수민, 성진, 재봉이형, 규선, 한국, 동현, 성범에게 감사의 말과 더불어 미안함을 전합니다. 또한, 연세대학교 NEMOLAB을 통해 인연을 맺게된 광우형, 태현이, 광이, 민철이형, 호성이형, 재훈이, 한솔이, 소의, 지원이, 유진이에게도 감사의 마음을 전합니다.

오랫동안 따뜻한 사랑으로 보살펴 주신 아버지와 어머니의 은혜에 한없는 감사를 드리며, 자주 대화하고 격려해주고 큰 힘이 되어준 누나에게도 깊은 감사의 마음을 전합니다. 성실함과 겸손함을 갖춘 사회인으로 열심히 생활할 것을 약속드립니다.

2013년 2월

이 동 명