Pricing and Consumer Behavior in the Wireless Telecommunications Industry

by

Markus von Wartburg

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate Studies

(Economics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August, 2009

Abstract

The wireless telecommunications industry has seen extraordinary growth over the last decade and associated with the widespread adoption of wireless phone service are peculiar pricing schemes such as three-part tariffs and on-net/off-net pricing. This dissertation examines the interaction of consumer behavior and pricing schemes in the wireless telecommunications industry.

Chapter 2 addresses in a theoretical model the interaction of consumers' consumption patterns over the billing cycle with the monopolist's provision of access. The service provider designs a menu of contracts to screen privately informed consumers who learn about their actual demand in a sequential manner over the billing period. The model shows that the distorted contracts in the profit-maximizing menu of tariff options are characterized by an increasing marginal price schedule. Three-part pricing schemes commonly observed in the wireless telecommunications industry consisting of a fixed monthly fee, an allowance of minutes and a positive marginal price for minutes consumed in excess of the allowance can be reconciled with rational consumer behavior if the consumer model accounts for the sequential consumption pattern over the billing cycle.

Chapter 3 examines termination-based price discrimination, where the price a mobile customers pays for a call to a subscriber on another network

(off-net) exceeds the price for a call to a subscriber on the same network (on-net). A standard Hotelling-type model of network competition is combined with closed user groups such as a family or a group of friends who are able to internalize tariff-mediated network externalities when choosing their network. The model results show that termination differentials can reduce social welfare and contradict the commonly held belief that the presence of closed user groups can mitigate networks' market power.

The empirical analysis in Chapter 4 presents a structural consumer model of tariff choice and consumption in the presence of three-part tariffs. Econometric results based on individual consumer records suggest that consumers tend to exhibit significant tariff-specific preferences and that the pricing parameters of three-part tariffs have much larger effect on cellular plan choice than on the consumption of cellular calling minutes.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to take this opportunity to thank my research advisors Prof. Thomas Ross and Prof. Patrick François for their encouragement, support and guidance throughout this venture. During my time at the University of British Columbia, I have greatly benefitted from conversations with Heng Ju, Steve Yong and many other scholars in the Department of Economics and the Sauder School of Business. In addition, I would also like thank the Department of Economics and the University of British Columbia for the indispensable financial support and the permission to carry out this research.

to Ingrid and my parents for all their support

Chapter 1

Introduction

There has been extraordinary growth in the wireless telecommunications industry in the 1990s and the early part of the new century. The average growth rate of wireless telecommunication revenue in OECD countries from 1995-2005 exceeded 18.8% and in many OECD countries mobile telecommunication revenue now exceeds fixed (land-line) telecommunication revenue (OECD, 2007). Cellular phones have become such an integral part of everyday life that market penetration rates in 2005 - defined as the number of active subscribers per inhabitants - exceeded 100% in 14 OECD countries (OECD, 2007).[1] Similar growth can be observed in developing countries, where for many people cellular phones are the only telecommunication means available. Associated with this significant industry growth and the widespread adoption of wireless telecommunication services has been the increased interest among economists to gain a better understanding of consumer behavior, firm pricing and competition in the wireless telecommunications industry. This thesis dissertation examines these three aspects both from a theoretical as well as from an empirical perspective.

[1] The mobile penetration rate in Canada was 52% in 2005 and has since increased to 67%, but the mobile penetration rate in Canada is still substantially below other OECD countries (OECD, 2007; CWTA, 2008). Mobile penetration rates over 100% can occur when subscribers have multiple subscription or prepaid accounts.

Some major features distinguish cellular telecommunication services from more traditional products, and these features give rise to distinctive pricing schemes, two of which are examined in more detail in this dissertation:

Monthly Contracts and Three-Part Tariffs: Most cellular communication services are subscription-based. Consumers choose at the beginning of the billing period a service option and tariff (cellular plan) that subsequently determines their bill payment at the end of the billing period. They often choose from a menu of tariff options offered by cellular service providers that comprises of cellular plans that feature a three-part pricing structure: For a fixed monthly fee, the consumer gets access to the communications network of the service provider and receives a plan allowance of calling minutes included with the fee. If the consumers decides to consume more than the chosen monthly plan allowance, a per-minute price (overage price) applies to any calling minutes consumed in excess of the plan allowance.

Network Interconnection and On-net/Off-net Differentials: If a subscriber of a mobile network places a call to a subscriber of a competitor's network, the two network operators need to agree on network interconnection. In the process, the originating network operator pays an interconnection or call termination fee to the receiving network operator to have the call placed through to the receiving caller. This interconnection fee often gives rise to on-net/off-net differentials whereby the price a mobile customers pays for a call to a subscriber on another mobile network (off-net call) exceeds the price for a call to a subscriber on the same network (on-net call).

Three-part tariffs are at the centre of Chapters 2 and 4, on-net/off-net

differentials are the focus of Chapter 3 of this dissertation.

Consumers who use cellular phone services generally sign up for a monthly plan with a cellular phone service provider. The pricing of wireless services is typically based on a monthly billing cycle. The price for consuming cellular services often depends on the amount of services already consumed within that billing cycle, for instance if the monthly plan includes an allowance of minutes for a fixed fee and a positive marginal price for consumption of minutes in excess of the allowance. If the consumer is still within the allowance of minutes, the marginal price of an additional minute is zero. If the allowance is exceeded, a positive marginal price applies for any additional minutes consumed. At the beginning of the billing cycle, the consumer is uncertain about the exact consumption needs throughout the month and must make decisions about when and how much to consume during the billing cycle.

On the other hand, cellular phone service providers offer a menu of contracts with different pricing features in order to differentiate among various consumer segments. Chapter 2 of the thesis develops a simple theoretical model of a rational consumer's timing of consumption decisions over the billing cycle to investigate the consequences of such consumer behavior on the profit-maximizing menu of pricing plans offered by the cellular phone service provider.

The results of the theoretical model in Chapter 2 show that the consumer's rational consumption behavior during the billing cycle can provide an explanation for why firms offer cellular plans with a three-part pricing

structure. The model's explanation does not rely on behavioral consumer anomalies that have dominated in the literature. The intuition behind the model is as follows: To differentiate among different consumer segments, the service provider needs to distort contracts for some consumers from cost in order to reduce the information rents accruing to other consumers. In essence, a high-volume consumer needs to be enticed to select a cellular plan with a high monthly fee. In order to do that, the firm has to make the plan with the smaller monthly fee less attractive to the high-volume consumer, a task that is best achieved with a three-part tariff. Consumption exceeding the allowance is discouraged through a high overage price for consumption in excess of the allowance, while consumption for the intended low-volume consumer is close to efficiency. Compared to a two-part tariff, a three-part tariff creates additional consumption inefficiencies for the high-volume consumer, thereby allowing for better differentiation among consumer segments, and ultimately higher firm profits. The sequential nature of consumption decisions over the billing cycle can reconcile three-part pricing within the rational consumer model.

Chapter 3 examines on-net/off-net price differentials in mobile telecommunication markets that follow the caller-pays principle (CPP).[2] On-net/off-net differentials can readily be found by examining calling plans offered by major cellular service providers in CPP-countries. They are extremely common in most European mobile markets for pre-pay as well as for monthly

[2]Most OECD countries follow the caller-pays principle (CPP) in which the caller pays for the cost of the mobile call. Notable exceptions to the CPP-regime are the United States, Canada, Hong Kong and Singapore, these countries follow the receiver-pays principle (RPP) in which the calling *and* the receiving party pay for the cost of the call.

packages (Harbord and Pagnozzi, 2008).[3] Such differentials between a call that terminates on the same network and one that terminates on another network give rise to network externalities: A consumer benefits if additional consumers subscribe to the same mobile network because the share of calls for which the lower on-net price applies increases. On-net/off-net differentials have been well-explored in the literature under uniform consumer calling patterns, that is, when each consumer is equally likely to call (or receive a call) from any other subscriber on any other network. However, most mobile subscribers tend to place a large fraction of their calls to a small select group of friends and family members with which they share repeat calling relationships. In addition, receiving calls conveys benefits, particularly in repeat calling relationships between a couple, close friends or business associates. Such receiving call externalities have been mostly ignored in the standard telecommunications literature.

The theoretical model presented in Chapter 3 extends the standard network competition model to incorporate closed user groups such as a couple, close friends, family or a small business. It adds two features to the standard model of network competition: (i) call externalities that arise from the concern of members of a closed user group about the cost to others of making a call to them and (ii) an own-network biased calling pattern arising from the large volume of within-group calls and the coordination of network

[3] On-net/off-net differentials are less prominent in RPP-countries such as Canada and the United States. Nonetheless, several cellular service providers - for example Fido and Rogers in Canada - offer cellular plans that include unlimited on-net calling (e.g. Fido-to-Fido/Rogers-to-Rogers plans). *Friends & Family* or *My Five/myFaves* plans where members enjoy better rates for calls to a small set of numbers share similar on-net/off-net differentials only if the set of numbers included in the group is restricted to be on the same network.

subscription choice.

The results from the model demonstrate that on-net/off-net differentials are not tied to termination charges for interconnection between networks, in fact, on-net/off-net differentials can exists even with termination charges at cost. Intuitively, networks fully internalize the call externalities of closed user group members for on-net calls, but not for off-net calls where the externality benefits primarily subscribers of the competitor's network. Hence, networks will charge a price differential even without any cost-based differences between these two types of calls.

If the two competing network operators are of different size, the model results show that the on-net/off-net differential of the larger network exceeds the differential charged by its smaller competitor. Compared to the standard model of network competition, the extended model with closed user group features even larger on-net/off-net differentials. If there exist anti-competitive concerns about the disadvantage of a smaller network if on-net/off-net differentials are large, the presence of closed user groups will only exacerbate the smaller network's disadvantage. This results stands in contrast to the commonly advanced argument that any anti-competitive effect of high termination charges and its resulting on-net/off-net differentials are mitigated by the presence of closed user groups.

The welfare results of the model show that the overall effect of on-net/off-net differentials is welfare-reducing in the presence of closed user groups. This result lends support to the contention that a ban on on-net/off-net differentials could improve welfare and replace cost-based regulation of termination charges for network interconnection.

Thesis Chapter 4 explores consumer behavior, cellular plan choice and the effects of demand uncertainty under three-part tariffs. While there exists an extensive literature on consumer behavior under two-part tariffs, there has been little exploration of choice behavior in the presence of three-part tariffs. Consumer behavior can substantially differ between two-part and three-part tariff structures in part due to the effects of demand uncertainty. The empirical analysis contributes to the literature on consumer behavior under three-part tariffs by estimating a detailed discrete/continuous demand model of cellular plan choice and consumption using a large consumer-level dataset from a major US cellular service provider. This data is combined with information on plan availability to account for temporal variation in consumers' choice sets resulting from the introduction or discontinuance of cellular plans and the practice of grandfathering cellular plans for current plan subscribers.

The estimated empirical model estimated identifies substantial and significant tariff-specific biases in consumer behavior. In particular, consumers tend to favor cellular plans with smaller allowances. This result stands in contrast to the bias previous studies in two-part tariff environments have found where consumers tend to buy communications services that they fail to consume afterwards, i.e. consumers choose plans that are "too big" given their consumption profile. Furthermore, the results suggest that the pricing parameters of three-part tariffs such as the monthly fee, the allowance of minutes included and the overage price affect consumer's plan choice to a much greater extent than their monthly consumption. Demand uncertainty is substantial among cellular consumers and is identified as one of the main

driving forces of plan-choice behavior in three-part tariff environments.

Chapter 2

The Billing Period and Consumption of Wireless Telecommunication Services

2.1 Introduction

Consumers who use cellular phone services generally sign up for a monthly plan with a cellular phone service provider. The entire length of the consumer's cellular contract might exceed one year depending on the upfront handset subsidy, but the pricing of cellular services is typically based on a billing cycle of a month. The price for consuming cellular services often depends on the amount of services already consumed within the billing cycle, for instance if the monthly plan includes an allowance of minutes for a fixed fee and a positive marginal price for consumption of minutes in excess of the allowance. Cellular phone service providers on the other hand tend to offer a menu of contracts with different (monthly) pricing features in order to differentiate among various consumer segments. This chapter develops a simple theoretical model of a rational consumer's timing of consumption

decisions over the billing cycle to investigate the consequences of such consumer behavior on the profit-maximizing menu of pricing plans offered by the cellular phone service provider. Can we reconcile commonly observed cellular pricing plans within a rational consumer model?

The model developed in this chapter combines the consumer's decision problem of when and how much to consume during the billing cycle with the screening problem of the cellular phone service provider who differentiates pricing plans among consumer segments. The consumer's rational consumption behavior during the billing cycle provides an explanation for three-part tariffs that does not rely on behavioral consumer anomalies. In essence, to screen different customer segments, the service provider needs to distort contracts for some agents in order to reduce the information rents accruing to other agents. A standard two-part tariff creates an above-marginal-cost consumption distortion, but it is not profit-maximizing since the service provider can utilize the entire price schedule to achieve the required distortion. Offering a contract with an increasing marginal price schedule implies that the effective marginal price of consumption at the beginning of the billing cycle depends on expected future consumption later in the billing cycle. Such a contract creates additional *sequential* consumption inefficiencies that allow the cellular phone service provider to achieve better type-separation, and hence higher profits.

In many settings, buyers face the choice from a set of pricing plans at a time when they are still uncertain about future consumption needs. In telecommunication markets, consumers choose among alternative long-distance or wireless plans and then - at a later stage - decide exactly how

much they wish to consume. More generally, subscription markets such as utilities, cable and telecommunications are characterized by time separation of subscription and consumption decisions, which gives rise to a multi-stage decision process. Internet access providers offer consumers a choice among several pricing plans depending on expected usage. Airlines and public transportation systems offer a range of advance purchase options of tickets with varying discount rates depending on expected travel needs. Financial service institutions offer a range of chequing and saving accounts depending on the average expected balance or the number of cheques drawn on the account. Fitness clubs charge different monthly rates depending on registration the duration of the contract.

In such markets, consumers first choose which service option they would like to sign up for based on their expected future consumption needs. Later on, as consumers learn their needs with certainty, they decide how much of the good to purchase conditional on the rates of the tariff option previously selected by them. In many instances, the service option chosen in the first stage will not be ex-post optimal. If consumers were to make service option and usage decisions simultaneously, all relevant information would be known at the time of consumption and the choice of service option would be just dual to the usage decision.

This two-stage decision process has been incorporated into the traditional nonlinear pricing literature. Courty and Li (2000) study monopolistic screening when consumers know at the time of contracting only the distribution of their valuations but subsequently learn their actual valuation for the good. In a general unit-demand framework in which the types of con-

sumers are characterized by different distributions, the authors show that it is optimal to charge a fixed fee and offer the good below marginal price to consumers with smaller valuation uncertainty in order to reduce the rents to consumers that face greater demand uncertainty. Miravete (2005) splits the asymmetric information parameter into an (additive) ex-ante type and an independent demand shock. The subscription choice is based only on the agent's knowledge of the ex-ante type while the consumption decision incorporates the independent demand shock. A range of screening mechanisms from standard nonlinear tariffs, optional two-part tariffs to fully nonlinear options are compared and their welfare performance empirically evaluated.

Although these (and other) studies separate the subscription from the consumption decision, they fail to account for the sequential nature of the consumption decision itself. In many industries, including telecommunication and utilities, consumers face nonlinear price schedules for consumption within a fixed billing period (typically a month). Within the billing period, consumers have to optimize relative to a pricing structure in which the marginal price depends on the amount already consumed in that particular period.[4] Hence, consumers' expectations regarding future consumption at a later stage in the billing period become relevant: a rational consumer will anticipate the impact of current consumption decisions on the marginal price of future consumption.

Consumption decisions made sequentially in the presence of substantial

[4] A number of studies investigate consumer behavior in the presence of nonlinear budget sets, treating the consumption decision in a one-shot framework. Hausman (1985) provides an overview of the different forms of uncertainty and the variety of econometric specifications with applications ranging from labor supply to electricity consumption and two-part tariffs.

uncertainty regarding future consumption lead to a within-period consumption pattern that cannot be ignored due to its implications on the profit-maximizing price schedule offered by the firm. Focusing exclusively on the consumers' decision problem, Keeler, Newhouse, and Phelps (1977) model the within-period consumption behavior in the presence of uncertainty in the context of health care services. Using dynamic programming, the authors show that the correct price for consumers to use when making health consumption decisions under insurance plans with deductibles, co-payment and coverage ceilings is the "effective price" - the shadow price of one more unit of consumption. In the absence of risk-aversion and income effects, Ellis (1986) establishes that the effective price is equal to the expected marginal price at the end of the accounting period.

The wireless telecommunications industry is characterized by firms offering a menu of non-linear tariffs from which the consumers select their preferred tariff option. Common cellular phone plans are variations of a three-part tariff consisting of a fixed fee, an allowance of minutes included with the fixed fee and a marginal price per minute for consumption in excess of the allowance. While cellular contracts frequently exceed one year in length, the billing period is typically a month, after which the counter is set back to zero in regards to the allowance.[5] Consumers learn about their consumption needs and make consumption choices in a sequential manner throughout the billing period. Consumption decisions at the beginning of the billing period are based on *expected* consumption later in the billing period.

[5]Some cellular plans offer "rollover" minutes. Unused minutes under the allowance can be accumulated and rolled over into the next billing periods (for up to one year).

Similar to the subscription choices in the more standard two-stage process, the consumption pattern might not be efficient ex-post. For example, a consumer might have foregone valuable consumption opportunities in the early stages of the billing period on the basis of an expected positive marginal price, but then subsequently stayed within the allowance (marginal price of zero) due to a lack of valuable consumption opportunities in the later stages of the billing period.

The model developed in this chapter is the simplest setup able to shed light on consumers' within-period consumption pattern and its effect on the structure of contracts offered by a monopolist screening agents under asymmetric information. It adapts traditional nonlinear pricing models to address the sequential nature of the consumption pattern. The main insights from the model are as follows:

Menu of Nonlinear Price Schedules: In nonlinear pricing models and in the absence of uncertainty, each type gets offered a specific quantity and a total payment associated with it. If the total payment function is concave in quantity, the profit-maximizing nonlinear schedule can be implemented through a menu of two-part tariffs.[6] In the presence of consumption uncertainty however, the entire marginal price schedule is relevant and a single nonlinear schedule is dominated by a menu of nonlinear price schedules. The simple two-type model shows that the distorted contract in the profit-maximizing menu of screening contracts exhibits increasing marginal prices, which allows for better type-separation than a standard two-part tariff.

[6] Only the total payment and its derivative are strategically relevant.

Average Price Within and Across Contracts: Sequential screening models with a one-shot consumption framework have been unable to explain the fact that firms offer a selection of tariff options for which the average per-minute price increases *within* a contract but decreases *across* the menu of contracts ordered by the quantity of minutes included in the plan allowance. The simple screening model developed in this chapter which focuses on the within-period consumption pattern can reconcile the two seemingly contradictory observations: The declining average price across contracts is equivalent to the standard screening result from optimal quantity discounts in monopolized markets Maskin and Riley (1984), while the increasing average cost within a contract is due to the within-period consumption pattern and allows for better type separation in a single-product monopolistic screening model.

Within-Period Consumption Behavior: The screening model developed in this chapter shows that agents will restrict their consumption when facing increasing marginal price schedules. This result has important implications for the empirical analysis of consumer behavior in the presence of three-part (hybrid) tariffs. Many empirical telecommunication studies have found biases in consumers' selection of tariff options, explaining them using various behavioral models ranging from flat-rate tariffs biases, to overconfidence (underestimation of variance), to naïve consumers with hyperbolic time preferences. Such reliance on behavioral explanations should not be surprising given the lack of a rational consumer model that can explain why firms find it profit-maximizing to offer three-part tariffs that exhibit increasing marginal price schedules. The model developed in this chapter suggests that if con-

sumption patterns within the billing cycle are economically relevant but are ignored in the empirical analysis, then the resulting price elasticities and tariff choices will be biased in the direction of recent empirical findings.

The rest of the chapter is organized as follows: Section 2.2 reviews some of the particular features of wireless telecommunications pricing that provide the motivation for the model that follows. The theoretical model establishes the focus on within-period consumption patterns, its resulting inefficiencies and characterization of the profit-maximizing screening contracts. Next, alternative explanations for three-part tariffs are outlined and critically discussed, followed by a section on the implications of the model of within-period consumption patterns on the empirical analysis of consumer behavior. This chapter ends with concluding remarks.

2.2 Wireless Telecommunications

In the wireless telecommunication industry, consumers are typically offered a choice among several tariff options that include a fixed fee, a monthly allowance of minutes and a price for minutes used in excess of the allowance. Various other features such as voice-mail, unlimited weekend calling, data, etc. might be offered along with the bundle, but the basic structure of cellular contracts is surprisingly consistent. Table 2.1 presents a selection of cellular phone contracts offered in Canada:[7]

[7]The cellular pricing information was taken from the websites of major Canadian cellular phone companies, as of April 2008. Cellular pricing is similar in the United States and the reader may confirm that these type of cellular tariffs still prevail. The broad structure of wireless contract pricing with an allowance and a marginal price in excess of the allowance can also be observed all across Europe, even though Europe - in contrast to

Company	**Allowance** monthly	**Fee** monthly	**Minimum Average Price**[A] per minute in cents	**Additional Time** per minute in cents	**Ratio** (Col 5/4)
Bell	150	30	20.0	30	1.50
	250	40	16.0	30	1.87
	500	60	12.0	25	2.08
	850	100	11.8	20	1.69
	1,250	150	12.0	20	1.67
Telus	150	30	20.0	30	1.50
	400	50	12.5	25	2.00
	650	75	11.5	25	2.17
	800	100	12.5	25	2.00
	1,250	150	12.0	25	2.08
Rogers	150	30	20.0	35	1.75
	250	40	16.0	30	1.86
	500	60	12.0	30	2.50
	800	100	12.5	20	1.60
	1,250	150	12.0	20	1.67

[A] The minimum average price occurs at the allowance for all contracts.
Allowances and rates are for daytime, weekday calls with domestic long distance included.
Source: Websites of Cellular Phone Companies as of April 2008.

Table 2.1: Cellular Phone Tariffs in Canada (2008)

Several characteristics of these contracts stand out: First, for a fixed fee, all of these contracts have an allowance of included minutes and a fixed marginal price for usage in excess of the allowance (overage price), which implies that marginal prices are (weakly) increasing with the quantity consumed. Second, the marginal price per minute for usage in excess of the monthly allowance is high and substantially exceeds the minimum average price per minute within the allowance. Consequently, *within* a contract, the

North America - follows the caller-pays-principle.

average price is increasing for consumption in excess of the allowance (quantity premiums). Nonetheless, the average per minute price *across* contracts (outer envelope) is declining, an observation in line with standard results from optimal quantity discounts in monopolized markets Maskin and Riley (1984). Both of these characteristics are illustrated in Figure 2.1 which plots the average price per minute for the cellular plans offered by Rogers.

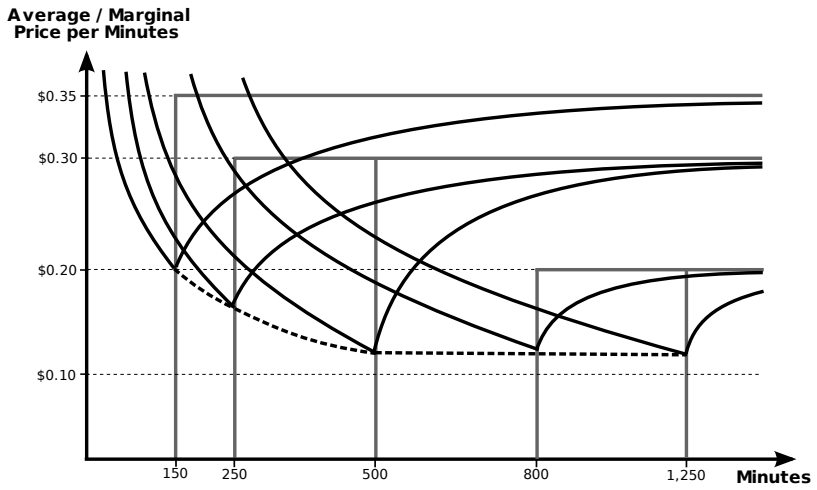

Figure 2.1: Cellular Plans offered by Rogers Wireless (2008).

Interestingly, cellular data plans for email and web use often have an identical pricing structure as illustrated in Table 2.2. The fixed monthly fee includes a data allowance coupled with high marginal rates for usage in excess of the allowance which leads to steeply increasing average prices once the data transfer allowance is used up:

TELUS Data Plans	**Fee** monthly	**Data** MB	**Minimum Average Price**[A] per MB in $	**Additional Data** per MB in $	**Ratio** (Col 5/4)
Email&Web 25	25	4	6.25	12	1.92
Email&Web 40	40	8	5.00	8	1.60
Email&Web 60	60	30	2.00	6	3.00
Email&Web 100	100	1024	0.098	3	30.61

[A] The minimum average price occurs at the data allowance for all data plan.
Source: http://www.telusmobility.com/bc/business_solutions/handheld_rate_plans.shtml(accessed May 4, 2008).

Table 2.2: Cellular Data Plans Offered by TELUS (2008)

Models with valuation heterogeneity across consumer types commonly result in a menu of tariff options offered but they fail to explain the fact that within a contract the per-minute price *increases* with consumption while the per-minute price decreases across contracts with larger allowances. In the literature, this phenomenon has been attributed to consumer irrationality of some form: naïveté, hyperbolic discounting, misperception of future consumption patterns or a systemic flat-rate bias.

Many field experts (and some economists) argue that there exist important "particularities" in telecommunication markets that make consumer susceptible to deviations from rational behavior. Most economists however are uncomfortable relinquishing the rational consumer model in which mistakes commonly can only be of transitory nature in a dynamic process in which consumers eventually learn about their consumption patterns. Under this view, frequent decisions made by millions of consumers in telecommunication markets should not be "anomalies" outside the realm of common economic principles. How can one then reconcile firms offering a selection

of tariff options for which the per-minute price increases *within* the contract but decreases *across* contracts? Can observed patterns of cellular pricing be explained within a rational consumer framework, or why don't cellular service providers simply price "per-minute" using two-part tariffs?

Although cellular contracts are the most prominent and widespread example of this particular pricing scheme involving allowances, similar pricing practices have been observed in the market for computer operating systems as well as in utility markets (electricity and water). By 1992, Microsoft's licensing agreements with original equipment manufacturers (OEM) for its MS-DOS operating system included a CPU-license and involved long-term contracts with minimum commitments. A minimum commitment on part of OEM's consisted of a fixed (upfront) payment for which units up to the minimum commitment are traded at zero marginal price and units in excess of the minimum commitment are traded at a positive marginal price. Such a minimum commitment is equivalent to an allowance-type contract observed in the telecommunications industry (Woroch, Warren-Boulton, and Baseman, 1998).[8] Similarly, utility contracts involving electricity and water pricing are often characterized by block-pricing including an allowance (minimum charge), where usage in excess of the allowance is priced based on a

[8] In 1994, the Department of Justice in the United States filed a civil antitrust complaint arguing that Microsoft used exclusionary and anti-competitive licensing practices for its MS-DOS operating system sold to original equipment manufacturers (OEM). The consent decree signed bans three types of licensing provisions: per-processor licenses, lump-sum pricing, and long-term contracts (exceeding one year) with large minimum commitments. Minimum commitments were banned based on the argument that if sales are below the minimum commitment, a double royalty effectively exists. Furthermore, the carry-over of unused portion of the minimum commitments prolongs the agreement by increasing switching cost. The final judgement determined that apart from foreclosure, no rationale for minimum commitments (allowance-type contracts) exists for a monopolist.

constant marginal price for a (quantity) block that might be increasing or decreasing across blocks. However, utilities typically offer only a single price schedule, potentially with lifeline rate discounts for households with low income. Despite the pricing similarities, the rationale is more likely based on efficient management of peak-loads and shortages rather than the screening of heterogeneous consumers.[9]

2.3 The Model

The monopolist (principal) produces a non-storable good with fixed cost f and constant marginal costs of production $c > 0$. The principal sells the good to a buyer (agent) whose type belongs to the set $\Delta = \{\theta_L, \theta_H\}$. The type is private information to the agent and the probability of each type λ_{θ_L} and $\lambda_{\theta_H} = 1 - \lambda_{\theta_L}$ respectively is common knowledge. There is no arbitrage possible among different types of agents. The monopolist offers a menu of contracts that specify a fixed (upfront) fee F and a marginal price schedule for consumption over the accounting period.[10] The exogenously fixed billing period is of length T and in order to keep the model simple and tractable, $T = 2$.[11] This setup gives the simplest, non-trivial model that allows for

[9] See for example Olmstead, Hanemann, and Stavins (2006), who estimate an empirical model of household water demand under increasing block rates.

[10] In standard screening models without uncertainty, the discussion often centers around a menu of tariffs that implements the (single) nonlinear pricing schedule. In the presence of consumption uncertainty, a single nonlinear price schedule however is not equivalent to a *menu* of price schedules due to the separation of tariff choice and consumption choice. Hence, the optimization problem has to allow for a nonlinear price schedule for each consumer type.

[11] The focus on the consumer's consumption pattern within the billing period raises the question of the optimal length of the billing period. One month is the standard billing period in telecommunications and utilities. Contracting cost and/or other institutional reasons might have given rise to this pattern and given its prevalence this chapter will abstract from that issue and simply treat the billing period as exogenous. More generally though, one could envision an environment in which the principal chooses the length of

examination of the within-period consumption pattern.[12] In each subperiod t, the consumer has unit demand and derives instantaneous utility, which is equal to the surplus derived from the good:

$$u_t = \begin{cases} v_t - p & \text{if she buys unit a price } p \\ 0 & \text{if she does not buy} \end{cases}$$

where the reservation utility is normalized to zero.[13] Both principal and agent are risk-neutral and there is no discounting within the billing period. Hence, consumer net utility for the entire billing period is simply the sum of instantaneous utility minus the fixed fee F:

$$U = \sum_t u_t - F.$$

For each type θ_i, the valuation in each period is an independent draw from a twice continuously differentiable cumulative distribution G_i on $[\underline{v}, \bar{v}]$ with density $g_i(v)$.[14]

the billing period.

[12] This simple model focuses exclusively on the effect of a (single) nonlinear pricing schedule on consumer behavior. It ignores additional details such as off-peak pricing (evenings and weekends) or bundle pricing where cellular services are sold as part of a bundle with other communications services (TV and internet).

[13] The utility is assumed to be of quasi-linear form in order to abstract from wealth effects.

[14] The valuations for the two agent-types are drawn from two different distributions, G_{θ_L} and G_{θ_H}. But each type draws from the same type-specific distribution in both periods, i.e. the agent's valuation *distributions* are perfectly correlated across the two subperiods. The model's results are expected to hold up if agents' valuations are imperfectly correlated over time (e.g. a high valuation in the first subperiod is likely to be followed by a low valuation in the second subperiod). If the valuations themselves however are perfectly correlated in the two subperiods, no consumption uncertainty remains after the decision on first subperiod consumption and the model would become identical to a one-shot consumption model.

Assumption 2.1 (First-Order Stochastic Dominance)

The distribution G_{θ_H} first-order stochastically dominates the distribution G_{θ_L}, that is, $G_{\theta_L}(v) \geq G_{\theta_H}(v)$ for every v.

The assumption of first-order stochastic dominance implies that the Spence-Mirrlees single crossing condition holds.[15] At the time of contracting, $t = 0$ (before the start of the billing period), the distribution of valuations for each type is common knowledge. In period $t = 1$, once the agent has chosen a contract (tariff option), she learns about her exact valuation v_1 for subperiod 1 consumption, and must make a decision on whether to consume the good. The valuation is time-specific to subperiod 1, i.e. the good is non-storable. If the consumer decides to forego consumption in subperiod 1, the valuation associated with that call opportunity is "lost" and cannot be transferred into the next subperiod. Again in subperiod $t = 2$, she learns her exact valuation for subperiod 2 consumption v_2 and must decide whether to consume the unit. The timing of the model is illustrated in Figure 2.2:

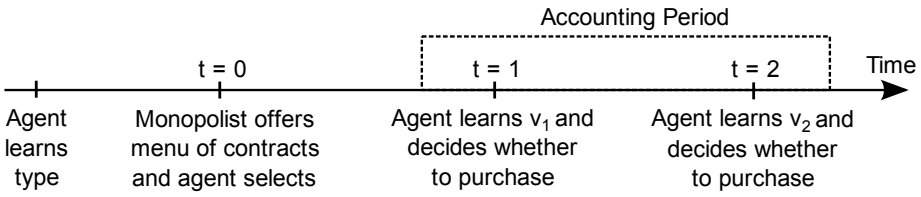

Figure 2.2: Timing of the Model

[15] The Spence-Mirrlees single-crossing condition (sorting condition) requires that the indifference curves of the different types cross only once. Consumer utility is quasi-linear and the first-order stochastic dominance assumption implies that the (ex-ante) marginal utility of type θ_H exceeds the marginal utility of type θ_L.

The monopolist is assumed to have full commitment power for the length of the billing period, which means it can offer contracts that cover the entire billing period and cannot be breached or renegotiated.[16] Furthermore, dynamic contracting issues arising from repeated interactions over multiple billing periods are ignored for the purpose of this study.

2.3.1 Consumer Behavior

Rational consumers anticipate the impact of current consumption decisions on the marginal price of future consumption within the billing period. Hence, their first subperiod decision ($t = 1$) is based on expectations about consumption in the second subperiod and follows a "threshold" rule. Given the model setup, agents consume at most two units of the good and the marginal price schedule is fully characterized by the two prices for the first and second unit, p_1 and p_2. To avoid any confusion, it is important to note that p_1 is the price for the *first unit* and if the consumer decides to forego consumption in the first sub-period, the applicable price in the second sub-period is (still) p_1 since the consumer purchases the *first* unit. If the price for the first and second unit are equal, $p_1 = p_2$, then the consumer will decide to buy a unit whenever the first realized valuation v_1 exceeds the marginal price p_1. If however, $p_1 < p_2$, then the consumer will decide to forgo the purchase on

[16] Full commitment guarantees that the revelation principle holds and leads to immediate revelation of information on part of the informed agent (consumer) (Myerson, 1979, 1986; Harris and Townsend, 1981). Screening contracts in this model are not renegotiation-proof. Optimal pricing for some agents will diverge from marginal cost, thereby giving rise to pareto-improving renegotiation opportunities within the billing period. It seems entirely plausible though that administration costs make it unprofitable for cellular service providers to take advantage of ex-post inefficiencies within the billing period through contract renegotiation.

the first subperiod if the valuation only marginally exceeds p_1 in order to increase the expected surplus from consumption in the second subperiod. If on the other hand $p_1 > p_2$, then the agent might decide to purchase the unit even though the realized valuation in the first subperiod is slightly below p_1 to increase the expected surplus from consumption of the second unit. The agent will consume the first unit in subperiod $t = 1$ whenever her valuation exceeds the threshold $\widehat{p}_i$ for $i = \theta_L, \theta_H$, which depends on the agent's type and the full marginal price schedule $p(p_1, p_2)$. The threshold is given as

$$\widehat{p}_i = p_1 + \int_{p_1}^{\overline{v}} (v - p_1)\, g_i(v)\, dv - \int_{p_2}^{\overline{v}} (v - p_2)\, g_i(v)\, dv \quad \text{for } i = \theta_L, \theta_H.$$

Lemma 2.1 (Threshold Consumption Behavior)

(i) $\widehat{p}_i > p_1$ *if and only if* $p_1 < p_2$.

(ii) $\widehat{p}_{\theta_H} > \widehat{p}_{\theta_L}$ *if and only if* $p_1 < p_2$.

Proof. *See Appendix A.*

The threshold behavior in Lemma 2.1(i) indicates that if $p_1 > p_2$, the agent might consume in the first period even though his realized valuation is below the marginal price. On the other hand, if $p_1 < p_2$, the consumer might not consume in the first period even though his valuation exceeds the marginal price. The second part of the lemma shows that the cutoff of the high-type agent (θ_H) is more responsive to the differential in marginal prices:

If the marginal price schedule is upward sloping ($p_1 < p_2$), then the high-type agent will have a higher first subperiod threshold than the low-type agent on the same price schedule and her consumption in the first subperiod will be more "cautious". On the other hand, faced with an downward sloping marginal price schedule, the high-type agent's threshold will be lower and her consumption in the first subperiod more "aggressive".

The consumer's first period threshold $\widehat{p}_i$ is the effective marginal price of consumption. Lemma 2.1 shows that the effective price of first-period consumption $\widehat{p}_i$ exceeds the marginal price p_1 whenever marginal prices are increasing ($p_1 < p_2$). On the other hand, the effective price $\widehat{p}_i$ is lower than the marginal price p_1 if marginal prices are decreasing ($p_1 > p_2$). The second part of Lemma 2.1 means that the effective price of first-period consumption for the high-type agent is higher (lower) than the effective price for the low-type agent whenever marginal prices are increasing (decreasing). In essence, heterogeneous agents face different effective prices of first-period consumption and the profit-maximizing menu of screening contracts will account for that.

Consumption inefficiencies clearly exist for a naïve consumer who consumers a unit whenever the valuation exceeds the marginal price. A rational consumer anticipates that the consumption of a unit will alter the marginal price of future consumption units and take this into account. However, an increasing marginal price schedule can create two types of consumption inefficiencies ex-post even with a fully rational consumer:

(i) Ex-post Overconsumption: If the first-period valuation exceeds the effective marginal price (threshold) but not the marginal price for the second unit, $\widehat{p}_i < v_1 < p_2$, and subsequently the second period valuation exceeds the marginal price for the second unit, $v_2 > p_2$, then the consumer's first period consumption valuation is below the (ex-post) marginal price of consumption. The marginal price of consumption ex-post is p_2, but the valuation of the unit consumed in the first period v_1 was below the marginal price ex-post. Such consumption behavior is fully rational since at the time the consumer makes the consumption decision in the first period, the second period valuation is unknown.

(ii) Ex-Post Underconsumption: If the first period valuation exceeds the price of the first unit but not the *effective* marginal price, $p_1 < v_1 < \widehat{p}_i$, and the second period valuation falls short of the price for the first unit, $v_2 < p_1$, then the consumer's (non-realized) first period consumption valuation exceeds the (ex-post) marginal price of consumption. The marginal price of consumption ex-post is p_1, but the valuation of the unit *not* consumed in the first period was above the marginal price ex-post. Again, such consumption behavior is fully rational due to the uncertainty of the second period valuation v_2 at the time of deciding whether to consume in the first period.

The same two types of consumption inefficiencies exists for decreasing marginal price schedules (only reversed). This distinctive consumption pattern over the two subperiods with its resulting consumption inefficiencies is at the heart of this sequential screening model. It is neither present with two-part tariffs nor in the standard model with a one-shot consumption

framework.[17]

2.3.2 Profit-Maximizing Menu of Screening Contracts

A screening contract consists of a fixed fee F paid upfront (independent of consumption) and marginal prices p_1 and p_2 to consume the first and second unit.[18] Hence, the set of screening contracts considered includes fixed fee contracts with marginal price equal to zero ($p_1 = p_2 = 0$), two-part tariff contracts for which the marginal price of the first unit is equal to the marginal price of the second unit ($p_1 = p_2$) and more general three-part tariff contracts for which the marginal prices for the first and the second unit are not equal. Regardless of the fixed fee F, the consumer will consume in the first sub-period if the valuation exceeds the cutoff $\widehat{p}$, and in the second sub-period if the realized valuation exceeds the applicable marginal price: If the agent did not consume in the first sub-period, the marginal price is p_1, if the agent did consume, the marginal price is p_2. The principal offers two contracts $\{F_{\theta_L}, p_{\theta_L}(p_1, p_2), F_{\theta_H}, p_{\theta_H}(p_1, p_2)\}$ and the profit maximization problem can be written as

[17] If the consumption decision takes place at a single point in time, the valuation of the marginal unit (last unit consumed) is always below the valuation of inframarginal units and above the valuation of all opportunities not realized. At the (single) time of the consumption, all demand uncertainty is assumed to be resolved in the standard model.

[18] The marginal prices form a quantity schedule and are not the prices applicable in the two subperiods. Given the two consumption subperiods are identical, the good non-storable and the valuation time-specific, setting different marginal prices for the two periods is not optimal.

$$\max_{\substack{F_{\theta_L}, p_{\theta_L}(\cdot) \\ F_{\theta_H}, p_{\theta_H}(\cdot)}} \sum_{i=\theta_L,\theta_H} \lambda_i \left[F_i + (p_{1i} - c)\left(1 - G_i\left(\widehat{p}_i\right) G_i\left(p_{1i}\right)\right) + (p_{2i} - c)\left(1 - G_i\left(\widehat{p}_i\right)\right)\left(1 - G_i\left(p_{2i}\right)\right)\right]$$

subject to

$$-F_i + \int_{\widehat{p}_i}^{\overline{v}} (v - p_{1i})\, dG_i(v) + G_i\left(\widehat{p}_i\right) \int_{p_{1i}}^{\overline{v}} (v - p_{1i})\, dG_i(v) + \left(1 - G_i\left(\widehat{p}_i\right)\right) \int_{p_{2i}}^{\overline{v}} (v - p_{2i})\, dG_i(v) \quad \geq \quad 0 \qquad \forall i = \theta_L, \theta_H$$

and for $\forall i \neq i'$

$$\begin{aligned} & -F_i + \int_{\widehat{p}_i}^{\overline{v}} (v - p_{1i})\, dG_i(v) + G_i\left(\widehat{p}_i\right) \int_{p_{1i}}^{\overline{v}} (v - p_{1i})\, dG_i(v) \\ & + \left(1 - G_i\left(\widehat{p}_i\right)\right) \int_{p_{2i}}^{\overline{v}} (v - p_{2i})\, dG_i(v) \\ \geq \quad & -F_{i'} + \int_{\widehat{p}_{i|i'}}^{\overline{v}} (v - p_{1i'})\, dG_i(v) + G_i\left(\widehat{p}_{i|i'}\right) \int_{p_{1i'}}^{\overline{v}} (v - p_{1i'})\, dG_i(v) \\ & + \left(1 - G_i\left(\widehat{p}_{i|i'}\right)\right) \int_{p_{2i'}}^{\overline{v}} (v - p_{2i'})\, dG_i(v) \end{aligned}$$

The first set of constraints are the ex-ante individual rationality constraints[19] and the second set of constraints are the (ex-ante) incentive com-

[19]Ex-ante individual rationality is weaker than ex-post individual rationality since the utility of the agent is allowed to be negative in certain states of the world as long as

patibility constraints.

Lemma 2.2 (Individual Rationality)

IR_{θ_L} and IC_{θ_H,θ_L} imply IR_{θ_H}.

Proof. *See Appendix A.*

The high-type θ_H gets more utility than low-type θ_L from any particular contract. Lemma 2.2 also implies that the individual rationality constraint of the low-type IR_{θ_L} holds with equality for the profit-maximizing screening contract, otherwise increasing the fixed fees F_{θ_H} and F_{θ_L} by the same amount would increase profits. In addition, the incentive compatibility constraint if the high-type IC_{θ_H,θ_L} holds with equality, otherwise profits could be increased by increasing F_{θ_H}. Substituting IR_{θ_L} and IC_{θ_H,θ_L} into the principal's objective function and ignoring IC_{θ_L,θ_H} results in the "relaxed" problem

$$
\begin{aligned}
\max_{p_{\theta_L}(\cdot),p_{\theta_H}(\cdot)} \sum_{L,H} \quad & \lambda_{\theta_i} \int_{\widehat{p}_{\theta_i}}^{\overline{v}} (v-c)\, dG_{\theta_i}(v) + \lambda_{\theta_i} G_{\theta_i}(\widehat{p}_{\theta_i}) \int_{p_{1\theta_i}}^{\overline{v}} (v-c)\, dG_{\theta_i}(v) \\
& + \lambda_{\theta_i} \left(1 - G_{\theta_i}(\widehat{p}_{\theta_i})\right) \int_{p_{2\theta_i}}^{\overline{v}} (v-c)\, dG_{\theta_i}(v) \\
& - \lambda_{\theta_H} \Bigg[\int_{\widehat{p}_{\theta_H|\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_H}(v) + G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right) \int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right) \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_H}(v) - \int_{\widehat{p}_{\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_L}(v) \\
& - G_{\theta_L}(\widehat{p}_{\theta_L}) \int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_L}(v) - \left(1 - G_{\theta_L}(\widehat{p}_{\theta_L})\right) \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_L}(v) \Bigg]
\end{aligned}
$$

non-negative ex-ante utility is guaranteed.

Denote the low-type's surplus by $S_{\theta_L}(p_{\theta_L})$ and the information rent for the high-type by $R_{\theta_H}(p_{\theta_L})$, both as functions of the marginal price schedule for the low-type:

$$S_{\theta_L}(p_{\theta_L}) = \int_{\widehat{p}_{\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L}(v) + G_{\theta_L}(\widehat{p}_{\theta_L}) \int_{p_{1\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L}(v) + (1 - G_{\theta_L}(\widehat{p}_{\theta_L})) \int_{p_{2\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L}(v)$$

$$\begin{aligned} R_{\theta_H}(p_{\theta_L}) = & \int_{\widehat{p}_{\theta_H|\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_H}(v) \\ & + G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right) \int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_H}(v) \\ & + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right) \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_H}(v) - \int_{\widehat{p}_{\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_L}(v) \\ & - G_{\theta_L}(\widehat{p}_{\theta_L}) \int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_L}(v) - (1 - G_{\theta_L}(\widehat{p}_{\theta_L})) \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_L}(v) \end{aligned}$$

Notice that the marginal price for the high-type agent θ_H is unrestricted and the profit-maximizing solution has therefore $p_{\theta_H}\left(p^1_{\theta_H}, p^2_{\theta_H}\right) = c$ to maximize the surplus to the high-type θ_H. The marginal price for type θ_L then maximizes the surplus from the low-type θ_L minus the rent to the high-type θ_H:

Proposition 2.1 (Profit-maximizing Screening Contract)
The profit-maximizing screening contracts are characterized by marginal price schedules

$$p^*_{\theta_H}\left(p_{1\theta_H}, p_{2\theta_H}\right) = (c, c)$$

and

$$p^*_{\theta_L}\left(p_{1\theta_L}, p_{2\theta_L}\right) = \arg\max_{p(\cdot)} \left[\lambda_{\theta_L} S_{\theta_L}\left(p_{\theta_L}\right) - \lambda_{\theta_H} R_{\theta_H}\left(p_{\theta_L}\right)\right].$$

The fixed fee of the low-type's contract extracts the entire ex-ante surplus (IR_{θ_L} binds) while the fixed fee of the high-type's contract is determined by the binding incentive compatibility constraint (IC_{θ_H,θ_L}).

Proof. *See Appendix A.*

In line with standard screening results in the literature, Proposition 2.1 shows that there is no consumption distortion "at the top", i.e. the contract for the high-type θ_H exhibits marginal cost pricing. The distorted contract offered to the low-type θ_L balances surplus extraction $S_{\theta_L}\left(p_{\theta_L}\right)$ from the low-type θ_L with the minimization of information rents $R_{\theta_H}\left(p_{\theta_L}\right)$ to the high-type θ_H.

In the absence of uncertainty, the restriction of profit-maximizing contracts to a menu of two-part tariffs is without loss of generality. However, given the agent's consumption uncertainty, two-part tariffs do not necessarily allow for optimal type separation. In fact, Proposition 2.2 indicates that the principal can achieve better type separation by adjusting the marginal

price schedule of the distorted contract:

Proposition 2.2 (Improvement on Profit-maximizing Two-Part Tariff)

(i) *The distorted contract in the profit-maximizing menu of contracts restricted to two-part tariffs sets a markup on the contract for low-type* θ_L *equivalent to*

$$p_{\theta_L}^{TPT} - c = \frac{\lambda_H \left(G_{\theta_L} \left(p_{\theta_L}^{TPT} \right) - G_{\theta_H} \left(p_{\theta_L}^{TPT} \right) \right)}{\lambda_L g_{\theta_L} \left(p_{\theta_L}^{TPT} \right)}.$$

(ii) *Starting from the profit-maximizing two-part tariff for the low-type* ($p_{\theta_L}^{TPT}$)*, lowering the marginal price on the first unit* $p_{1\theta_L}$ *and increasing the marginal price of the second unit* $p_{2\theta_L}$ *increases the principal's profits, that is,*

$$\left.\frac{d\Pi}{dp_{1\theta_L}}\right|_{p_{\theta_L}^{TPT}} < 0 \quad \textit{and} \quad \left.\frac{d\Pi}{dp_{2\theta_L}}\right|_{p_{\theta_L}^{TPT}} > 0.$$

Proof. *See Appendix A.*

The low-type agent's contract needs to be distorted in order to reduce the information rents accruing to the high-type agent. However, since the principal can utilize the whole schedule to achieve this distortion, a two-part tariff is not profit-maximizing. Offering the low-type agent a contract with an upward sloping marginal price schedule achieves optimal type-separation. The two-part tariff does not create any *sequential* consumption inefficiencies since first subperiod consumption does not affect the marginal price of future consumption, It only creates an above-marginal-cost consumption distortion for the low-type agent. An upward sloping marginal price schedule however creates sequential consumption inefficiencies that allow for better

type-separation. Consequently, screening different consumers with upward sloping marginal price schedules also enlarges the parameter range for which screening (rather than pooling) is profit-maximizing for the monopolist.

Numerical Example of Profit-maximizing Screening Contracts

To illustrate the profit-maximizing screening contracts, consider the following simple numerical example: Suppose that in each subperiod, a consumption opportunity arises with probability ρ_i with $i = \theta_L, \theta_H$ and $\rho_H > \rho_L$ and the valuation v is uniformly distributed on the interval $[0, 1]$. Figure 2.3 demonstrates that the valuations of the high-type agent θ_H first-order stochastically dominate the valuations of the low-type agent θ_L, consistent with the Assumption 2.1 of the theoretical model developed above:

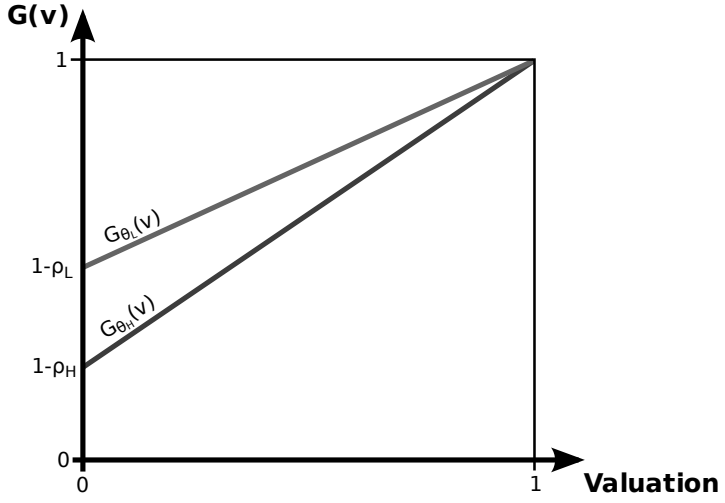

Figure 2.3: Distribution of Consumption Valuations v

Table 2.3 presents the profit-maximizing two-part tariff and compares it with the profit-maximizing nonlinear screening contract:

Numerical Example	Two-Part Tariff	Optimal Screening
F_{θ_H}	0.5731	0.5787
p_{θ_H}	0	0
F_{θ_L}	0.3781	0.4050
$p_{1\theta_L}$	0.1304	0.0746
$p_{2\theta_L}$	0.1304	0.1939
Π	0.5078	0.5097

Parameters: $\lambda_{\theta_L} = 0.8$, $c = 0$, $\rho_H = 0.8$, $\rho_L = 0.5$

Table 2.3: Profit-maximizing Two-Part Tariff and the Profit-maximizing Nonlinear Screening Contract

The optimal nonlinear screening contracts maximize the monopolist's profits: As the numerical example demonstrates, the undistorted contract for the high-type agent θ_H features marginal cost pricing, while the distorted contract of the low-type agent θ_L features a lower price on the first unit compared to the profit-maximizing two-part tariff, but a substantial markup on the second unit. This increasing marginal price schedule allows for better type-separation and thereby increases the monopolist's profits compared to screening agents with a simple menu of two-part tariffs.

This theoretical screening model is related to dynamic models of adverse selection with full commitment.[20] The two consumption subperiods of the model can be interpreted as two separate periods with the nonlinear tariff as a long-term contract spanning the entire billing period. In adverse selection models with permanent (fixed) types, the optimal long-term contract is a replica of the one-period optimal contract (Laffont and Martimort, 2002).[21]

[20] See Laffont and Martimort (2002) for a general treatment of dynamic contracting models, Baron and Besanko (1984) analyze optimal contracting with correlated types and full commitment and Laffont and Tirole (1996) provide an application to the regulation of pollution.

[21] Such models are often referred to under the terminology of false dynamics since the

Since consumers have unit demand in each subperiod, the optimal one-shot contract is a menu of two-part tariffs: The high-type agent θ_H will be able to purchase a single unit at marginal cost while the marginal price for the low-type agent θ_L will be distorted upwards from marginal cost and the fixed fees set such that all surplus is extracted from the low-type agent and the high-type agent's incentive compatibility constraint is binding. The replica of the one-shot optimal contract spanning the entire billing period is the optimal two-part tariff characterized in Proposition 2.2(i).

However, the valuation uncertainty present with sequential consumption decisions implies that this model is much more closely related to adverse selection models with correlated "types". A high first subperiod valuation makes it more likely that the agent is of the high-type θ_H, which then means that the second subperiod valuation is also likely to be high since the agent's type is fixed. The realization of consumption in the first subperiod conveys information about future (second subperiod) consumption opportunities. This gives rise to dynamic incentive constraints spanning the entire billing period and allows the monopolist to improve the terms of the rent extraction - efficiency trade-off. The optimal long-term contract will no longer be the simple replica of the one-shot contract (two-part tariff), but rather involves an increasing marginal price schedule in the distorted contract for the low-type agent.

optimal dynamic contract is analogous to the optimal static contract.

2.3.3 Discussion

The Scope of Three-Part Tariffs

While three-part tariffs are ubiquitous in the wireless telecommunications industry and in some other subscription markets (e.g. internet access), few other products and services feature similar pricing. If three-part tariffs are such a prominent pricing feature to screen different customer segments, why don't we observe more services and goods priced using three-part tariffs with an allowance and a high marginal price for consumption in excess of the allowance? For the implementation of three-part tariff pricing one needs to be able to measure individual consumption and preclude resale opportunities, but these basic requirements are fulfilled in many other instances yet there are relatively few goods and services priced using three-part tariffs. The screening model with sequential consumption decisions developed in this study provides an explanation for three-part tariffs based on rational consumer behavior, but it can also help explain the *absence* of a similar pricing structure for other products and services. The following thought experiment will illustrate the limited scope of three-part pricing outside telecommunications and other subscription markets and highlight the distinctive feature of cellular services that could explain the widespread adoption in this industry.

Consider an amusement park that offers a number of different rides or other entertainment attractions. To screen heterogeneous tourists (and their children), the park could potentially offer various pricing options resembling three-part tariffs: For a fixed entry fee, a number of rides could be included (allowance), coupled with a high marginal price for additional rides to induce

more enthusiastic amusement park visitors to purchase the more comprehensive package that features a higher fixed entry fee but includes more rides. Whether such a menu of three-part tariffs leads to increased profitability compared to a more simpler menu of two-part tariffs hinges on the presence consumption inefficiencies.

There are no consumption inefficiencies as long as visitors can choose among the rides in a way that allows them to pick the rides they find the most exciting ones. This is clearly the case if tourists have visited the amusement park before, or if the park offers brochures or maps outlining the various rides available. Even if tourists have little information on the rides prior to entry of the park, typical tourist behavior would lead them to walk around the park, take a look at the different rides before deciding which ones to enjoy. In either case, the tourists will choose the most valuable rides and pass on rides for which their value is below the marginal price. Tourist may - upon seeing the type of rides offered - wish to have purchased a different entry ticket, but the consumption of rides is efficient *given* the chosen entry ticket. If tourist are rational, three-part pricing offers little benefit over simpler two-part pricing.

Under what circumstances could three-part tariffs provide better screening opportunities? Suppose that tourists do not know the type of rides, but only the number of total different rides offered in the park. Furthermore, assume that tourists have to decide whether to take the next ride before seeing other rides and they won't be able to take that particular ride later if they pass it up initially. In this situation, consumption can exhibit inefficiencies: At the exit of the park (ex-post), tourists might wish to have taken a dif-

ferent selection of rides *even given their chosen entry ticket.* In essence, the sequential decision process with uncertainty about the value of future rides implies that the value of rides chosen could be below the marginal price, or that the value of missed rides could be above the marginal price upon exiting the park.

Three-part tariffs, or increasing marginal price schedules in general, can be beneficial to a firm's profitability only if they can create sequential consumption inefficiencies for a *given tariff plan* identical to the ones described in Section 2.3.1. While the amusement park environment does not lend itself particularly well for three-part pricing, the cellular industry and generally subscription markets that feature billing periods extending over a considerable time frame with consumption uncertainty provides the optimal environment for three-part pricing.

Length of Billing Period

The theoretical model illustrates that the monopolist can increase profits by offering a "long-term" contract spanning the entire billing period rather than separately contracting in each subperiod. In essence, the dynamic incentive conditions improve the rent extraction-efficiency trade-off of screening and the monopolist is expected to benefit even further from extending the billing period. Why don't we then observe contracts that feature billing periods extending past a single month?

The profit-maximizing screening contracts in the model are not renegotiation-proof. Since the pricing of the distorted contract diverges from marginal cost pricing, pareto-improving renegotiation opportunities arise within the billing

period. For such screening contracts to be viable, the monopolist needs the ability to fully commit to a set of contracts for the entire billing period and refuse to renegotiate at intermediate stages. It seems entirely plausible that administrative costs make it unprofitable for cellular service providers to take advantage of ex-post inefficiencies within a monthly billing period.

Automatic Tariff Adjustment

Uncertainty is at the heart of the model's explanation of observed cellular pricing based on the sequential nature of consumption decisions over the billing period. Three-part tariffs are not just implementing the (single) profit-maximizing nonlinear price schedule the way two-part tariffs do in the standard screening model without uncertainty. In fact, the *menu* of nonlinear tariffs is fundamentally different from a single nonlinear tariff.

The tariff option chosen by the consumer may not be cost-minimizing ex-post, sometimes the consumer could have saved money if she had chosen a different tariff option at the beginning of the period. The fact that consumers can switch their tariff option for the next billing period or that some cellular service providers even inform customers if their tariff option is non-optimal given their consumption pattern does not invalidate the model's explanation of observed cellular pricing. However, the sequential within-period consumption pattern fails as an explanation for three-part tariffs if cellular service providers were to offer concurrent tariff plan adjustments such that the customer is automatically adjusted to the cost-minimizing tariff for that particular billing period. Although some cellular companies have run trials with automatically adjusting tariff options, such automatic tariff adjustment

- despite its simplicity - is very rarely observed.[22]

Rollover Minutes

Some cellular companies offer so-called "rollover" minutes or an additional one-time allowance of minutes upon signing of a long-term contract lasting one year or more. Rollover minutes are unused minutes under the allowance that can be accumulated and rolled over into the next billing period (for up to one year), while a one-time allowance of minutes is a separate allocation of free (included) minutes that can be used if the monthly allowance has been exhausted. The high overage rates then only apply if the monthly *and* the one-time allowances of minutes have been used up.

Rollover minutes are not widespread in the cellular industry and one-time allowances are used primarily as a promotional tool. Both have the effect of "smoothing" the total payment over multiple billing cycles when customers face high overage charges for consumption in excess of the monthly allowance. In essence, they allow for increased consumption variability without increasing the total payment. While such arrangements do affect how consumption patterns of different consumer-types translate into the profit-maximizing menu of screening contracts, the underlying structure of the within-period consumption problem is unchanged: Even with rollover minutes and one-time allowances, the high-type consumer is more likely to face substantial overage charges and hence has a higher expected marginal price relative to the low-type consumer and the rationale for increasing marginal

[22] For example, Sprint/Nextel offered an *Auto Adjust Fair & Flexible Plan* in some trial markets in the United States in 2006, but the tariff trial was discontinued and the cellular plan with automatic tariff adjustment was never offered nationwide.

price schedules and three-part tariffs persist.

This simple theoretical model illustrates the rationale behind contracts for which the marginal price increases with the amount of units consumed. Observed cellular contract pricing with an allowance of minutes included with the monthly fee and a high marginal price for consumption in excess of the allowance can be seen as an simple real-world approximation to the profit-maximizing nonlinear screening contract of a extended model with many subperiods, downward sloping demand and complicated within-period consumption patterns. Such consumption patterns within the billing cycle build the foundation for a three-part tariff rationale that does not resort to irrational consumer behavior or other behavioral anomalies.

2.4 Alternative Explanations for Three-Part Tariffs

This section discusses five alternative explanations for three-part tariffs and critically contrast them against the within-period consumption explanation from the theoretical model developed in this study.

2.4.1 Standard Price Discrimination

Nonlinear pricing schemes and its welfare implications have been thoroughly researched (Wilson, 1993). Since consumers first choose their preferred pricing plan and only later learn about their exact consumption needs, any potential screening model has to account for the sequential decision process.

Such sequential screening has been analyzed by Courty and Li (2000). Ignoring the within-period consumption behavior though, such a model can explain observed cellular pricing only for demand distributions that are not supported by empirical evidence.

To illustrate this argument, consider the following simple situation: Consumer type-I has a high valuation for the good, $v = 2$, that is also highly variable, the consumer demands 0 or 2 units with equal probability. The equally likely consumer type-II has a lower valuation, $v = 1$ that is not variable, i.e. the consumer always demands one unit. The monopolist with marginal costs $c = 0.5$ finds profit-maximizing to offer the type-I consumer unlimited usage at a price equal to marginal cost for a high monthly fee, $F_I = 1.5$, while the type-II consumer pays a lower fee, $F_{II} = 1$, gets the first unit for free and pays a high marginal price ($p = 2$) for units in excess of the allowance. The high marginal price for the second unit has no impact on the type-II consumer, but makes the type-II pricing plan much less attractive to type-I consumer.

Hence, to support such pricing resembling observed tariffs in a screening model without within-period consumption patterns along the lines of Courty and Li (2000) requires that consumers who select the higher allowance $A_2 > A_1$ would be more likely to consume strictly less than A_1 minutes than would consumers who selected the tariff with the allowance A_1. As illustrated with the simple numerical example above, the type-I consumer is more likely to consume less than one unit than the type-II consumer (who consumes one unit with certainty).

The empirical evidence however does not support such demand distribu-

tion: (i) Grubb (2007) finds that the cumulative usage distribution of the three cellular plans offered to the university students is consistent with strict first-order stochastic dominance ordering, which - ignoring within-period consumption patterns - rules out allowance-type pricing plans with a high marginal price for consumption in excess of the allowance. (ii) Similarly, the sample of over 12,000 customers that chose among four plans offered by a major US cellular service provider is also consistent with first-order stochastic dominance ordering in the range of the respective allowances as shown in Figure 2.4:[23]

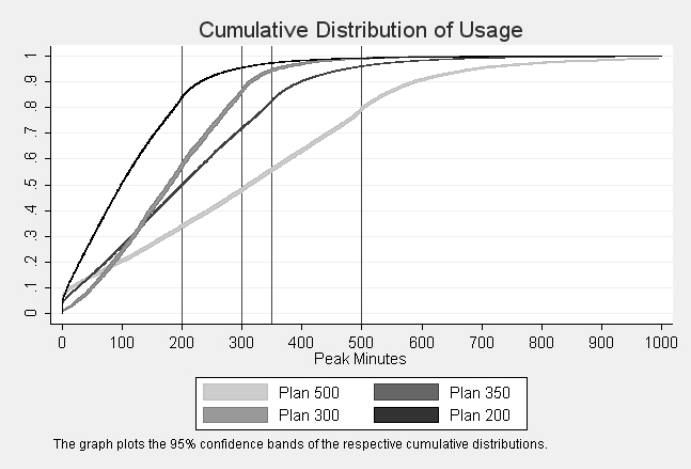

Figure 2.4: Cumulative Usage Distribution

[23]The wireless telecom sample is provided by the Teradata Center at Duke University and covers customers of a major US cellular provider from September 2001 to March 2003. The plans have allowances of 200, 300, 350 and 500 minutes included respectively, which are indicated in Figure 2.4 by vertical lines. The distributions are indistinguishable around zero and above 1,000 minutes, but are consistent with first-order stochastic dominance ordering in the range from 150 to 750 minutes (around the respective allowances). If $\widehat{F}(q)$ is the sample cumulative density distribution for N observations, the point-wise 95% confidence interval is $\widehat{F}(q) = \pm 1.96\sqrt{\frac{\left(1-\widehat{F}(q)\right)\widehat{F}(q)}{N}}$ (For large N, $\widehat{F}(q)$ is approximately normal with mean $F(q)$ and variance $\sqrt{\frac{(1-F(q))F(q)}{N}}$.

Screening models that ignore the within-period consumption patterns cannot simultaneously explain both observed cellular pricing and observed usage patterns of consumers. In contrast, the rational consumer screening model extended to account for within-period consumption patterns can explain observed three-part tariffs without imposing unrealistic consumer usage patterns. In fact, first-order stochastic dominance ordering across pricing plans with varying allowances is consistent with the screening model focusing on within-period consumption patterns.

2.4.2 Implementation of Competitive Nonlinear Pricing

In monopolistic nonlinear pricing models, the monotone hazard rate ensures that the payment (outlay) function is concave and implementable with a menu of two-part tariffs. Under duopoly competition however, Jensen (2006) shows that under certain conditions, the profit-maximizing nonlinear tariff is convex for low quantity purchases and can therefore not be implemented using a simple menu of two-part tariffs. For consumers with low demand, firms then optimally offer three-part tariffs as part of a larger menu of tariffs that still includes two-part tariffs for consumers with high demand as illustrated in Figure 2.5. The three-part tariffs which include an allowance are a moderated version of a knife-edge mechanism to implement the convex section of the outlay function.

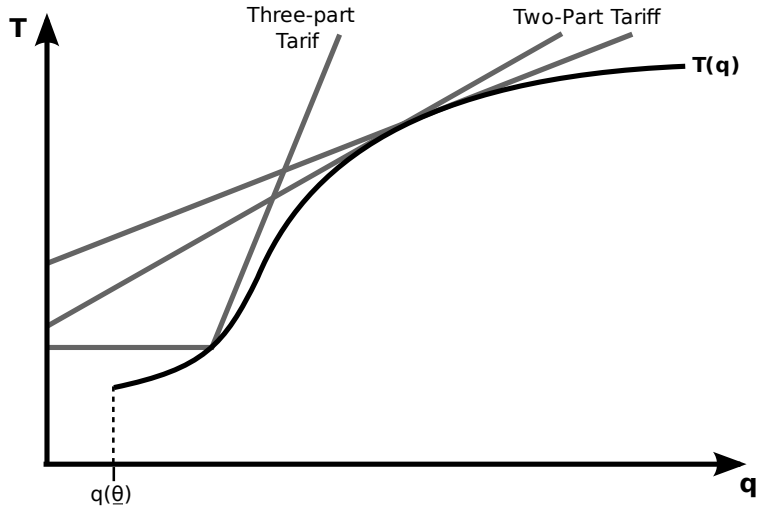

Figure 2.5: Implementation of Profit-maximizing Nonlinear Tariff under Duopoly Competition

Jensen's theoretical duopoly model provides three main predictions:[24] i) the lower envelope of the menu of three-part tariffs offered should in fact be convex, since this is the suggested prime driving force for three-part tariffs; ii) three-part tariffs are more likely to be offered for low quantity pricing plans, that is, the three-part tariff plans should be optimal (cheapest) for low quantities consumed and iii) there is no rationale for three-part tariffs to be offered by a monopolist. Empirical evidence does not seem to support these main predictions. The sample of cellular contracts offered in Canada shown in Table 2.1 results in a set of lower envelopes that contain convex and concave parts, most likely a result of firms attempting to offer a simple

[24]The condition under which the profit-maximizing nonlinear tariff with duopolistic competition is convex for small quantities depends on a rather complicated relationship among third derivatives of the consumer's utility function. The condition is not only hard to evaluate in practice, but impossible to understand intuitively. Nevertheless, Jensen (2006) does confirm that the optimal nonlinear price schedule is convex in the lower part for quadratic and logarithmic consumer utility.

set of tariffs.[25] Furthermore, most cellular service providers do offer a menu consisting of two- and three part tariffs. However, the two-part tariffs contracts (mostly pay-as-you-go) are optimal for a rather narrow range of low consumption. Hence, rather than offering three-part tariffs for the customer with low demand as the duopoly model predicts, this customer segment is typically covered by two-part tariffs in the form of pay-as-you-go pricing plans.

While the duopoly model by Jensen (2006) illustrates the scope of three-part tariffs in the implementation of the profit-maximizing nonlinear tariff, there is little plausible evidence supporting its theoretical predictions.

2.4.3 Flat-Rate Bias

The underlying assumption of standard screening models is that consumers maximize their (ex-ante) surplus which implies that they choose - at least on average over a large set of consumers - the optimal tariff. Nonetheless, numerous studies of the telecommunications industry based on transactional data indicate that consumers prefer a flat-rate tariff even though they could save money by choosing a pay-per-use tariff instead. This inability to anticipate future consumption and minimize expenditure accordingly is commonly referred to as the *flat-rate bias*.[26] Additional corroborating evidence for a

[25] The lower envelope of the total payment for all three carriers has concave and convex parts. The Bell-envelope is concave at small quantities and then convex, the Telus-envelope is convex at low quantities and then becomes concave and the Rogers-envelope is concave for small and large quantities but convex in the middle. This limited sample suggests that it is unlikely that the convexity of the profit-maximizing nonlinear price schedule at small quantities could be the (sole) rationale behind three-part tariffs.

[26] Numerous studies have investigated and found evidence of the flat rate bias in the telecommunications industry (Mitchell and Vogelsang, 1991; Kridel, Lehman, and Weisman, 1993; Kling and Van Der Ploeg, 1990; Train, McFadden, and Ben-Akiva, 1987; Train,

flat-rate bias as a behavioral phenomenon that extends past telecommunications comes from health club usage data (DellaVigna and Malmendier, 2004), online grocery stores and swimming pools (Nunes, 2000). Lambrecht and Skiera (2006) provide a comprehensive overview of the evidence of the flat-rate bias and suggest risk aversion, demand overestimation and the "taxi-meter" effect as potential causes of such a bias.

A systematic flat-rate bias would undoubtedly affect the terms under which three-part tariffs are offered, but the suggested causes of such a bias are inconsistent with three-part tariffs: Risk-averse consumers with demand variability will particularly suffer from high overage charges on their monthly bill and service providers would prefer to offer tariffs that provide consumers with less variability in their payment. Similarly, overestimation of demand cannot account for the three-part tariffs since firms could extract more surplus from consumers who systematically overestimate demand by raising the fixed fee rather than the marginal price for consumption in excess of the allowance.[27] Even if consumers suffer from the "taxi-meter" effect and derive less pleasure from consumption units for which marginal charges apply compared to units pre-paid by a fixed fee, the fact that the minute allowance in cellular contracts is limited implies that the "taxi-meter" is still running. In summary, the mere existence of a flat-rate bias cannot plausibly explain the three-part tariffs offered in the wireless telecommunications market.

Ben-Akiva, and Atherton, 1989; Lambrecht, Seim, and Skiera, 2007).

[27] A consumer who systematically overestimate her demand is excessively hurt by a high marginal price for consumption in excess of the allowance, with little resulting revenue to the firm since the consumer rarely exceeds the allowance given the overestimation of demand. The firm could increase profits from such consumers by raising the fixed fee rather than the overage price.

2.4.4 Underestimation of Future Demand

If consumers are in fact subject to behavioral anomalies, the pricing of profit-maximizing firms could then reflect such non-standard features of consumer behavior. DellaVigna and Malmendier (2004) study contract design when consumers have time-inconsistent preferences and are partially naïve about it. The profit-maximizing contract targets the consumer's misperception of future consumption. The authors consider cellular phone service a leisure good with immediate benefits and delayed costs and suggests that naïve users underestimate the number of future calls when choosing a monthly airtime package. Cellular phone companies can then extract profits from naïveté by setting high marginal prices for minutes beyond the monthly allowance, which could potentially explain the phenomenon of increasing per-minute prices within a pricing plan but decreasing per-minute prices across pricing plans.

Similarly, Gabaix and Laibson (2006) study profit-maximizing pricing of goods with add-ons (printers, hotel rooms, credit cards, etc.) when some consumers are myopic or unaware of the need to purchase the add-on. Cellular pricing plans could be interpreted as a basic good in the form of a minute allowance for a fixed monthly fee and add-ons in the form of overage minutes. When sophisticated consumers can substitute away from shrouded high overage rates through proper selection of pricing plans, the myopic consumers subsidize the low monthly fee of the included allowance by paying high overage rates.

Naïve hyperbolic discounting consumers and consumers myopic about

add-on purchases essentially underestimate their demand at the time of tariff selection. Those consumers will tend to choose pricing plans with too small of an allowance. That is, they would have been better off with a pricing plan that includes a larger allowance of minutes. If underestimation of demand is the primary rationale behind firms offering three-part tariffs, one would expect that consumers often choose cellular plans that are "too small" for their consumption profile, i.e. a situation in which they could obtain the same consumption profile at lower costs if they signed up to a plan with a larger allowance. There is no evidence of such a bias for small allowance plans, rather the evidence points to a systematic consumer bias towards three-part tariffs under which the allowance significantly exceeds the expected average usage.[28] While behavioral underestimation of future demand could rationalize three-part tariffs with high overage rates, the underlying behavioral driving force implies tariff choice patterns that are not supported by empirical evidence.

2.4.5 Overconfidence (Underestimation of the Variance of Demand)

Rather than underestimating future demand, Grubb (2007) develops a model in which consumers underestimate the *variance* of future demand because of overconfidence and shows that the model can explain observed cellular phone plans containing a minute allowance.

For a rationale of three-part tariffs based on consumer overconfidence

[28] See Train, Ben-Akiva, and Atherton (1989); Miravete (2002); Guo and Erdem (2006); Lambrecht, Seim, and Skiera (2007).

to be plausible, overage minutes and revenues must be substantial. Grubb (2007) provides empirical evidence from a sample of students of a major U.S. university to support the argument. However, a comparison of Grubb's sample with a much larger and more general sample of customers from a major US cellular provider in Table indicates that the evidence is much weaker: The high overage rate only applies to 9.62% of cellular service minutes consumed and overage revenue makes up only 6.6% of average subscriber revenue.

	Grubb (2007) Students (US University)	Teradata Customers (US Cellular Provider)
Individual Bills	18,064	195,956
Unique Customers	1,484	12,499
Bills with Overage	19%	17%
Overage as % of Monthly Fee	44%	19%
(conditional on overage occurring)	(229%)	(111.7%)
Overage as a % of Average	23%	6.6%

Table 2.4: The Importance of Overage Revenue

Furthermore, Grubb (2007) argues that the large fraction of consumers for which an alternative plan would have resulted in a lower bill for the same usage over the duration of the sample serves as evidence in favor of overconfidence (underestimation of variance). This argument however suffers from two major shortcomings: (i) Overconfident consumers who underestimate the variance of demand tend to choose pricing plans with too small of an allowance. As mentioned in Section 2.4.4, there is no evidence supporting such a bias for small allowance plans. To the contrary, a systematic overchoice of tariffs under which the allowance significantly exceeds the expected average usage is well-documented. (ii) In addition, an explanation based

on consumer overconfidence leaves no room for learning. Consumers will eventually learn about their demand variation and adjust their pricing plan accordingly. Miravete (2003) documents that cellular consumers respond even to small cost differences by switching between service plans.

Alternative explanation of observed cellular pricing, whether based on rational agents (Sections 2.4.1 and 2.4.2) or some behavioral anomalies (Section 2.4.3-2.4.5), are either inconsistent with observed pricing and demand distributions or would imply tariff choice biases that run opposite to empirical evidence. The screening model incorporating within-period consumption as presented in this study is not only compatible with the rational consumer model, but also has implications for the empirical analysis of consumption and tariff choice behavior in the telecommunications industry, as the next section will argue.

2.5 Empirical Implications of Within-Period Consumption Patterns

To date, no empirical study has investigated the within-period consumption patterns of three-part tariffs. Suppose that the underlying (true) consumption model is in fact sequential and consumers take into account the effect of current consumption decisions on the marginal price of future consumption within the billing period. Modeling the consumption decision as a one-time decision then ignores such within-period patterns and its resulting consumption inefficiencies. If the sequential nature of the consumption decision is

(economically) relevant in describing consumer behavior, empirical findings on demand elasticities and tariff choice biases will be affected, potentially biased. The potential bias influences the estimation of price elasticities of the various components of a three-part tariff structure as well as consumer preferences across a menu of three-part tariff choices. This section discusses the direction of the bias and how it can help explain empirical findings.

2.5.1 Estimation of Price Elasticities

Cellular service providers have a keen interest in understanding how different components of the pricing structure - the fixed fee, the minute allowance and the marginal price for overage consumption - impact customers' decision of plan choice, consumption of services and potential defection to a competitor.

In the standard, one-shot consumption model, the value of the marginal unit consumed is lower than any inframarginal units consumed, i.e. the consumer can perfectly prioritize calls. This severely restricts the range of the demand response to shifts in the allowance and/or the price for overage consumption. For example, a change in the overage price has zero effect on consumption in demand states in which the allowance is not exceeded. Consequently, the estimated elasticity of the overage price will be low considering the average consumers exceeds the allowance only about 20% of the time. But if consumers make the consumption decision sequentially, an increase of the overage price will not only affect demand states in which consumption exceeds the allowance: A higher overage price (price of the second unit in the simple model) increases the effective price of first-period consumption within the allowance, creating additional consumption inefficiencies and

lowering usage even in demand states that eventually remain within the allowance. Since the standard model does not capture the increase in the effective price of first-period consumption and the consumption inefficiencies resulting from it, it tends to underestimate usage elasticity and the loss of consumer surplus from an increase of the overage price.

Chapter 4 of this thesis and Lambrecht, Seim, and Skiera (2007) estimate discrete/continuous models of choice among three-part tariffs using consumer-level data on cellular phone plans and internet usage respectively and analyze the consumers' responsiveness to the different elements of a three-part tariff under demand uncertainty. Both studies find that usage is relatively inelastic to changes in the overage price. The empirical study in Chapter 4 finds an average usage elasticity of the overage price of -0.005 in the context of cellular phone plans, Lambrecht, Seim, and Skiera (2007) estimate the average usage elasticity across consumers at -0.076 for internet usage, substantially below elasticities found on two-part tariff pricing of local telephone service.[29]

These empirical findings are entirely consistent with a model of sequential consumption decisions. Particularly, the observation that changes in the fixed fee and the overage price have a differential impact on the probability a particular three-part tariff is chosen *holding constant the change in the consumer's bill* indicates that the standard model misses the increased consumption inefficiency and reduced consumer surplus resulting from an

[29] Typical estimates of usage elasticity with respect to marginal prices for local telephone service under two-part pricing range from -0.10 to -0.75 (Train, McFadden, and Ben-Akiva, 1987; Hobson and Spady, 1988; Kling and Van Der Ploeg, 1990), but can go as high as -1.70 to -2.50 (Narayanan, Chintagunta, and Miravete, 2007).

increase in the overage price.

2.5.2 Tariff Choice Bias

There exists a substantial literature on the *flat-rate bias*, a well-documented overchoice of tariffs under which the allowance significantly exceeds the expected average use.[30] Such a tendency of consumers to pick plans that are "too large" given their consumption pattern manifests itself through tariff-specific dummies in empirical tariff choice models.

The standard, one-shot consumption model does not account for the consumption inefficiencies created by the sequential consumption decisions throughout the billing period. Hence, on any three-part tariff, the consumer faces a higher effective marginal price for first-period consumption than the standard model accounts for. Consequently, such a three-part tariff is less appealing to the consumer compared to a flat-rate tariff, or a three-part tariff with larger allowance. If the within-period consumption pattern is in fact economically relevant, a standard model ignoring the sequential nature of consumption decisions will show tariff-specific preferences for flat-rate plans or three-part tariffs with larger allowances that seem non-optimal.

The empirical presence of a flat-rate bias in standard, one-shot consumption models is consistent with sequential consumption decisions over the billing period. The flat-rate bias might simply be an indication of an incorrectly specified consumer model rather than an irrational bias resulting from cognitive limitations of consumers.

[30] See Mitchell and Vogelsang (1991); Kridel, Lehman, and Weisman (1993); Kling and Van Der Ploeg (1990); Train, Ben-Akiva, and Atherton (1989); Train, McFadden, and Ben-Akiva (1987); Lambrecht, Seim, and Skiera (2007).

Whether an empirical consumer model that focuses on within-period consumption patterns specific to three-part tariffs has any success in explaining consumer behavior in the wireless telecommunications industry remains an open question. The simple model of sequential consumption decisions within the billing period provides a rational consumer explanation for the presence of three-part tariffs and can potentially help explain recent empirical findings regarding tariff choice biases and price elasticities.

2.6 Concluding Remarks

This chapter introduced a simple theoretical model focusing on the interaction of consumers' consumption patterns within the billing cycle and the firm's pricing of tariff options. The model illustrates how the effective marginal price of consumption in the first subperiod depends on the entire marginal price schedule and the distribution of demand valuations. Increasing marginal price schedules similar to commonly observed cellular contracts allow for better type separation in a menu of screening contracts. The sequential nature of consumption decisions over the billing cycle can reconcile three-part pricing within a rational consumer model while alternative explanations primarily rely on behavioral consumer anomalies and are not well-supported by evidence. The model incorporating the sequential consumption decisions could potentially provide consistent explanations of empirical findings regarding tariff choice biases and price elasticities. Eventually though, the merit of this particular rational consumer model will have to be judged by the success of its empirical application, a task for future research.

Chapter 3

Phone a Friend: Closed User Groups and Termination-based Price Discrimination

3.1 Introduction

The price a mobile customers pays for a call to a subscriber on another mobile network (off-net call) can substantially exceed the price for a call to a subscriber on the same network (on-net call). These on-net/off-net price differentials are a form of termination-based price discrimination, the price paid for a call differs depending on which telecommunication network the call terminates on. The study of on-net/off-net differentials is interesting since such differentials give rise to network externalities: A consumer benefits if additional consumers subscribe to the same mobile network because the share of calls for which the lower on-net price applies increases. These externalities caused by on-net/off-net differentials are called *tariff-mediated network externalities* since they are caused by pricing rather than technolog-

ical constraints.[31]

On-net/off-net differentials have been well-explored in the literature under uniform consumer calling patterns, that is, each consumer is equally likely to call (or receive a call) from any other subscriber on any other network. However, most mobile subscribers tend to place a large fraction of their calls to a small select group of friends and family members with which they share repeat calling relationships. Furthermore, receiving calls conveys benefits, particularly for calls between a couple, close friends or business associates. Such receiving call externalities tend to be ignored in the standard telecommunications literature. This thesis chapter presents a theoretical model that extends the standard network competition model to incorporate closed user groups such as a couple, friends, family or a small business. The study contrasts the results of the extended closed-user group model with the standard model of network competition and evaluates regulatory policy descriptions from cost-based regulation of interconnection among different networks to a ban of on-net/off-net differentials.

Due to the widespread use of on-net/off-net differentials by wireless operators, regulatory authorities have taken interest in termination-based price discrimination based on its potential collusion and foreclosure effects: First, since a large fraction of calls occur between subscribers to different networks, networks have to agree on a interconnection or call termination rate that the calling network operator has to pay the receiving network operator to terminate a call from its subscriber to a subscriber on the competitor's network.

[31] A *network externality* exists if the value to a subscriber of being connected to a specific network increases as the total number of subscribers increases.

In many countries, such termination rates are set bilaterally between network operators, in others they are set by regulatory agencies. The allegation against large wireless operators is that they employ on-net/off-net differentials in conjunction with high mobile-to-mobile termination rates to foreclose markets against smaller competitors. Essentially, the argument advanced by regulatory bodies such as the European Regulators Group (ERG) states that mobile-to-mobile termination rates above cost imply that off-net calls are more costly than on-net calls. Subscribers of the large network operator will make proportionately more on-net calls which places the smaller network at a competitive disadvantage as it is burdened by higher average calling costs. Second, since mobile network operators compete for subscribers, preliminary negotiations among networks about (reciprocal) call termination rates could potentially lead to softened price competition.[32] The (alleged) foreclosure and the collusion effect has led to numerous investigations into on-net/off-net price differentials and the regulation of termination charges around the world.[33]

During those investigations, the argument has been advanced that a substantial presence of closed user groups in mobile telecommunication markets can undermine the network carriers' incentive to keep termination charges high. Since members of closed user groups are often in close repeat call-

[32]The seminal articles in this field are Laffont, Rey, and Tirole (1998a,b).

[33]The list of countries in which investigations on on-net/off-net differentials and termination charges have been launched includes among others the UK, France, Germany, Ireland, the Netherlands, Italy, Japan and Australia. In the United Kingdom there have been thousands of pages of publicly released submissions and government studies on mobile call termination with all mobile network operators subject to price caps for call termination. In the United States, termination charges are indirectly regulated through reciprocity requirements with fixed networks (Armstrong and Wright, 2007).

ing relationships, they care about the price it costs other members of the group to place a call to them. Furthermore, closed user groups can much easier coordinate the network subscription decisions, for example in the case of a couple, a family or a small business enterprise. It is argued that the sensitivity to incoming call charges of closed user groups and the ability to take advantage of lower on-net prices by coordinating subscription decisions places a competitive constraint on the level of termination charges (Crandall and Sidak, 2004). In fact, mobile operators have made the case that the large on-net/off-net differentials arise primarily from low on-net charges by which networks try to attract new subscribers and in particular closed user groups.[34] Rather than a sign of high termination charges then, large on-net/off-net differentials could originate from networks' strategy to attract consumers in close-knit repeat calling relationships.

The main contribution of this study is the incorporation of heterogeneous closed user groups into a standard model of network competition. In communication markets in which direct consumer interaction is limited to a few people, it is more likely that the *social* network of a consumer determines subscription choice. Mobile telecommunications networks are highly compatible from a technological point of view, network effects are mainly induced by networks' pricing of on-net and off-net calls. A consumer whose friends and contacts use the same network operator does not suffer the high bills of a consumer whose friends and contacts all use different network operators. Such tariff-mediated network effects have two consequences: First, consumers may limit the frequency and length of off-net calls. Second, consumers may try

[34] See (Competition Commission, 2003, paras. 2.114, 2.124).

to coordinate the choice of network operator with their friends and peers in order to shift calls to the same network.

For the purpose of this model, the social network forming the closed user group is treated as a black box. Rather, the model introduces two features that characterize the behavior of subscribers that are part of a closed user group: First, consumers in close-knit repeat relationships benefit from receiving a call from other members of their social network. They are negatively affected if the price other group members pay to call them increases. That is, they are concerned about the cost to others within their circle of making a call to them. The strength of such externalities from receiving calls will depend on the particular social network ties and vary across consumers. Second, calling patterns of subscribers that are part of a closed user group tend to be disproportionally biased towards on-net calls - even when accounting for price differentials - due to either coordination of subscription decisions or shared idiosyncratic preferences. Incorporating the call externalities and own-network bias of closed user groups can lead to surprising insights into the rationale, benefits and welfare effects of termination-based price discrimination. The model disentangles on-net/off-net differentials from the termination markup and can thereby explain why on-net/off-net differentials are often a multiple of the termination charges. The main insights of this study are as follows:

No Termination-Based Price Discrimination: If networks are prohibited or refrain from using differential on-net/off-net pricing and charge a single usage price indepedent of terminating network, the own-network bias of closed

user groups leads networks to price usage below the average cost of calls originating on the network.[35] If the own-network bias is large, the marginal price can even be below resource cost. The marginal price will however always exceed the social-welfare maximizing price, which is below cost due to the call externalities of closed user groups. Furthermore, network profits are neither affected by call externalities nor by the own-network bias, that is, closed user groups have no effect on network profits under a uniform marginal price.

With Termination-Based Price Discrimination: The call externalities of closed user groups disentangles on-net/off-net differentials from the termination markup. The on-net markup over marginal cost is negative as networks fully internalize the call externality subscribers obtain from receiving calls from its own network. On the other hand, the off-net markup is positive since the call externalities benefit subscribers of rival networks. The on-net/off-net differential is positive even in the absence of a termination markup (cost-based access).

Asymmetric Network Competition: With an ex-ante asymmetry and competition between networks of unequal size, the difference between the marginal price of the larger and the smaller network in the absence of termination-based price discrimination depends on the sign of the termination markup. If the termination markup is positive, the smaller network has a higher usage

[35] The own-network bias of closed user groups stems from two sources: termination differentials and shared preferences. Closed-user groups exhibit an own-network bias even in the absence of termination differentials due to similar idiosyncratic preferences that leads group members to subscribe to similar network services. Birke and Swann (2006) report evidence of such an own-network bias unrelated to the existence of price differentials between on-net and off-net calls.

price. With on-net/off-net differentials, the larger network will always have a larger spread between its on-net and off-net price compared to the smaller competitor.

Welfare Effects of Termination-Based Price Discrimination: On-net/off-net differentials unambiguously reduce network profits in equilibrium. Furthermore, consumers as a group lose from termination differentials unless the elasticity of demand or the termination markup is high, although subscribers with large call externalities could benefit over a larger parameter range. Overall welfare is reduced and a ban on termination-based price discrimination improves welfare. Essentially, on-net/off-net differentials force networks into a prisoner's dilemma, in which they increase the termination differential and the pricing inefficiency, resulting in reduced welfare overall.

3.2 Mobile Telecommunications Markets

The focus of this study are mobile telecommunication markets in which only the caller pays for the call, a regulatory regime referred to as the caller-pays principle (CPP). Most anti-competitive investigations and regulatory issues have occurred in countries with CPP, a regime that is in place in the majority of OECD countries. Notable exceptions to the CPP-regime are the United States, Canada, Hong Kong and Singapore, these countries follow the receiver-pays principle (RPP) in which the calling *and* the receiving party pay for the cost of the call.

Evidence of substantial on-net/off-net price differentials can readily be found by examining calling plans offered by major network operators in

countries operating under the caller-pays principle. Cellular tariffs involving a fixed fee and a positive differential between the price of a call to a subscriber on another network and the price of a call to a subscriber on the same network are a general characteristic of cellular pricing in these markets, although exceptions certainly exist.[36] Table 3.1 illustrates average per-minute call charges for the UK:[37]

Year	Off-Net Calls	On-Net Calls	Termination Rate
2001	26.2	5.9	11.1
2005	11.3	4.2	5.9

The termination rate refers to the interconnection charge a network has to pay to have a competitor complete an off-net call.
Source: Ofcom (2006, Figure 3.38, 3.39)

Table 3.1: Average Call Charges (pence per minute)

Although the price differential has been narrowing over time, the difference between off-net and on-net charges remains prominent. The network of a subscriber making an off-net call must pay the competitor network a termination charge to complete the call on its network. High termination rates for network interconnection could be an explanation for on-net/off-net differentials as off-net calls have to bear termination charges while on-net calls do not. Interestingly, as Table 3.1 shows, off-net calls are more expensive

[36] On-net/off-net differentials are extremely common in most European mobile markets for pre-pay as well as monthly packages (Harbord and Pagnozzi, 2008). Such differentials are less prominent in RPP-countries such as Canada and the United States. Nonetheless, several cellular service providers - for example Fido and Rogers in Canada - offer cellular plans that include unlimited on-net calling (e.g. Fido-to-Fido/Rogers-to-Rogers plans). *Friends & Family* or *My Five/myFaves* plans where members enjoy better rates for calls to a small set of numbers share similar on-net/off-net differentials only if the set of numbers included in the group is restricted to be on the same network.

[37] Empirical evidence is provided for the United Kingdom because many well-documented analytical reports and investigations exist for the UK that are in the public domain.

than on-net calls *even after deducting termination charges* from the price of off-net calls, which indicates that the differential is not entirely cost-based.[38]

In addition to a substantial on-net/off-net differential, Table 3.2 illustrates the rather unbalanced calling patterns between on-net and off-net calls.

Year	**Off-Net Calls**	**On-Net Calls**
2001	67.5	32.5
2005	57.5	42.5

Source: Ofcom (2006, Figure 3.50)

Table 3.2: Shares of Types of Calls (in %)

Calling patterns are said to be balanced if the ratio of on-net to off-net calls corresponds proportionally to the market shares of the networks. In the absence of a termination differential between the price of an on-net and an off-net calls and the four roughly symmetric network operators in the United Kingdom, off-net traffic should be three times larger than on-net traffic if calling patterns were balanced. Termination-price differentials are one explanation for unbalanced calling patterns, although the termination-price differential has been narrowing at the same time as calling patterns became more tilted towards on-net calls. Another explanation for calling patterns biased towards on-net calls are closed user groups that make a large share of calls within their own group. If network subscription decisions are coordinated, either due to termination differentials or shared preferences, a substantial presence of closed user groups will tilt calling patterns towards a disproportionate share of on-net calls.[39] The theoretical model developed

[38] See also Competition Commission (2003, para 2.126 and table 5.22).

[39] See Competition Commission (2003, paras. 2.113-2.121) and Armstrong and Wright

in this paper incorporates the latter explanation, the unbalanced calling patterns arising from closed user groups, into a model of network competition to analyze the effect on on-net/off-net differentials and welfare.

3.3 Related Literature

Two different research strands are tied to the model of network competition developed in this paper. First, there is an extensive literature on network interconnection and two-way access in the telecommunication industry, anchored by the seminal work of Armstrong (1998) and Laffont, Rey, and Tirole (1998a) that has led to extensions in various directions.[40] Both papers show that network operators can use the reciprocal termination charge as an instrument of collusion. Laffont, Rey, and Tirole (1998b) and Gans and King (2001) extend the basic model to allow for termination-based price discrimination and two-part tariffs and find that the collusionary motive behind high termination charges does not extend to nonlinear pricing. Carter and Wright (1999, 2003) and Behringer (2006) analyze termination charges under asymmetric network competition with brand loyalty and Calzada and Valletti (2005) study the effect of termination rates on entry. Consumer demand heterogeneity is considered by Hahn (2004) and Dessein (2004) analyzes price discrimination with heterogeneous calling patterns of light and heavy users of telecommunication services. Calling patterns remains uniform across networks but are biased in terms of traffic direction (originating/receiving

(2007).

[40] See Laffont and Tirole (2000) and Armstrong (2002) for excellent summaries on the interconnection and access pricing literature.

calls). Armstrong (2006) integrates mobile termination into a larger model that includes fixed networks to study wholesale arbitrage and demand-side substitution.

Jeon, Laffont, and Tirole (2004) incorporate the utility from receiving a call in an analysis related to the receiver-pays regime in which the originator and the receiver of a call is charged. Call externalities have been predominantly studied in an RPP-regime where the receiving party pays for part of the call and the necessity to include some type of benefits to the receiver is obvious. Berger (2004, 2005) studies call externalities in a caller-pays regime with linear and two-part tariffs and Hoernig (2007) extends the analysis to asymmetric competition.

The second research strand studies price discrimination with social networks and personal communication ties: Shi (2003) studies monopolistic price discrimination based on social communication ties. The monopolist's social network-based discriminatory pricing strategy consists of a menu of price plans that offers discounts to subscribers with close communication networks (friends and family members) but charges higher prices to subscribers with less dense, more spread-out communications networks. The absence of competing firms however ignores the effect of social networks on interconnection and access pricing. Cherdron (2001) models calling clubs with some captive (infinite switching cost) and some perfectly mobile members and firms able to discriminate between captive and mobile members. He shows that network operators can endogenously differentiate their networks by setting high termination charges and raise profitability when consumers' calling patterns are sufficiently biased towards their peer group. Similarly,

Gabrielsen and Vagstad (2007) incorporate (identical) exogenous switching costs into a simple model of social networks in which members do not coordinate their switching behavior and find that networks then have an incentive to charge a markup on access leading to on-net/off-net differentials. Both models imply that a simple ban on termination-based price discrimination prevents collusion and restores the first-best equilibrium without any further regulatory intervention.

3.4 Closed User Groups

The most general definition of a closed user group is a group of subscribers who are concerned not only about the price of making a mobile call, but also about the price of receiving a call.[41] More detailed definitions distinguish between a narrow closed user group, in which the mobile subscriber is also the party who pays for incoming calls to its mobile, and a wide closed user group, where a group of friends and family have an interest in keeping call cost down in general. A family whose children call their parents' mobile phones and all call charges paid by the parents, or business employees calling mobiles of other employees and the company paying for both the calling party's and the called party's phone bill are examples of narrow closed user groups. A group of friends or family who do not want to impose high costs on one another or a business who has an interest in keeping rates paid by calling parties down who are clients or potential sources of business are examples of wide closed user groups.[42]

[41] Crandall and Sidak (2004).
[42] Oftel (2001, p.6).

High mobile termination charges have been a widespread concern among regulators and have lead to many investigations. In this process, it has been suggested that the presence of closed-user groups with a concern for the price of incoming calls and biased calling patterns place competitive pressure on networks to keep termination charges low.[43] Although there is evidence that mobile subscribers are less concerned about the price of incoming calls relative to the price of outgoing calls, it is doubtful that they are generally ignorant about the dampening effect of high charges for incoming calls.[44] Mobile network operators Orange and T-Mobile in their submission to the Competition Commission (UK)[45] have in fact argued that closed user groups place a constraint on the level of termination charges that network operators can impose since such closed groups can take advantage of the fact that on-net calls are priced much lower than off-net calls.[46] Since termination rates affect on-net/off-net differentials and such termination differentials have raised concerns with competition authorities, it is of interest to examine network competition with closed-user groups and their effect on termination rates and price differentials.

The formal model developed in the next section treats the social network

[43] Crandall and Sidak (2004) and Competition Commission (2003).

[44] Evidence by Oftel suggests that 13% of residential mobile subscribers take incoming call charges into account when choosing a network. Additional evidence submitted by mobile operators suggests that this share could be higher. Furthermore, small and medium business subscribers appear to be more concerned about the cost of others calling them (31%). On balance however, Oftel concluded in 2001 that termination charges substantially above cost indicate that the ability of closed user groups to constrain termination charges is limited (Oftel, 2001, pp.5-7).

[45] Competition Commission (2003, 2.114).

[46] An argument against the constraining effect of closed user groups on termination rates and charges for incoming calls is that the mobile operators can differentiate these particular groups by charging them lower prices.

that makes up a closed user groups as a black box. To reduce complexity, the internal dynamics of the social network, the ties among members and the coordination of sign-up and switching behavior are ignored. Instead assumption are placed on the behavior of consumers that reflect the fact that they are members of a closed user group.[47] In particular, consumers that are members of closed user groups are assumed to be subject to call externalities and a biased calling pattern tilted towards on-net calls:

Call Externalities: The benefits of a mobile call do not exclusively fall on the calling party. Receiving calls conveys benefits, particularly for calls between a couple, among friends, business associates or generally, individual members of closed user groups in repeat calling relationships. With call externalities falling on the receiving party, the off-net call charge of other members of the closed user group exerts an externality on the receiving party. The model incorporates heterogeneous closed user groups that differ in the magnitude of the call externalities to reflect the various types of tightly or more loosely connected social networks.

Biased Calling Patterns: The calling behavior of closed-user groups breaks down into two categories: Within-group calls and general calls that exhibit a balanced calling pattern.[48] If individual members are predominantly sub-

[47] Given the plethora of calling plans that target specific social networks (couples, friends & family, *myFaves*, business, etc.), it is an interesting avenue for future research to model the detailed workings of social networks and their implications on the pricing of mobile telecommunication services.

[48] The standard analysis of network competition assumes that calling patterns are balanced overall, that is, any two consumers receive the same number of calls from each other. Balanced calling patterns imply that the percentage of calls originating on a network and terminating on the same network is equal to the fraction of consumers subscribing to this network. This means that any given customer is equally likely to call any other user,

scribing to the same network, the resulting overall calling pattern of closed-user groups exhibits a disproportionate fraction of on-net calls. This assumption can be motivated based on two arguments: First, in the presence of termination differentials, individual members of closed user groups have an incentive to sort themselves when choosing their network and (a majority of) members of the closed user group are likely join the same network to save on calling expenditures even without explicit coordination. Network coordination among a closed-user group could be strengthened if decisions are made sequentially, but the own-network bias is present even if complete sorting fails. Second, members of a closed user group may be more likely to share similar idiosyncratic preferences and consequently prefer to subscribe to similar network services.

In essence, the relevance of network competition models that incorporate social networks and closed user groups hinges on the extent of consumer awareness and sensitivity to on-net/off-net differentials as well as on the opportunities for members of closed user groups to coordinate their network choice, thereby internalizing tariff-mediated network externalities. Birke and Swann (2006) find that the observed ratio of off-net to on-net calls is reasonably sensitive to the price premium for off-net calls. Furthermore, their estimation results suggest that even in the absence of a premium on off-net calls, there would be a disproportionately large share of on-net calls. This indicates the presence of households and closed user groups who have

regardless of the network that user is on. The assumption of balanced calling patterns is strong and often violated in reality. Estimation results by Birke and Swann (2006) suggest that overall calling patterns exhibit a disproportionately large share of on-net calls.

already coordinated their network operator choice.[49] The impact from a household member on the own operator choice is substantial and highly significant: Roughly 9.2 million network subscribers have the same impact on a household member's network choice as one additional member from the same household being on the same network.[50] In a related study, Birke and Swann (2005) find that second-year undergraduate students at the University of Nottingham Business School strongly coordinate their choice of mobile phone operators, but do so only for network operators that have on-net/off-net price differentials. Coordination on network choice is strongest within groups of students that frequently interact with each other, but weaker with students from outside their peer group.[51]

In the standard network interconnection model without call externalities or biased calling patterns (Laffont, Rey, and Tirole, 1998a,b), networks price at perceived marginal cost and consequently on-net/off-net differentials

[49] A precondition for coordinating operator choice is some knowledge about what operators other members of the closed user group subscribe to. It is not always possible to identify the operator from the telephone number alone. The availability of information on operator choice is directly linked to the closeness of two individuals in a social network, and could be obtained through direct conversation or through identifying the operator from a mobile phone (exclusive handset, logo, etc.). Birke and Swann (2005) report that in their sample, over 75% reported that they have knowledge about the operator choice of their partner and all family member and 45%/90% have knowledge about the operator choice of all/some of their friends.

[50] The total UK market of mobile subscribers during the period of their analysis is around 40 million.

[51] The authors provide evidence that this result is unlikely due to friends sharing the same (unobserved) characteristic as students did not coordinate their choice of handsets; rather they tend to choose a different handset than the one used by their friends. Coordination on network choice is strongest among Chinese students who largely use Vodafone, despite the fact that it neither offers special tariffs targeting Chinese (e.g. cheap calls to China) nor does it operate a network in the PRC. A student asked as to why Chinese students choose Vodafone replied that other Chinese students told her on her arrival that all Chinese students use Vodafone and that she should also use it in order for other people to call her.

are determined by the termination markup m. Most regulatory arguments about termination differentials henceforth centered primarily around the termination markup, large on-net/off-net differentials were synonym with high termination rates. The introduction of closed user groups with call externalities disentangles the on-net/off-net differential from the termination markup since they are driven by different pricing considerations. Hence, the benefits and drawbacks of termination-based price discrimination differ from the effects of high termination rates that have so far been the primary concern of regulatory agencies.

3.5 The Model

The network competition model is structured based on the seminal work of Laffont, Rey, and Tirole (1998a) and its companion article Laffont, Rey, and Tirole (1998b). It extends the framework by incorporating heterogeneous closed user groups with call externalities and a biased calling pattern. It is modeled based on the caller-pays principle (CPP) which is the most widespread arrangement around the world and in place in the majority of OECD countries.[52]

[52]Notable exceptions from CPP are the United States, Canada, Hong Kong and Singapore who follow the receiver-pays-principle (RPP/MPP). It has been mostly implemented due to technological reasons since mobile service providers in those countries do not have distinct access codes and a consumer could not tell whether the call is onto a fixed (landline) network or onto a mobile network. It may be considered unfair to charge a high price for a call onto a mobile network to a subscriber who is not aware what type of network she is calling. Modeling an RPP-regime is more involved since call volume is not exclusively determined by the calling party but also by receiver sovereignty: the receiver can affect the volume of calls (for which she pays) by hanging up (Jeon, Laffont, and Tirole, 2004).

Cost

The two full coverage networks have the same cost structure: There is a traffic-insensitive cost f of connecting a customer to the network for one period, which includes the cost of billing, servicing and any other cost unrelated to call volume.[53] As Figure 3.1 illustrates, a call requires the use of a switch at the originating and terminating ends of the call with marginal costs c_0 per call, and a trunk segment that connects the two switches with marginal costs c_1 per call for the transport in-between switches.[54] The total marginal cost per call is thus

$$c = 2c_0 + c_1.$$

A network operator has to pay a reciprocal unit charge a for interconnection access to the competitor's network.[55] Since the cost of completing the call for the terminating network is c_0, the termination markup related to interconnection access relative to the total cost of a call is

$$m \equiv \frac{a - c_0}{c}.$$

[53] In the case of wireless telecommunications, the fixed costs f may also include handset subsidies and might therefore be relatively large.

[54] On average, on-net calls will have a slightly lower expected cost for two reasons: (i) some on-net calls occur between subscribers on the same switch and require no trunking and (ii) on-net calls do not require communication between networks and any duplication or inefficiencies in routing can be avoided. Furthermore, incoming call costs tend to be higher than outgoing call costs since the originating network has no knowledge of the location of the called party and hands over the call to the terminating network at the most convenient interconnection point for the originating network. The terminating network then has to locate the called party to terminate the call (Competition Commission, 2003, 5.127). For simplicity and to maintain consistency with the previous literature, such minor cost differences are omitted from the model.

[55] The access charge a is assumed to be non-negative. With negative access charges, a network could obtain an infinite amount of money by installing a computer that calls customers of the other network. Hence, the termination markup is restricted to $m \geq -\frac{c_0}{c}$.

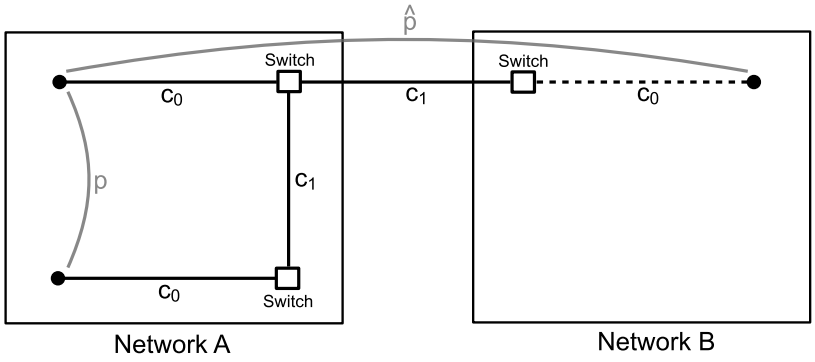

Figure 3.1: Network Interconnection

On-net and off-net calls incur the same resource cost if call termination is priced at cost, $a = c_0$, or equivalently by definition $m = 0$. [56] Network competition is initially solved for any arbitrary reciprocal access charge determined by regulators. Section 3.7 extends the analysis to include access negotiations between interconnecting firms in advance of the price setting game. It is important to note that the termination markup refers to the markup of the interconnection charge over the real resource cost of interconnection (which for the most part of the paper is exogenous) and not to a possible markup of the off-net price $\widehat{p}$ charged by the network, which is endogenously determined under network competition with termination-based price discrimination.

Demand

The two networks are horizontally differentiated, the consumers are uniformly located on the segment $[0, 1]$ with density 1 and the networks are

[56] This is consistent with findings of the UK Competition Commission (Competition Commission, 2003, para. 2.126).

located at the endpoints, $x_1 = 0$ and $x_2 = 1$. Consumers face a discrete-choice problem of deciding which (single) network they will subscribe to.[57] Consumer s located at x_s joining network i has utility

$$y + u(q) + \overline{u(q)} + v_0 - \tau |x_s - x_i|$$

where y is the exogenously given income and q is the consumption of telecommunication services. There exists a fixed surplus v_0 from being connected to the network, which is assumed to be "large enough" so that all consumers are connected in equilibrium (fully covered market).[58] The cost of not being connected to their most preferred network is $\tau |x_s - x_i|$, where $\sigma = \frac{1}{2\tau}$ is the measure of substitutability between network. When a mobile subscriber is called, she also receives (passive) utility that is proportionate the caller's utility, $\overline{u(q)} = \beta_s u(q)$. The individual parameter β_s is a measure of the extent to which a consumer is part of a closed user group and reflect the strength of social network ties. It represents the fact that calls are usually part of an underlying relationship, personal or economic, between the calling parties. β_s is uniformly distributed on $[0, 1]$ and independent of the consumer's network preference x_s. If $\beta_s = 0$, as in the standard model of network competition, the consumer does not get any utility from being called and hence is not concerned about the price other users pay to call her. Higher values of β_s

[57] In practice, consumers often belong to a single network either because it minimizes transaction costs or because networks charge two-part tariffs for price discrimination purposes reflecting the connection-, billing-, and servicing costs.

[58] A fully covered market implies that the market shares α_i and α_j sum to one. The term $v_0 - \tau |x_s - x_i|$ can be thought of as the net utility of having access to basic mobile services such as being able to make (or receive) a call from family members or doctor's in an emergency. Such phone services provide high utilities relative to cost and are demanded inelastically compared to regular consumption represented by the utility functions $u(q)$ and $\overline{u(q)}$.

represents subscribers who receive utility from being called and consequently are concerned about the price others pay to call. Networks are assumed to be unable to price discriminate among subscribers based on the strength of the externality β_s.[59]

The functional form of the variable surplus is

$$u(q) = \frac{q^{1-\frac{1}{\eta}}}{1-\frac{1}{\eta}} \qquad \eta > 1$$

which yields a constant elasticity demand function[60]

$$u'(q) = p \qquad \Longleftrightarrow \qquad q = p^{-\eta} \quad \text{and} \quad u(p) = \frac{\eta}{\eta - 1} p^{-(\eta-1)}$$

where p is the uniform price and the consumer's variable net surplus is

$$v(p) = \max_q \{u(q) - pq\} = \frac{p^{-(\eta-1)}}{\eta - 1}.$$

Individuals subscribing to network i make a fraction of α_i^γ on-net calls and $1-\alpha_i^\gamma$ off-net calls with $\gamma \in (0, 1]$. The parameter γ reflects the own-network bias of closed user groups, with $\gamma = 1$ resulting in a balanced calling pattern.

[59] In the chosen specification, the call utility function of the call externality is proportional to the subscriber's utility function for making outbound calls and features diminishing marginal utility. Without closed user group heterogeneity, Jeon, Laffont, and Tirole (2004) and Berger (2005) show that as the benefit of receiving calls tends towards the benefit of making calls ($\beta \to 1$), the equilibrium off-net price becomes arbitrarily high and leads to a "connectivity breakdown". Heterogeneity with respect to the call externality in this specification avoids a connectivity breakdown in the symmetric equilibrium. Alternatively, Armstrong and Wright (2007) employ a linear specification based on the argument that subscribers generally do not control which calls to them are made and each call should be taken as a random draw from the recipient's willingness-to-pay for incoming calls. The linear specification avoids the connectivity breakdown since the harm to a network's own subscribers eventually dominates the harm done to a rival's subscribers.

[60] The assumption of constant elasticity is made for technical convenience.

3.5.1 No On-Net/Off-Net Price Discrimination

Suppose first that networks are unable to employ termination-based price discrimination. Networks then compete by offering a single two-part tariff each:

$$T_i(q) = F_i + p_i q(p_i) \qquad i = 1, 2$$

where the fixed fee F_i is a subscriber charge and p_i is the marginal price or usage fee for a call (independent of the terminating network). After subtracting the fixed fee F_i, the net surplus offered to network i's consumers becomes

$$w_i = \alpha_i^{\gamma} [v(p_i) + \overline{u}(q(p_i))] + (1 - \alpha_i^{\gamma}) [v(p_i) + \overline{u}(q(p_j))] - F_i.$$

With uniform usage fees independent of the terminating network, firms are unable to target closed user groups. Since $\overline{u}(q(p_i)) = \beta u(q(p_i))$, the consumer located at $x(\beta)$ that is indifferent between subscribing to network i and j is given by

$$x(\beta) = \frac{1}{2} + \sigma [v(p_i) - F_i - (v(p_j) - F_j)] - \left(1 - \alpha_i^{\gamma} - \alpha_j^{\gamma}\right) \beta [u(q(p_i)) - u(q(p_j))]$$

and market share of network i is given by

$$\alpha_i = \int_0^1 x(\beta)\, d\beta,$$

or

$$\alpha_i = \frac{1}{2} + \sigma [v(p_i) - F_i - (v(p_j) - F_j)] - \frac{1}{2} \left(1 - \alpha_i^{\gamma} - \alpha_j^{\gamma}\right) [u(q(p_i)) - u(q(p_j))].$$

(3.1)

Given the fraction of on-net calls α_i^γ, network profits are

$$\pi_i = \alpha_i (F_i - f) + \alpha_i (p_i - c) q(p_i) + \alpha_i (1 - \alpha_i^\gamma) mc (q(p_j) - q(p_i)).$$

Network competition in the presence of closed user groups but without termination-based price discrimination is characterized in the next proposition:

Proposition 3.1 (No Termination-Based Price Discrimination)
If the degree of substitutability between networks (σ) is not too high, then there exists a unique and symmetric equilibrium in nonlinear tariffs that is characterized by the following properties:

(i) *The markup of the marginal price over perceived marginal cost (Lerner index) is*

$$L = \frac{p - \left(1 + \left(1 - \left(\frac{1}{2}\right)^\gamma\right) m\right) c}{p} = -\kappa(\gamma) \leq 0$$

where $\kappa(\gamma) = \left(\frac{1}{2}\right)^\gamma - \frac{1}{2}$ is the own-network bias in the symmetric equilibrium and $L_i^* = 0$ if and only if $\gamma = 1$.

(ii) *The fixed subscription fee in the symmetric equilibrium, F^*, is equal to the standard Hotelling mark-up $\frac{1}{2\sigma}$ plus the connection cost f minus the net marginal revenue if all calls were on-net $(p - c) q(p)$, or*

$$F = \frac{1}{2\sigma} + f - (p - c) q(p).$$

(iii) *The termination markup m and the network bias γ do not affect the*

symmetric equilibrium profit, which is equal to the profit that networks would obtain under unit demands:[61]

$$\pi = \frac{1}{4\sigma}$$

Proof. *See Appendix B.*

In standard network competition models, two-part tariffs lead to marginal cost pricing, although the relevant cost is not the resource cost c, but rather the perceived marginal cost faced by individual network i. In this model, the perceived marginal cost depend on the termination markup m and the network bias γ and equal $\left(1 + \left(1 - \left(\frac{1}{2}\right)^{\gamma}\right) m\right) c$ in the symmetric equilibrium. In the absence of the own-network bias, networks will price at perceived marginal cost even if call externalities are present. The own-network bias inherent in closed user groups though pushes the usage price below perceived marginal cost. The marginal price p charged could even fall below resource cost c, which happens if the termination markup m is small and the own-network bias strong (γ close to zero or $\kappa(\gamma)$ close to $\frac{1}{2}$), more specifically if

$$m < \frac{\kappa(\gamma)}{\frac{1}{2} - \kappa(\gamma)} = \frac{\left(\frac{1}{2}\right)^{\gamma} - \frac{1}{2}}{1 - \left(\frac{1}{2}\right)^{\gamma}}.$$

The fixed fee F charged by the networks in the symmetric equilibrium reflects the markup $\frac{1}{2\sigma}$ from standard (unit-demand) Hotelling competition, the traffic-insensitive cost of connecting a subscriber to the network f and an

[61] Laffont, Rey, and Tirole (1998a) show that when two-part tariffs are available, network competition with multi-unit demand closely resembles competition with unit demand, where equilibrium prices are equal to $c + \tau$ and hence $\pi = \frac{\tau}{2} = \frac{1}{4\sigma}$ (for this computation, see e.g. Tirole (1988)).

adjustment term. The adjustment term is equal to the net marginal revenue if all calls were on-net calls. A decrease of the equilibrium marginal price due to a stronger own-network bias (higher $\kappa(\gamma)$) or an increased termination markup m will lead to an increased fixed fee. In fact, the increase in the fixed fee just compensates the decreased marginal price such that network profits in equilibrium are unaffected by either the own-network bias or the termination markup, they exactly equal network profits without closed user groups.[62] Suppose that the termination markup m is raised. This has a direct positive effect on network profits as it generates additional access revenues. However, competition forces networks to adjust the marginal price and the fixed subscription fee F, since they must compensate subscribers for the loss in net utility from making off-net calls. Overall the two effects exactly cancel.[63]

[62] See Laffont, Rey, and Tirole (1998a).

[63] The result that profits are independent from the termination markup is similar to Laffont, Rey, and Tirole (1998a), even though networks do not price at perceived marginal cost unless there is no own-network bias. To gain some intuition with respect to the independence of the termination charge, suppose the termination markup m is raised by dm. This increases each network's marginal cost by $\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)cdm$, the marginal price p goes up by

$$\frac{\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)c}{\frac{1}{2}+\left(\frac{1}{2}\right)^{\gamma}}dm$$

and marginal revenue is increased by

$$q\left(\frac{\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)}{\frac{1}{2}+\left(\frac{1}{2}\right)^{\gamma}}-\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)\right)cdm.$$

In order to keep consumers' net surplus (market shares) constant, a network must reduce the fixed fee by

$$dF=q\frac{\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)c}{\frac{1}{2}+\left(\frac{1}{2}\right)^{\gamma}}dm.$$

The net effect of the lowered fixed-fee gain coupled with increased marginal revenue from attracting a new consumer is $-q\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)cdm$. On the other hand, the increase in the

The effect of closed user groups in the absence of termination-based discrimination is to push marginal prices below perceived marginal cost as a result of the own-network bias. If the own-network bias is large, this can result in pricing below marginal resource cost c. Networks compensate lower marginal prices with increased fixed fees and overall profits are unaffected by the own-network bias or the termination markup. Not surprisingly, with an inability to specifically target closed user groups with termination differentials, the results of Proposition 3.1 mirror standard network competition analyzed by Laffont, Rey, and Tirole (1998a), apart from the trade-off between lower marginal prices and increased fixed fees.

3.5.2 Termination-Based Price Discrimination

With termination-based price discrimination, the call externality coupled with the own-network bias hands networks more strategic pricing flexibility and alters the competitive environment: A network's ability to raise off-net prices imposes a negative externality on subscribers to its competitor's network and large price differentials between on-net and off-net calls might be appealing to subscribers exhibiting large call externalities given the own-network bias.

Consumers are assumed to be aware of the identity of the mobile network to which the call is being made.[64] Networks compete by offering a single two-

termination markup provides an additional incentive to attract a customer, as this saves an extra amount in access payment to the competing network equal to $q\left(1-\left(\frac{1}{2}\right)^{\gamma}\right)cdm$ in the symmetric equilibrium. The two effects cancel, and the intensity of competition does not vary with the termination markup.

[64] Since the advent of mobile number portability (MNP), the number prefix does not automatically indicate the network assignment of a phone number and subscribers may suffer from a customer ignorance problem about outbound calling charges if they are unable to

part tariff each: [65]

$$T_i(q) = F_i + p_i q(p_i) + \widehat{p}_i q(\widehat{p}_i) \qquad i = 1, 2$$

where p_i, $q(p_i)$ and $\widehat{p}_i$, $q(\widehat{p}_i)$ are the marginal prices and quantities of on- and off-net calls and F_i is the fixed fee. The consumer $x(\beta)$ that is indifferent between subscribing to network i and j is given by

$$\begin{aligned} x(\beta) =& \frac{1}{2} - \sigma(F_i - F_j) + \sigma\alpha_i^{\gamma}\left[v(p_i) + \beta u(q(p_i))\right] + \sigma(1-\alpha_i^{\gamma})\, v(\widehat{p}_i) \\ & - \sigma\left(1-\alpha_j^{\gamma}\right)\beta u(q(\widehat{p}_i)) - \sigma\alpha_j^{\gamma}\left[v(p_j) + \beta u(q(p_j))\right] \\ & - \sigma\left(1-\alpha_j^{\gamma}\right) v(\widehat{p}_j) + \sigma(1-\alpha_i^{\gamma})\,\beta u(q(\widehat{p}_j)) \end{aligned}$$

Market share for network i is then

$$\begin{aligned} \alpha_i = \int_0^1 x(\beta)\, d\beta =& \frac{1}{2} - \sigma(F_i - F_j) + \sigma\alpha_i^{\gamma}\left[v(p_i) + \frac{1}{2}u(q(p_i))\right] \\ & + \sigma(1-\alpha_i^{\gamma})\, v(\widehat{p}_i) - \frac{1}{2}\sigma\left(1-\alpha_j^{\gamma}\right) u(q(\widehat{p}_i)) \\ & - \sigma\alpha_j^{\gamma}\left[v(p_j) + \frac{1}{2}u(q(p_j))\right] - \sigma\left(1-\alpha_j^{\gamma}\right) v(\widehat{p}_j) \\ & + \frac{1}{2}\sigma(1-\alpha_i^{\gamma})\, u(q(\widehat{p}_j)) \end{aligned} \tag{3.2}$$

identify the network assignment of a particular number. Gans and King (2000) and Wright (2002) show that mobile operators have an incentive to increase termination charges if customers are only aware of *average* prices. Various countries have mandated measures to overcome the loss in tariff transparency through acoustic signals or verbal announcements for off-net calls or a toll-free inquiry number to identify a particular number's network assignment (Buehler, Dewenter, and Haucap, 2006). Although number prefixes never identified mobile carriers in receiving-party-pay regimes (e.g. United States, Canada), the consumer ignorance problem does not vanish as long as on-net/off-net differentials exist for the calling party.

[65] This assumes that due to incomplete information on consumers' tastes, the networks lack the ability to price discriminate according to a subscriber's strength of the call externality β_s.

and network profits are

$$\begin{aligned}\pi_i =&\alpha_i \left(F_i - f\right) + \alpha_i^{1+\gamma} \left(p_i - c\right) q\left(p_i\right) \\ &+ \alpha_i \left(1 - \alpha_i^{\gamma}\right) \left(\left(\widehat{p}_i - c\right) q\left(\widehat{p}_i\right) + mc \left(q\left(\widehat{p}_j\right) - q\left(\widehat{p}_i\right)\right)\right).\end{aligned}$$

Competition with two-part tariffs and termination-based price discrimination is characterized in the next proposition:

Proposition 3.2 (On-Net/Off-Net Differentials)

As long as the degree of substitutability between networks (σ) is not too high, there exists a unique symmetric equilibrium in nonlinear tariffs with the following properties:

(i) *The optimal marginal prices exhibit a negative markup (Lerner index) for on-net calls, $\frac{p-c}{p} = -\frac{1}{2}$, and a positive markup (over perceived marginal cost) for off-net calls, $\frac{\widehat{p}-(1+m)c}{\widehat{p}} = \frac{1}{2}$.*

(ii) *The on-net/off-net differential in the symmetric equilibrium is*

$$\delta = \widehat{p} - p = \left(\frac{4}{3} + 2m\right) c = 2\left(p + mc\right) > 0.$$

There exists a positive on-net/off-net differential even if the access markup is zero, $d_{m=0} = 2p > 0$.

(iii) *The fixed subscription fee in the symmetric equilibrium is*

$$\begin{aligned}F =&\frac{1}{2\sigma} + f - \gamma \left(\frac{1}{2}\right)^{1+\gamma} \left[w\left(p\right) - w\left(\widehat{p}\right)\right] - \left(1+\gamma\right) \left(\frac{1}{2}\right)^{\gamma} \left(p - c\right) q\left(p\right) \\ &- \left(1 - \left(1+\gamma\right) \left(\frac{1}{2}\right)^{\gamma}\right) \left[\left(\widehat{p} - c\right) q\left(\widehat{p}\right)\right]\end{aligned}$$

where $w(p) = v(p) + \frac{1}{2}u(q(p))$ *and* $w(\widehat{p}) = v(\widehat{p}) + \frac{1}{2}u(q(\widehat{p}))$.

(iv) *Equilibrium profits are equal to*

$$\pi = \frac{1}{4\sigma} - \gamma\left(\frac{1}{2}\right)^{1+\gamma}\left(\frac{1}{2}\left[w(p) - w(\widehat{p})\right] + (p-c)\,q(p) - (\widehat{p}-c)\,q(\widehat{p})\right)$$

and are decreasing in the network bias (increasing in γ*).*

Proof. *See Appendix B.*

While the markup of the usage price in the absence of termination-based price discrimination was primarily determined by the own-network bias, biased calling patterns have little influence on usage prices if networks use termination differentials. Rather, the usage prices and the resulting on-net/off-net differential are driven by the call externalities. With termination-based price discrimination, networks fully internalize the call externality of members of closed user groups for on-net calls but not for off-net calls. The on-net price is adjusted downwards below resource cost to reflect the call externality subscribers enjoy from being called by other subscribers on the same network. The on-net price is unaffected by the termination markup and socially efficient, that is, it equals the welfare-maximizing price $\widetilde{p}$.[66] If a network lowers its on-net price, only its own customers benefit directly, or indirectly through the call externality. For off-net calls however, call externalities give an indirect benefit to customers of the rival network who receive

[66] The socially optimal marginal price $\widetilde{p}$ maximizes

$$\int_0^1 (1+\beta)\,u(q(\tilde{p})) - cq(\tilde{p})\,d\beta$$

and is equal to $\widetilde{p} = \frac{2}{3}c$.

cross-network calls. The off-net price is adjusted upwards because fewer calls to subscribers of the rival network reduces surplus to its subscribers and benefits the original network. Hence, networks find it optimal to have a positive on-net/off-net differential even in the absence of a termination markup.

The usage price for on-net calls is below the marginal cost ($p < c$), and the usage price for off-net calls is above perceived marginal cost ($\widehat{p} > (1+m)\,c$); on-net calls are cheaper and off-net calls more expensive compared to the equilibrium without termination-based price discrimination. The average usage price with termination-based price discrimination is $\bar{p} = \left(\frac{1}{2}\right)^{\gamma} \frac{2}{3} c + \left(1 - \left(\frac{1}{2}\right)^{\gamma}\right) 2\,(1+m)\,c$ and exceeds the non-discriminatory price in the absence of termination-based price discrimination. Although marginal prices are unaffected by the strength of the own-network bias (γ), the fixed fee (and network profits) are decreasing in the own-network bias (increasing in γ) since average usage prices paid depend on the share of on-net calls.[67]

The equilibrium on-net/off-net differential is $\delta = 2\,(p + mc)$, increasing in the termination markup m. The own-network bias does not affect the equilibrium differential, call externalities affect the differential indirectly through the on-net price. With call externalities, the on-net/off-net differential is always positive for any feasible termination markup ($m > -\frac{c_0}{c}$) which implies that termination differentials exists even if access is priced at cost ($m = 0$). The call externality present in closed user group separates the on-

[67] Contrary to Jeon, Laffont, and Tirole (2004) and Berger (2005), this model does not result in a connectivity breakdown in the symmetric equilibrium for any values of the network bias γ. This is due to subscriber heterogeneity with respect to the call externality and the networks inability to screen subscribers. However, connectivity breakdowns could arise with asymmetric networks as the larger network l will charge prohibitively high off-net prices $\hat{p}$. Indeed, as $\alpha_l \to \frac{2}{3}$, the off-net price $\hat{p}_l$ goes to $+\infty$.

net/off-net differential from being directly tied to the termination charges. Observed termination differentials are not solely due to above-cost termination charges, a result that can reconcile the empirical observation that termination differentials often exceed cost differences between on-net and an off-net calls.

3.6 Competition Between Asymmetric Networks

Smaller mobile competitors have raised concerns that on-net/off-net price differentials place them at a disadvantage relative to larger competitors, or even that larger operators strategically use on-net/off-net price differentials to induce the exit of a smaller competitor. In the absence of call externalities and biased calling patterns, the termination differentials are directly tied to the termination charge (m) between network.[68] The large on-net/off-net price differentials are merely a manifestation of the underlying termination markup.[69]

Closed user groups however give rise to on-net/off-net differentials even in the absence of a termination markup ($m = 0$) based on entirely different considerations. This section introduces a (small) asymmetry and presents comparative statics results of usage prices, and on-net/off-net differentials with respect to the level of asymmetry in market share with and without termination-based price discrimination, ignoring potential exclusionary mo-

[68] See Laffont, Rey, and Tirole (1998b).

[69] Stennek and Tangerås (2006) argue for a ban on termination-based price discrimination in mobile telecommunications in order to negate the networks' incentives for high termination rates. The particularly appealing feature of such a ban is that it requires neither cost nor demand information.

tives of high termination markups.

To account for asymmetric markets, subscribers are assumed to receive an additional utility of $\frac{A}{\sigma}$ if they join network 1, where A measures the ex-ante asymmetry in market share before equilibrium effects.[70] The extra utility can be thought of as an incumbency or reputational advantage of network 1 and serves to make equilibrium market shares asymmetric, with $\alpha_1 > \alpha_2$.[71]

The equilibrium under asymmetric competition cannot be solved analytically. Rather, Proposition 3.3 presents comparative statics results from introducing a small ex-ante asymmetry A to the symmetric Nash equilibrium:

Proposition 3.3 (Comparative Statics of Ex-ante Asymmetry A) *Starting from the symmetric equilibrium, the following comparative statics results hold with respect to a small ex-ante asymmetry A:*

No Termination-Based Price Discrimination: The marginal price p of the large (small) network decreases (increases) with the ex-ante asymmetry parameter A if the termination markup m is positive. If $m < 0$, the results are reversed. The (absolute) difference in marginal prices between the large and the small network is increasing in cost c, the termination markup m and the network bias (decreasing in γ).

With Termination-Based Price Discrimination: Networks charge the same on-net price (p) independent of size, but the larger network has a larger

[70] This modeling approach follows Carter and Wright (1999, 2003) with the extra benefit additive to the subscriber's location. This implies that all subscribers to that network are equally affected. An alternative approach employed by Behringer (2006) incorporates an asymmetric multiplicative utility amplifier which confers larger benefits to subscribers who are closer to their preferred brand.

[71] The market is still assumed to be fully covered (v_0 is "large enough").

markup on its off-net price ($\widehat{p}$) compared to the smaller network. Hence, the on-net/off-net differential δ is larger for the bigger network. The difference between the networks' on-net/off-net differentials ($\delta_1 - \delta_2$) is increasing in cost c, the termination markup m and the network bias (decreasing in γ).

Proof. *See Appendix B.*

In the absence of termination-based price discrimination, the smaller network will charge a higher usage price p whenever the termination markup m is positive but will undercut the larger network if $m < 0$. While the markup over "perceived" cost is identical for both networks, price differences arise since the smaller network has higher perceived marginal cost if $m > 0$ and lower perceived marginal cost if $m < 0$ (a larger fraction of its call are off-net). The own-network bias γ amplifies the relative share of on-net calls of the larger network for a given ex-ante market share asymmetry A (relative to the smaller network) and the absolute difference in marginal price is increased independent of the value of m.

With termination-based price discrimination on the other hand, the on-net price p is unaffected by the ex-ante market share asymmetry A (although the share of on-net calls will differ). Network internalize the full externality arising from receiving on-net calls. The off-net price is partly driven by the call externality arising to subscribers of the competitor's network. Subscribers of the large network benefit less from a low off-net price of the small network relative to the benefit subscribers of the small network receive from a low off-net price of the large network. This is entirely due to the lower share of incoming off-net calls. Consequently, the larger network optimally

chooses a higher off-net price than the smaller network. Given the on-net price is the same for both networks, this translates directly into a larger on-net/off-net differential for the larger network. Similar to the situation without termination-based price discrimination, the network bias amplifies differences in the relative share of incoming off-net calls, which increases the difference between the networks' on-net/off-net differentials $\delta_1 - \delta_2$.

3.7 Negotiations on Access Pricing

Up to this point, the termination charge has been assumed to be exogenously determined before the price competition stage. This characterization reflects the regulatory situation in the European Union where termination charges are set by an industry regulator. However, in other jurisdictions, termination charges are negotiated among competitors in a preliminary stage with or without further restrictions such as reciprocity imposed by law.[72] This section analyzes networks (cooperatively) negotiating a reciprocal termination markup m that maximizes their joint profits and the implications of such access negotiation for on-net/off-net price differentials are considered.

The reciprocal termination markup negotiated between networks could be negative ($m < 0$). The markup is restricted though by technology to be above $m > \overline{m} = -\frac{c_0}{c}$, since a termination markup below $\overline{m}$ would imply that networks could obtain an infinite amount of money by installing a com-

[72] Reciprocity is imposed by law in the United States, but is not imposed for access negotiations among networks in Korea, Japan or New Zealand. Imposing reciprocity can eliminate the problem of double marginalization that appears under non-reciprocal access pricing (Laffont, Rey, and Tirole, 1998a).

puter that calls customers of the other network.[73] A termination markup $\overline{m}$ implements a *bill-and-keep* interconnection arrangement in which no settlement payments are exchanged between networks for terminating calls from the other network. Such a bill-and-keep arrangement allows the networks to implement the lowest feasible termination markup $\overline{m}$ and could potentially save on accounting and other transaction costs.

Laffont, Rey, and Tirole (1998a,b) and Gans and King (2001) analyze reciprocal access negotiations in the absence of call externalities and the own-network bias of closed user groups with the following conclusions:

(i) In the absence of price differentials, network profits are independent of the termination markup.[74]

(ii) With termination-based price differentials, the joint-profit maximizing termination charge is negative, $m < 0$. Intuitively, when m is negative, off-net calls will be priced *below* on-net calls since the absence of a network bias or call externalities implies that networks will employ usage prices reflecting perceived marginal cost. Hence, subscribers benefit from belonging to the smaller network as a larger share of the calls will be cheaper off-net calls. Lowering the fixed fee F to attract marginal

[73] The total cost of a call are $c = 2c_0 + c_1$, where c_0 is the cost from the switch to the caller and c_1 is the cost of using the trunk to connect from switch to switch. Hence, the lowest potentially feasible termination markup is $\overline{m} = -\frac{1}{2}$ and occurs when trunk cost c_1 are zero.

[74] Laffont, Rey, and Tirole (1998a) show that when the access charge is too large, no equilibrium exists in pure strategies and attention is therefore restricted to reciprocal access charges for which an equilibrium in pure strategies does exist. While network profits are independent of the termination markup, cooperation (reciprocity) is still required among the networks: Although they are indifferent between all symmetric access charges, they are not indifferent with respect to unilateral increases in their own access charges. A small unilateral increase in one network's termination markup does increase that network's equilibrium profit.

subscribers then becomes a less attractive option and networks become less interested in building market share. Price competition is softened and networks obtain higher profits.[75]

Proposition 3.4 characterizes the optimal termination markup in the presence of closed-user groups exhibiting call externalities and biased calling patterns:

Proposition 3.4 (Negotiations of Termination Markup)

No Termination-Based Price Discrimination:

(i) *Network profits are independent of the termination markup m.*

(ii) *Social welfare is maximized if $\widetilde{m} = -\frac{1+\kappa(\gamma)}{3(1-\kappa(\gamma))}$. The social welfare-maximizing termination markup is always negative, the first-best outcome is only feasible though if $\widetilde{m} > \overline{m} = -\frac{c_0}{c}$.*

With Termination-Based Price Discrimination:

(iii) *The profit-maximizing termination-markup is always negative and equal to*

$$m^* = \max\left(-\frac{3\eta-2}{5\eta-2}, \overline{m}\right).$$

If $m^* = \overline{m}$, joint-network profits are maximized with a bill-and-keep system.

(iv) *The first-best outcome can never be achieved under termination-based price discrimination.*

Proof. *See Appendix B.*

[75] Calzada and Valletti (2005) generalize these findings to a multi-firm industry with logit demand.

Proposition 3.4(i) simply restates the result from Proposition 3.1 that in the absence of termination-based price discrimination, network profits are independent of the termination markup. It seems reasonable to argue then that if networks negotiate a reciprocal termination charge in the preliminary stage, they will be able to agree on a non-positive termination markup m resulting from either cost-based access ($m = 0$), a bill-and-keep system ($\overline{m}$) or from the social welfare maximizing termination charge ($\widetilde{m}$ if feasible).[76]

In the absence of closed-user groups and call externalities, networks price at perceived marginal cost. Consequently, on-net/off-net differentials are directly determined by the termination markup m. Introducing call externalities leads to two opposite pricing forces for off-net calls: (i) The inability to internalize call externalities for off-net calls leads networks to substantially markup off-net calls. This results in enlarged termination differentials and networks compete more aggressively for market share as subscribers have an incentive to belong to the larger network and it becomes more attractive for networks to lower the fixed fee F and attract marginal consumers. (ii) On the other hand, the termination markup affects the perceived marginal cost of off-net calls and through a negative termination markup, networks can reduce termination differentials and soften competition for market share.[77] Hence, the profit-maximizing termination charge is always negative. The profit-maximizing markup m is constrained if trunk cost c_1 or demand elas-

[76] While both networks charge a uniform price independent of terminating network, if there exists an ex-ante asymmetry A, a non-positive termination markup would imply that the smaller network has a (weakly) lower usage price.

[77] Competition laws prevent networks from fixing off-net prices. Interestingly though, such an anti-competitive agreement to fix the off-net price would involve *not* charging more than an agreed price for off-net calls and might look suspiciously pro-competitive.

ticity are high and a bill-and-keep system is then optimal from the networks' perspective.

Networks find it optimal to settle on a negative termination charge since large termination differentials give rise to strong (positive) network effects. Markets with positive network effects are characterized by fierce competition and low profits.[78] A negative termination markup overturns the price differential in the absence of closed user groups, and narrows the differential in the presence of closed-user groups, thereby leading to relaxed competition between networks and increased profits.

Achieving the first-best outcome requires that the termination markup m to be such that it results in the welfare-maximizing usage price $\widetilde{p} = \frac{2}{3}$. Without price-differentials, network set the marginal price at

$$p^* = \frac{\left(1 + \left(\frac{1}{2} - \kappa(\gamma)\right) m\right) c}{1 + \kappa(\gamma)}$$

and the first best outcome can be achieved if $\frac{c_0}{c}$ is large and the extra mass of on-net calls $\kappa(\gamma)$ arising from the own-network bias of closed user groups is small in the symmetric equilibrium ($\kappa(\gamma) = \left(\frac{1}{2}\right)^{\gamma} - \frac{1}{2}$), i.e. γ close to 1. If networks employ termination differentials, they fully internalize the call externality for on-net calls, which implies that the price of on-net calls is always at its first-best level. On the other hand, networks heavily markup their price for off-net calls, in the symmetric equilibrium networks set $\widehat{p} = 2(1 + m)c$. Since the lowest potentially feasible markup is $\overline{m} = -\frac{1}{2}$ (trunk cost $c_1 = 0$) and the first-best usage price is below cost, the first-best outcome

[78] See for example Grilo, Shy, and Thisse (2001) or Farrell and Klemperer (2007) for a general overview of markets with network effects.

can never be reached with any feasible termination markup. Although the on-net price is efficient ($p = \frac{2}{3}c$), the on-net price is always distorted upwards from first-best for any feasible termination markup m, with a bill-and-keep system networks can implement the second-best outcome.

For a large set of situations, networks will benefit from agreeing on a bill-and-keep interconnection arrangement with a termination markup $m = -\frac{c_0}{c}$. Such an agreement sets the termination charge below cost and implies that networks make losses on the interconnection of calls. Whether consumers benefit from such a bill-and-keep agreement is not clear: It leads to smaller termination differentials which is beneficial to consumers due to the presence of call externalities, but softened network competition leads to larger fixed fees.

It has been suggested that the presence of closed-user groups with a concern for the price of incoming calls and biased calling patterns place competitive pressure on networks to keep termination charges low.[79] As the above analysis shows, closed-user groups exhibiting call externalities and biased calling patterns have little effect on the negotiated reciprocal termination charges: With closed-user groups, networks have an incentive to agree on a *non-positive* termination markup to avoid intense competition arising from strong positive network effects, a result identical to network competition without closed user groups.[80] On the other hand, the presence of the call externalities in closed-user groups will lead to enlarged termination differentials since externalities arising from off-net calls are not internalized.

[79] See Crandall and Sidak (2004); Competition Commission (2003).
[80] See Gans and King (2001) and Armstrong (2006).

This suggest that the regulatory focus on the mobile termination rate is misplaced if the competitive disadvantage of a smaller network primarily arises from subscribers' unwillingness to sign up with the small network in the face of large termination differentials. A substantial mass of closed-user groups should then only intensify concerns about on-net/off-net differentials.

3.8 Welfare Effects of Termination-Based Price Discrimination

The welfare effect of price discrimination is ambiguous in monopolistic as well as in competitive environments.[81] Although the principal aim of price discrimination typically is to segment markets, in this application it arises from network interconnection (termination markup on an intermediate price) and the call externalities of closed user groups. It is of particular interest to investigate the welfare effects of termination-based price discrimination considering that a suggested ban on termination-based price discrimination could render complicated regulation on termination charges obsolete (Stennek and Tangerås, 2006; Gabrielsen and Vagstad, 2007).[82]

While Section 3.5.2 showed that on-net/off-net differentials lead to higher average prices compared to a uniform marginal price, some subscribers that

[81] See Armstrong (2006) for an overview of the literature on price discrimination.

[82] As Gabrielsen and Vagstad (2007) argue, the network's market power does not arise from the termination markup per se, but rather from the effect of the access markup on equilibrium on- and off-net prices. A ban on termination-based price discrimination essentially cuts of this source of market power and networks have an incentive to price access at cost, leading to efficient consumer prices. Stennek and Tangerås (2006) argue for a ban on termination-based price discrimination as part of a structural-rules regulation in order to negate the networks' incentives for high termination rates. The particularly appealing feature of such a ban is that it requires neither cost nor demand information.

exhibit large call externalities coupled with a own-network bias could still be better off despite the higher average price - primarily due to lower on-net charges. Then again, termination differentials tend to intensify competition among networks, thereby reducing the fixed fee consumers pay. This section investigates the welfare effects on consumers and networks and finds that for any given termination markup m and demand elasticity η, termination-based price discrimination reduces social welfare when network competition takes place among equals.

Network Profits

The difference between network profits without (π_N) and with termination-based price discrimination (π_D) under symmetry is

$$\pi_N - \pi_D = \gamma \left(\frac{1}{2}\right)^{\gamma} \left(\frac{1}{2}\left[w\left(p_D\right) - w\left(\hat{p}_D\right)\right] + \left(p_D - c\right) q\left(p_D\right) - \left(\hat{p}_D - c\right) q\left(\hat{p}_D\right)\right).$$

The term in square brackets evaluated at m^* is always positive and hence $\pi_N - \pi_D > 0$. Termination-based price discrimination in the presence of closed user groups results in lower network profits. On-net/off-net differentials intensify network competition since subscribers gain from being part of the larger network which raises the network's gain from additional market share. Lower network profits are entirely due to the call externalities of closed user groups which give rise to the price differential, in fact the own-network bias of closed user groups dampens the magnitude of the profit difference: the term $\gamma\left(\frac{1}{2}\right)^{\gamma}$ is increasing in γ for $0 < \gamma < 1$ and an increased

own-network bias (reduced γ) is beneficial for the network.[83]

Consumer Surplus

The difference in consumer surplus arising from termination-based price discrimination is

$$\begin{aligned} CS_N - CS_D =& v\left(p_N\right) + \frac{1}{2}u\left(q\left(p_N\right)\right) - \left(\frac{1}{2}\right)^{\gamma}\left(v\left(p_D\right) + \frac{1}{2}\left(q\left(p_D\right)\right)\right) \\ & - \left(1 - \left(\frac{1}{2}\right)^{\gamma}\right)\left(\hat{v}\left(\hat{p}_D\right) + \frac{1}{2}u\left(q\left(\hat{p}_D\right)\right)\right) - \left(F_N - F_D\right). \end{aligned}$$

Figure 3.2 illustrates the welfare effects of termination-based price discrimination on consumer surplus for different parameter values of the termination markup m and demand elasticity η. The own-network bias parameter γ is set at 0.8, which implies that in the symmetric equilibrium roughly 57% of the calls are on-net calls.[84]

[83] In the absence of closed user groups, networks benefit from termination-based price discrimination as long as they can negotiate a reciprocal termination markup below cost. If however a regulatory authority sets a positive termination markup, networks invariably lose from termination-based price discrimination.

[84] The thin dotted line indicates how the demarcation line is shifted if the value of γ is changed to 0.4 (76% on-net calls). The own-network bias has only a small effect on the welfare comparison for consumers, a larger bias shrinking the parameter range over which consumers benefit from termination-based price discrimination.

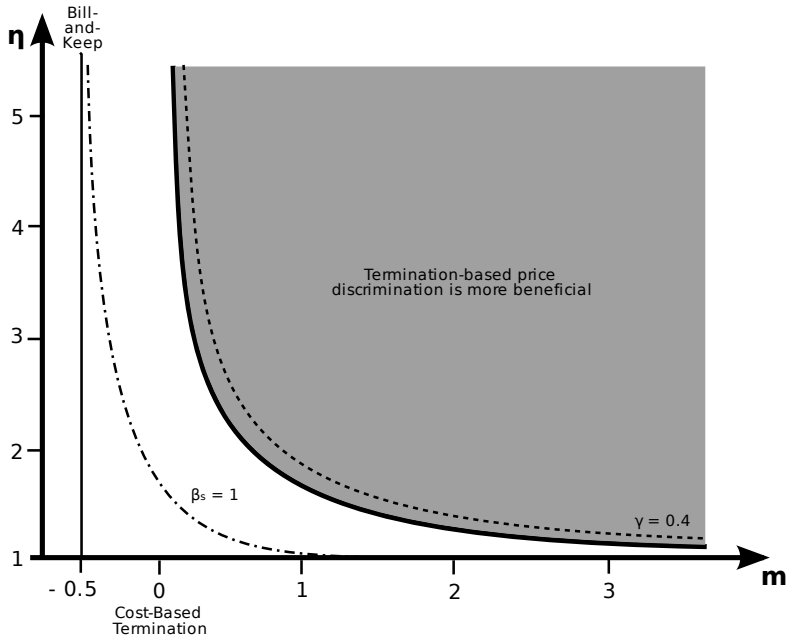

Figure 3.2: Welfare Effects on Consumer

Consumers benefit from termination-based price discrimination if either the demand elasticity η or the termination markup m is high. Due to heterogeneity with respect to the call externality, not all consumers benefit equally. Subscribers without call externalities ($\beta_s = 0$) never benefit from price differentials since the average price in equilibrium under termination-based price discrimination is higher than the uniform usage price. However, subscribers with large call externalities ($\beta_s = 1$, they share the utility of the call with the calling party) could benefit from on-net/off-net differentials even though consumers as a group lose out. Interestingly, in the most reasonable parameter range, that is, for negative termination markups negotiated between networks or regulated cost-based termination, consumers surplus overall is reduced by termination-based price discrimination.

Social Welfare

Whether price discrimination is beneficial from a social welfare point of view depends on the sign of the following term:

$$
\begin{aligned}
&CS_N - CS_D + \Pi_N - \Pi_D = \\
&\quad w(p_N) - \left(\frac{1}{2}\right)^{\gamma} w(p_D) - \left(1 - \left(\frac{1}{2}\right)^{\gamma}\right) w(\hat{p}_D) - (F_N - F_D) \\
&\quad - \gamma \left(\frac{1}{2}\right)^{\gamma} \left(\frac{1}{2}\left[w(p_D) - w(\hat{p}_D)\right] + (p_D - c)\, q(p_D) - (\hat{p}_D - c)\, q(\hat{p}_D)\right).
\end{aligned}
$$

It turns out that over the entire relevant parameter range depicted in Figure 3.2, the negative welfare effects on network profits dominate the potentially positive welfare effects on consumers surplus and social welfare is unambiguously reduced under termination-based price discrimination.

If termination-based price discrimination were welfare-improving, one would expect this to be particularly true in the presence of closed user groups: Call externalities and an own-network biased calling pattern would suggest that consumers could benefit from a larger fraction of calls made at the socially efficient on-net price that fully internalize call externalities. This model suggest otherwise: network profits are reduced due to intensified competition arising from on-net/off-net differentials but consumers only benefit if demand elasticity or termination markups are high, although the consumers with the strongest call externalities can benefit over a larger parameter range.[85] Overall social welfare is reduced compared to a situation in which termination-based price discrimination is banned.

[85] The lower fixed fees that result from intensified competition among networks do not affect social welfare comparisons since they are simply surplus/rent transfers from networks to consumers.

Essentially, on-net/off-net differentials force the networks into a prisoner's dilemma: To gain market share and attract closed-user groups, networks lower the on-net price and raise the off-net price, thereby increasing the termination differential. Not surprisingly, increased competition among networks reduces their profits. However, the increased on-net/off-net differentials also increase the pricing inefficiency, particularly on off-net calls which in term leads to reduced overall welfare.

To gain some intuition for this result, suppose that the own-network bias is large (γ is close to zero): With discrimination, the on-net price is socially efficient while the off-net price exhibits a substantial markup. Since most calls are on-net calls, one would expect social welfare to be higher compared to a uniform usage price. However, as Section 3.5.2 shows, the *average* price paid by the consumer under discriminatory pricing is still higher than under a uniform usage price. Although the uniform usage price will never be socially efficient, the call externalities coupled with a strong own-network bias narrows the difference between the uniform price and the socially efficient price and overall welfare is higher without termination differentials. This welfare analysis lends support to the contention that a ban on termination-based price discrimination could be welfare-improving and replace cost-based regulation as an alternative measure to address anti-competitive concerns about mobile network interconnection.

3.9 Concluding Remarks

This thesis chapter extended the standard model of network competition in the mobile telecommunications industry to incorporate two major features of consumers that share close repeat calling relationships within a closed user group: (i) call externalities that arise from the concern of members of a closed user group about the cost to others of making a call to them and (ii) an own-network biased calling pattern arising from the large volume of within-group calls and the coordination of network subscription choice.

The presence of closed user groups in the network competition model disentangles on-net/off-net differentials from the termination charge for network interconnection. In particular, the model shows that on-net/off-net differentials can exist even if termination charges are set at cost. Intuitively, networks fully internalize the call externalities of closed user group members for on-net calls, but not for off-net calls where the externality benefits primarily subscribers of the competitor's network. Hence, networks will charge a price differential even without any cost-based differences between these two types of calls. This result can explain why on-net/off-net differentials in the mobile telecommunications industry often substantially exceed the cost differences between an on-net and an off-net call. Furthermore, the model shows that in the presence of closed user groups, cost-based regulation of termination rates is not equivalent to the elimination of on-net/off-net differentials.

If two network operators of different size compete, the model shows that the on-net/off-net differential of the larger network exceeds the differential

charged by its smaller competitor. Compared to the standard model of network competition, the incorporation of closed user groups with call externalities increases on-net/off-net differentials even further. If there exist anti-competitive concerns about smaller network disadvantaged by large on-net/off-net differentials, the presence of closed user groups will only exacerbate the smaller networks disadvantage. This results stands in contrast to the commonly advanced argument that any anti-competitive effect of high termination charges and its resulting on-net/off-net differentials are mitigated by the presence of closed user groups.

Additional results from the welfare analysis suggest that the overall effect of termination-based price discrimination is welfare-reducing even in the presence of closed user groups. The majority of consumers and both network operators are worse off with termination-based price discrimination. This result lends additional support to the contention that a ban on termination-based price discrimination could improve welfare and replace cost-based regulation of termination charges.

Future Research

A major drawback of this study is the treatment of closed user groups as a black box. The model does not incorporate the coordination of subscription decisions - sign-up and switching behavior - and how they depend on network pricing, nor does it distinguish between receiving calls from within the social network from calls outside the social network. A more realistic model of social networks would build closed user groups from the ground up, thereby incorporating the feature that closed user groups differ in their ability to

coordinate network subscription behavior and hence differ in their relative shares of on-net and off-net calls. By targeting different consumer segments through on-net/off-net price differentials, a network could then influence the distribution of on-net/off-net calls even while market share remains fixed. Such an analysis is complicated by the fact that the share of on-net calls now depends on the overall market share of the network *and* the particular composition of closed user groups of its subscribers.

A comprehensive model of social networks would also be more applicable to real-world mobile telecom pricing characterized by a multitude of plans (couple, friends & family, myFaves, business, etc.) that target particular types of social networks and closed user groups, a screening strategy that is explicitly ruled out in the model presented here. The inner workings of such social networks and their coordination of network subscription decisions are a promising area of future research but are outside the scope of this thesis chapter.

Chapter 4

An Empirical Investigation of Consumer Behavior and Choice of Nonlinear Cellphone Tariff [86]

4.1 Introduction

Significant growth in the wireless telecommunications industry over the last decade has raised the need and interest in understanding consumer behavior and demand for cellular services. Two major features of cellular telecommunication services distinguish it from more traditional products, features that are also prevalent in other services such as internet access, car rental, or health club services: Firstly, most cellular communication services are subscription-based, that is, consumers choose at the beginning of the billing period a service option and tariff that will subsequently determine their bill payment. Secondly, cellular firms generally offer a menu of pricing options (plans), each of which gives consumers access to virtually the same com-

[86]I would like to thank Carl Mela, Raghuram Iyengar and the Teradata Center at Duke University for providing the wireless subscriber data as well as Allan Keiter and MyRatePlan.com, LLC for providing data on the availability and characteristics of cellular phone plans in the United States.

munication services on the firm's wireless network. It is fairly common in the industry that the menu of tariff options comprises of plans that feature a three-part tariff structure consisting of a fixed monthly fee, an included allowance of minutes and an overage price for consumption in excess of the monthly allowance. This thesis chapter explores consumer behavior, cellular plan choice and the effects of demand uncertainty under three-part tariffs based on a large consumer-level dataset from a major cellular service provider in the United States.

Consumer behavior is likely to be different under three-part tariffs compared to the more common two-part pricing structure observed in many other industries. Two-part tariffs include a fixed fee but no "free" allowance, that is, the consumer encounters a constant marginal price independent of the level of consumption. In contrast, a consumer subscribing to a plan with a three-part pricing structure faces a marginal price of zero if usage remains within the monthly allowance, but pays a positive marginal price (overage price) for consumption in excess of the allowance. The varying marginal price has different implications on consumer bills if the tariff choice and consumption decisions are temporally separated.

Cellular consumers choose to subscribe to a particular tariff option before they know their exact consumption needs. Consumption is subsequently determined over the course of the billing period based on the consumers' chosen cellular plan. Since there exists demand uncertainty over the usage at the time of tariff subscription, the billing implications of two-part tariffs differ from three-part tariff. For a two-part tariff, the total monthly bill fluctuates linearly with usage and is unaffected by (symmetrically) distributed demand

variation. On the other hand, under three-part pricing, the relationship between usage and total monthly bill payment is convex since usage below the allowance does not affect the bill. Consequently, the expected bill payment is affected by consumers' demand variation. Demand uncertainty does not only affect consumer behavior, but it is also a major factor in cellular providers' design of the menu of tariff options as the additional allowance parameter provides additional pricing flexibility, with the potential to increase providers' profit.

While the determinants of consumer choice behavior under two-part tariffs have been thoroughly researched for fixed (land-line) telecommunications (Train, McFadden, and Ben-Akiva, 1987; Train, Ben-Akiva, and Atherton, 1989; Kling and Van Der Ploeg, 1990; Mitchell and Vogelsang, 1991; Miravete, 2002; Narayanan, Chintagunta, and Miravete, 2007) and in the early US cellular telecommunications industry (Miravete, 2009), there exists very limited empirical evidence on consumer behavior in the presence of three-part tariffs. Huang (2008) estimates demand for cellular phone service under three-part pricing using firm-level data to analyze the cellular market in Taiwan. Using household-level data, Economides, Seim, and Viard (2008) study entry into residential local telephone service employing a discrete/continuous demand model that accounts for three-part tariffs and Lambrecht, Seim, and Skiera (2007) study consumer choice of internet service provider when consumers can choose between a flat-rate option and three-part tariff plans. More closely related to this study of the wireless telecommunications industry are Iyengar (2004) who examines plan switching and customer churn under three-part cellular tariffs and Guo and Erdem (2006) who explore the

distinction between usage variability and usage uncertainty of consumer behavior in the cellular industry. This study contributes to the literature by extending the standard discrete-continuous model of consumer plan-choice and usage decisions (Hanemann, 1984) using detailed consumer-level data from a major US cellular service provider. The analysis takes account of the particular pricing features of three-part tariffs and the availability of cellular plans. The model makes allowance for demand uncertainty through the temporal separation and interdependence of the discrete tariff choice and the continuous usage decision and investigates its consequences on consumer behavior and tariff choice. Furthermore, tariff-specific preferences are included in the model to explore the magnitude of consumer biases previously identified in empirical studies, particularly in the telecommunications industry (Train, McFadden, and Ben-Akiva, 1987; Kling and Van Der Ploeg, 1990; Mitchell and Vogelsang, 1991; Kridel, Lehman, and Weisman, 1993; Lambrecht and Skiera, 2006).

The remainder of this chapter is organized as follows: Section 4.2 describes the data and provides summary statistics of the consumer-level data. Next, the discrete/continuous model of cellular plan choice is developed and the estimation method discussed. Section 4.4 presents the estimation results and policy simulations, and discusses the implications of the findings. The chapter ends with concluding remarks.

4.2 Data

The dataset used in this study contains subscriber-level monthly billing records from a major national U.S. wireless service provider (Iyengar, 2004). Billing observations range from September 2001 to April 2003, with no subscriber billing information available for February 2002.[87] The dataset contains consumers (excluding business clients) who started to subscribe to wireless calling services between August and December 2001 who choose among four cellular calling plans offered by the service provider. Consumers have no contractual relationship with the service provider and are free to switch among the plans offered by the service provider. The monthly bill is determined at the end of the month based on actual cellular minutes consumed and the subscribed cellular plan.

Cellular phone plans generally distinguish between peak and off-peak (evening and weekend) minutes and may come bundled with other add-on services such as short messaging system (SMS), long-distance and/or roaming packages. Offpeak minutes are often included in the plan (or are charged a flat fee) and it is fairly common for cellular firms to give unlimited offpeak minutes. Furthermore, the use of add-on features is negligible and this study therefore focuses exclusively on the consumption of peak minutes under the three-part pricing structure of cellular plans that is composed of the monthly fixed fee, the included peak minutes (allowance) and the per-minute price for consumption in excess of the allowance (overage price).

[87] The first month of billing information for each subscriber is ignored since the billing period does not span an entire month and the monthly plan fee and allowance of minutes are prorated.

In addition to information on subscribers' usage and plan choice, the dataset contains information indicating whether the consumer is a new cellphone user as well as demographic information such as age and the consumer's PRIZM code.[88] Using the PRIZM code, consumers were classified into three categories based on their neighborhood characteristics - urban, suburban and rural - and into four categories based on their income - rich, upper middle class, middle class and poor.[89]

The study limits the set of plans in the consumers' choice set to the four cellular plans observed in the dataset. This approach ignores the influence on consumers' choice behavior of competitors' plan offerings or additional plans offered by the same cellular service provider. Despite this restriction, the choice set nevertheless varies across consumers and months for following two reasons:

Temporal Availability: Information on the availability of cellular phone plans in the 25 largest metropolitan markets in the United States obtained from MyRatePlan.com reveals that the four cellular plans in the dataset were offered nationwide by the cellular service provider, but not all four plans were offered over the entire data range. This implies that consumers face

[88] PRIZM is a household-level consumer segmentation system by Nielsen Claritas that classifies consumers into categories based on income and neighborhood characteristics.

[89] Urban areas have high populations density. Included in the category are downtowns of major cities and surrounding neighborhoods, urban areas often extend beyond the city limits and into surrounding jurisdictions. Suburban areas are characterized by moderate population density and the category includes moderately-sized independent cities, satellites cities in major metropolitan areas and suburbs of larger cities. Rural areas exhibit low population density and the category includes exurbs, towns, farming communities and a wide range of other rural areas. Households with annual household income above $50,000 are classified as rich, households with annual income in the range from $37,000 to $50,000 are considered upper middle class, households with annual income between $20,000 and $37,000 are considered middle class and households with annual income below $20,000 are considered poor.

different choice sets at different points in time.[90]

Grandfathering of Cellular Plans: Tariff plans in the wireless telecommunications industry are generally grandfathered if the service provider pulls a particular tariff option from the market. This means that consumers who subscribe to a plan that is being pulled from the market can continue to stay on that plan even though the same plan is not offered anymore to new subscribers (or customers who would like to switch cellular plans).[91] Hence, at any particular point in time, consumers face different choice sets depending on the plan they are currently subscribing to. Since the dataset contains information about the consumer's plan choice in the previous month, the empirical model in Section 4.3 adjusts consumers' choice sets to account for the grandfathering of cellular phone plans.

The temporal availability of the four cellular plans and the grandfathering clause are illustrated in Figure 4.1:

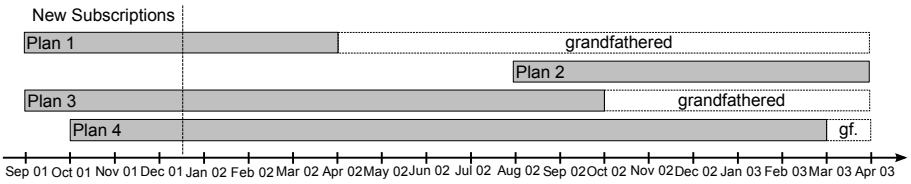

Figure 4.1: Temporal Availability of Cellular Plans

Due to computational intensity, a cross-sectional estimation is chosen

[90] A few observations for which the plan choice is inconsistent with plan availability information obtained from MyRatePlan.com have been eliminated from the selected study sample. The timing of market introduction and discontinuance of cellular plans is identical in all metropolitan areas and the empirical model is therefore estimated as one single (national) market.

[91] Cross-checking consumers' plan choices in the dataset with information on cellular plan availability from MyRatePlan.com confirms that the cellular service provider indeed employs such a grandfathering clause.

rather than a more flexible panel estimation with fixed effects.[92] The selected sample dataset employed in the estimation of the discrete-continuous model in this study thus contains a cross-section of consumers with a total 11,479 observations. For each consumers, one month was randomly selected over the time period the consumer is observed in the dataset. Table 4.1 presents summary information and characteristics of the four cellular plans offered by the provider. The four plans differ in their allowance of minutes included, which ranges from 200 minutes for Plan 1 to 500 minutes for Plan 4. The fixed monthly fee increases with the number of minutes included in the plan allowance, whereas the per-minute price for consumption in excess of the allowance (overage price) is the same across all four cellular plans, 40 cents per minute.

Cellular Plans	**Fee** $/month	**Allowance** minutes	**Overage Price** $/minute	**Sample Observations**	**Mean Usage** minutes
Plan 1	30	200	0.40	5,817 (50.7%)	124.19
Plan 2	35	300	0.40	386 (3.3%)	183.01
Plan 3	40	350	0.40	4,275 (37.2%)	232.03
Plan 4	50	500	0.40	1,001 (8.7%)	335.54

Table 4.1: Summary Information on Cellular Plans

The descriptive statistics show that the average number of monthly min-

[92] Demographic information contained in the dataset allows to control for observed demand heterogeneity in the chosen cross-sectional estimation, while a panel estimation could also control of unobserved demand heterogeneity. In a similar study on internet service under three-part tariffs, Lambrecht, Seim, and Skiera (2007) compare a cross-sectional estimation with a panel estimation on a sub-sample of their data and find that consumer-specific, time-invariant unobserved differences are not very important and their results are rather driven by high within-consumer usage variation. Guo and Erdem (2006) and Iyengar (2004) make use of the panel structure of the same dataset used in this study, but they either focus on the small subset of consumers who switch among the four plans offered or the model (incorrectly) assumes that all four plans were offered at all times during the period (choice set is not anymore individual and time-specific).

utes consumed increase with cellular plans that include a larger allowance, but fall significantly short of each plan's allowance. In fact, the average number of minutes consumed on a particular plan is even below the allowance of the plan with the next smaller allowance.

To illustrate the distribution of monthly usage of minutes, Figure 4.2 displays the histogram of the ratio of monthly consumption of minutes relative to the plan allowance. It is apparent that the distribution is right-skewed and a significant share of consumers use less than their monthly allowance. 80.6% of consumers in the selected sample consume less than their plan allowance, 44.2% consume less than half their monthly allowance of minutes. On the other end, the consumption of minutes drops sharply once the plan allowance is exceeded, only 4.9% of consumers exceed their minute allowance by more than 50% while only 1.8% of consumers use more than twice their monthly allowance of minutes.

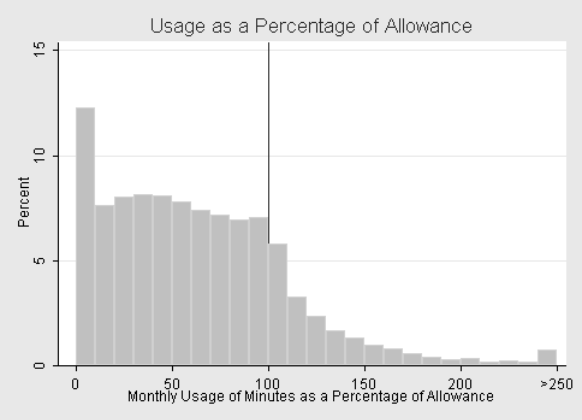

Figure 4.2: Monthly Usage of Minutes as a Percentage of the Plan Allowance

Taking into account temporal availability and grandfathering of cellular

plans, one can evaluate for each consumer in the selected sample whether the chosen cellular plan minimizes calling costs based on actual usage. Table 4.2 shows the plan choice matrix where rows represent the chosen cellular plan and columns represent the ex-post cost-minimizing plan that is available to the consumer. In the diagonal cells are consumers for which the chosen cellular plan is cost-minimizing given their actual (ex-post) consumption of minutes. Observations for which consumers would have had lower ex-post calling costs had they chosen a plan with a larger allowance of minutes are represented by cells above the diagonal. In cells below the diagonal are consumers for which a plan with a smaller allowance would have resulted in lower ex-post calling costs given their actual usage of minutes.

(N=11,479)	**Ex-Post Cost-Minimizing Plan**			
Chosen Plan	Plan 1	Plan 2	Plan 3	Plan 4
Plan 1	**86.8%**	3.8%	6.5%	2.8%
Plan 2	11.7%	**83.9%**	1.0%	3.4%
Plan 3	23.6%	21.1%	**41.0%**	14.4%
Plan 4	17.0%	15.4%	22.2%	**45.5%**

Table 4.2: Chosen versus Cost-Minimizing Cellular Plan

It becomes apparent from inspecting the cellular plan choice matrix that there seems to exist a tendency of consumers to subscribe to a plan that includes an allowance that is "too big", that is, a plan with a smaller allowance would have resulted - ex-post - in lower calling costs. For example, while most consumers subscribing to Plan 1 and Plan 2 have chosen the ex-post cost-minimizing plan, a large share consumers that subscribe to Plan 3 and Plan 4 would have incurred lower calling costs ex-post had they subscribed to a plan with a smaller allowance.

This puzzling empirical regularity has been well-documented in the telecommunications industry. Numerous studies based on transactional data with two-part tariffs indicate that consumers prefer a flat-rate tariff even though they could save money by choosing a pay-per-use tariff instead. This inability to anticipate future consumption and minimize expenditure accordingly is commonly referred to as the "flat-rate bias" (Train, McFadden, and Ben-Akiva, 1987; Mitchell and Vogelsang, 1991; Miravete, 2003; Lambrecht and Skiera, 2006). In the context of three-part tariffs, this systematic tendency of consumers to over-buy services (e.g. cellular calling minutes) that they subsequently fail to consume would manifest itself in a systematic bias towards cellular plans whose allowance significantly exceed usage.

One should be careful not to rush to premature conclusions about systematic consumer biases from Table 4.2 presented above. There is a potential rational explanation for the finding that the ex-ante chosen cellular plan is not ex-post cost-minimizing: Given the consumption uncertainty at the time of cellular plan choice, the ex-ante optimal plan might not be cost-minimizing ex-post if realized consumption deviates from expected consumption.[93] While consumption uncertainty in itself is insufficient to explain an asymmetry in ex-post deviations from the cost-minimizing cellular plan, three-part pricing with an increasing per-minute price can provide an explanation: If the per-minute price is constant - either zero (unlimited plan) or

[93]Given the widespread practice of grandfathering of the consumer's current cellular plan, it might not be optimal for consumers to switch to the cost-minimizing tariff in any given month *even if* consumption is perfectly known at the time of plan choice. If consumers expect consumption in future months to be different, switching to the cost-minimizing tariff might reduce the plan choice set in future months if the current plan is not offered anymore by the service provider.

positive (two-part tariff) - actual consumption realizations above and below expectations cancel each other out and no plan choice asymmetry arises. A three-part tariff however has the property that the marginal price depends on ex-post usage. With the per-minute marginal price increasing, actual consumption realizations above expectations are more costly then realizations below expectations. Consumption uncertainty may then lead to a bias towards larger allowances as consumers try to maintain usage flexibility and avoid paying high overage rates.

The descriptive evidence of cellular plan choice highlighted certain behavioral patterns of consumers, but a thorough analysis requires a econometric model that explicitly incorporates the sequentiality and interdependence between tariff choice and usage. Such a structural demand model is presented in the next section.

4.3 Empirical Demand Model Specification

The structural demand model developed in this section is an extension of the standard discrete/continuous demand model Hanemann (1984) and builds off similar models by Economides, Seim, and Viard (2008) who study entry into residential local phone service and Lambrecht, Seim, and Skiera (2007) who study the plan choice and consumption for internet access. Cellular plan subscribers indexed by $i = 1, 2, \ldots, I$ choose in month t a single plan from the set of available plans $j = 1, 2, \ldots J$ and a quantity of peak minutes q_{ijt} to consume on that particular plan.[94] By choosing plan j, the consumer must

[94] In the data, a consumer account is associated with a unique cellular plan. It is theoretically possible that consumers consume on multiple plans by opening more than

pay a fixed fee F_j and receives an allowance A_j of free minutes included with the plan. The overage price for consumption in excess of the plan allowance is denoted p_j. Hence, the marginal price of consumption on three-part tariff j is equal to zero for consumption below the allowance A_j and p_j for minutes consumed above the plan allowance. Subscribers spend the remaining part of their income y_{it} on the numeraire good z_{it}, whose price is normalized to 1. Subscriber i chooses the plan j that maximizes utility subject to the budget constraint.

4.3.1 Consumer Utility

Utility is specified to be quadratic, a simple functional form that accommodates the main features of telecommunications demand: Marginal utility of cellular phone usage declines with usage, which implies that demand is bounded even on flat-rate plans (unlimited minutes included in the allowance) where the marginal price of consumption is zero.[95] In addition, a quadratic utility specification allows for zero consumption of the good in a situation where the marginal price is zero.[96]

The consumer's optimization problem consists of two steps: First, the consumer chooses consumption to maximize utility on a *given* cellular plan. In a second step, the consumer then chooses the cellular plan that yields the highest *expected* utility among all the cellular plans offered. The utility

one account, but this is not a widespread phenomenon in the cellular industry and is ignored for the purpose of this study.

[95] At some point, subscribers spends so much time on the phone that it crowds out time spent on outside activities.

[96] In the selected sample dataset, around 5% of monthly consumer bills show no consumption of peak-minutes even though the marginal price of consumption is zero (i.e., the subscriber is well within the allowance).

maximization problem of the consumer for a particular cellular plan can be written as

$$\max_{q_{ij},z_i} U\left(q_{ijt},z_i\right) = z_i + \frac{1}{\beta}\left(\alpha_{it}q_{ijt} - \frac{q_{ijt}^2}{2}\right) - \frac{\alpha_{it}^2}{2\beta} + \xi_{ij}$$

subject to

$$y_i \geq z_i + F_j + \max\left(q_{ijt} - A_j, 0\right)p_j.$$

The error term ξ_{ij} reflects the subscriber's tariff specific preferences. Conditional on the choice of plan j, the associated conditional demand function for usage q_{ijt} is

$$q_{ijt} = \begin{cases} \alpha_{it} - \beta p_j & \text{if } p_j < \frac{\alpha_{it}-A_j}{\beta} \\ \alpha_{it} & \text{otherwise} \end{cases}. \tag{4.1}$$

Note that if the subscriber remains within the included allowance A_j (or were to subscribe to a plan that includes unlimited minutes), demand simplifies to $q_{ijt} = \alpha_{it}$.

Since only the total monthly minutes consumed by a subscriber (q_{ijt}) are observed, it is implicitly assumed that there is only a single consumption decision made during a particular month. Specifically, the estimated model cannot account for sequential consumption decisions over the billing cycle that were explored in Chapter 2 as a potential explanation for the prevalence of three-part tariffs in the cellular telecommunications industry.

Substituting the conditional demand functions back into the subscriber's utility function yields a set of conditional indirect utility function that vary

by type of plan and usage

$$V_{ijt}\left(p_j, F_j, y_i\right) = \begin{cases} y_i - F_j + \xi_{ij} & \text{if } q^*_{ijt} \leq A_j \\ y_i - F_j - \left(\alpha_{it} - A_j - \frac{1}{2}\beta p_j\right) p_j + \xi_{ij} & \text{if } q^*_{ijt} > A_j \end{cases}$$

The demand intercept is allowed to vary with the subscriber's observable and unobservable characteristics. For the demand to be well-specified, the demand intercept α_i is restricted to be positive by specifying it as an exponential function:

$$\alpha_{it} = e^{\kappa + \delta_D D_i + \delta_T T_t + \upsilon_i} \tag{4.2}$$

The vector D_i contains demographic characteristics of the individual (such as income class, age, population density), while the vector T_t includes time-specific dummy variables to account for seasonal variation.[97] The error term υ_i, accounts for demand uncertainty and is assumed to be normally distributed with mean zero and standard deviation σ_υ. The usage shock affects conditional demand by shifting the demand intercept α_{it}. At the time of choosing the cellular plan, the consumer only knows the distribution of the consumption shock but not the actual realization of the shock.

The subscriber's tariff specific preference ξ_{ij} is decomposed into

$$\xi_{ij} = \gamma Z_T + \epsilon_{ij}$$

where the vector Z_T includes plan-specific dummies that can account for biases towards specific cellular plans. As previously explained, such biases

[97] The sample mean age has been imputed for observations for which information on age was missing.

towards specific plans have been well-documented in the telecommunications literature. In addition, the plan-specific dummies are interacted with the new cellular subscriber indicator variable to explore whether subscribers without much experience with cellular services and pricing exhibit different behavior. Furthermore, the vector Z_T also includes a status quo dummy variable that measures the subscribers preference for the currently chosen plan. The status quo dummy variable takes on the value 1 if the consumer is still subscribing to the same cellular plan as in the previous month.[98] The error term ϵ_{ij} is a vector of unobservable plan preferences, which is assumed to be distributed according to a Type-1 extreme-value distribution (logit).

In the specified consumer demand model, the unobservable component in ξ_{ij} varies by cellular plan and affects only the discrete choice but not the quantity choice (demand). The unobservable component in the demand intercept α_{it} in contrast affects the quantity consumed, but the discrete choice only indirectly through the quantity choice. Hence, the two unobservables characteristics ϵ_{ij} and υ_i are assumed to be independent. A specification using this type of model error structure assumes that there are no unobservable components of the cellular plan that affect the quantity choice. This assumption can be justified based on the fact that cellular plans offered by the service provider grant access to the same network with identical call quality, network coverage and customer service.[99]

Given the consumer's demographic characteristics and plan pricing, the

[98] The status quo plan chosen in the previous month is known since the original data observes consumers each month as long as they subscribe to one of the four plans.

[99] Correlation between unobservable plan choice characteristics and the quantity choice could arise if the provider ran user and plan-specific advertising campaigns and decided to promote certain plans specifically to consumers who exhibit a certain demand behavior.

realized consumption on three-part cellular plans q_{ijt} is a nonlinear function of the usage shock υ_i. For realized consumption shocks $\upsilon_i < \ln(A_j) - \kappa - \delta_D D_i - \delta_T T_t$, consumption occurs below the allowance where the marginal price per-minute is zero. For values $\upsilon_i < \ln(A_j + \beta p_j) - \kappa - \delta_D D_i - \delta_T T_t$, consumption exceeds the plan allowance A_j and the consumer pays a positive overage price p_j. For realized consumption shocks in the intermediate interval $\ln(A_j) < \upsilon_i - \kappa - \delta_D D_i - \delta_T T_t < \ln(A_j + \beta p_j)$, consumption is equal to the allowance A_j. Hence the model predicts a mass point in the distribution of consumption at the plan allowance due to the three-part nature of cellular plan pricing. The magnitude of the mass point will depend on the standard deviation of the consumption shock. Figure 4.2 suggests that the bump or mass point in the consumption distribution at the allowance is relatively small, or in other words, the standard deviation of the consumption shock is relatively large. Another mass point in the distribution of usage occurs due to the fact that peak minutes consumed (including the allowance) under all plans offered by the cellular service provider are capped at 1,500 minutes per month.

The consumer chooses a cellular plan at the beginning of the month that yields the highest *expected* utility. While tariff specific preferences (ξ_{ij}) are unobserved by the researcher but known to the consumer, expectations are only taken with respect to the consumption shocks υ_i. Expected indirect

utility on a three part tariff is composed of two terms given as

$$\begin{aligned} E[V_{ij}^*] &= \Pr(q_{ijt}^* \leq A_j) E[V_{ij}^* | q_{ijt}^* \leq A_j] + \Pr(q_{ijt}^* > A_j) E[V_{ij}^* | q_{ijt}^* > A_j] \\ &= \Pr(q_{ijt}^* \leq A_j)(y_i - F_j) + \Pr(q_{ijt}^* > A_j)\Big[y_i - F_j + p_j A_j + \frac{1}{2}\beta p_j^2 \\ &\qquad - e^{\kappa + \delta_D D_i + \delta_T T_t + \upsilon_i} E(e^{\upsilon_{it} | q_{ijt}^* > A_j}) p_j\Big] \\ &= \overline{V}_{ijt}^* + \xi_{ij} \end{aligned} \tag{4.3}$$

If consumption remains below the plan allowance A_j, the consumer only pays the monthly fixed fee and no overage charges. If however consumption exceeds the plan allowance A_j, the consumer pays additional overage charges for consumption in excess of the allowance. The demand uncertainty (σ_υ) affects the tariff choice through the likelihood of incurring overage charges and the additional payment if such overage charges are incurred. The consumer chooses the cellular plan that maximizes expected utility in Equation 4.3 from the set of available plans, and then subsequently chooses consumption of minutes where demand follows Equation 4.1.

4.3.2 Model Estimation

Estimation of the model is based on two observed consumer decisions, the choice of cellular plan d_i and the corresponding usage choice q_{ijt}. The structural model predicts the optimal plan and usage as a function of a consumer's observable and unobservable characteristics and the plan's observable and unobservable attributes. Following the empirical studies by Economides, Seim, and Viard (2008) and Lambrecht, Seim, and Skiera (2007), the model is estimated using maximum likelihood methods:

θ	Vector of model parameters
d_i	Vector of subscriber's plan choices
q_{ijt}	Vector of subscriber's usage choice
$P(F_j, p_j)$	Vector of cellular plan characteristics
D	Vector of consumer characteristics
Z	Vector of plan-specific dummy variables
T	Vector of time dummy variables

Denote the entire variable vector by $X(D, P, Z, T)$. The full likelihood is the product of the likelihood for each consumer:

$$\mathcal{L}(\theta|\, d, q, X) = \prod_{i=1}^{I} \mathcal{L}_i(\Theta|\, d_i, q_{ijt}, X_i)$$

The log-likelihood of a subscriber is the joint probability of the subscriber's plan choice d_i and its quantity choice q_{ijt}. This joint probability can be written as the product of the probability that the subscriber i chooses plan j conditional on the usage shocks v_i and the probability distribution of q_{ijt}. Subscriber i's contribution to the log-likelihood equals

$$l_i(\theta|\, d_i, q_i, X_i) = \sum_{j=1}^{J} \mathcal{I}_{d_{ij}} \ln\left(f(d_i|v_i, X_i; \theta)\, g(q_{ijt}|\, X_i; \Theta)\right) \tag{4.4}$$

where $\mathcal{I}_{d_{ij}}$ is an indicator variable equal to 1 if subscriber i chooses plan j and zero otherwise. $f(d_{ij}|v_i, X_i; \Theta)$ is the conditional likelihood of observing subscriber i choosing plan j, while $g(q_{ijt}|X_i; \Theta)$ is the likelihood of usage q_{ijt}.

The contribution of usage to the likelihood is given by

$$g\left(q_{ijt}|X_i;\theta\right)=\begin{cases}\Phi\left(\upsilon_i^0\right) & \text{if } q_{ijt}=0\\ \phi\left(\upsilon_i\right)\mathcal{J}_i & \text{if } 0<q_{ijt}<1,500 \text{ and } q_{ijt}\neq A_j\\ \Phi\left(\overline{\upsilon}_i^{A_j}\right)-\Phi\left(\underline{\upsilon}_i^{A_j}\right) & \text{if } q_{ijt}=A_j\\ 1-\Phi\left(\upsilon_i^{1,500}\right) & \text{if } q_{ijt}=1,500\end{cases}$$

where ϕ and Φ denote the normal probability density and distribution functions of υ and $\mathcal{J}_{it}$ is the Jacobian of the transformation of υ_i to q_{it}. The values $\upsilon_i^0 \geq -\kappa-\delta_D D_i-\delta_T T_t$, $\overline{\upsilon}_i^{A_j}$ and $\underline{\upsilon}_i^{A_j}$ respectively, and $\upsilon_i^{1,500}$ are the cutoff values of υ_i related to the three mass points of the usage distribution that entail consumption at zero, at the allowance A_j, and at the peak minute cutoff of 1,500 minutes.

For any potential vector of parameter values θ, the probability that consumer i chooses plan j in month t is given by the integral of the distribution of plan preferences over the choice shock ϵ_{ij} such that cellular plan j gives maximal expected utility. Expected utility of a cellular plan is set to the average utility obtained from 50 random draws of the demand shock υ_i from the normal distribution with mean zero and standard deviation σ_υ. Due to the independence assumption of choice and usage shock, the conditional distribution of the choice shock given the usage shock remains an type-1 extreme value distribution (logit), and the probability that subscriber i chooses plan j is simply

$$f\left(d_{ij}|\upsilon_i,X_i;\theta\right)=\frac{\exp^{\overline{V}_{ij}^*(\upsilon_i,X_i;\theta)}}{\sum_{k=1}^{J}\exp^{\overline{V}_{ik}^*(\upsilon_i,X_i;\theta)}}$$

The parameters in the model, θ, are identified through both, the discrete choice of a cellular plan and the continuous usage choice. The parameters of the demand function, α_{it} and β are identified through the variation of usage. The coefficients of the individual subscriber characteristics, D_i, are identified by systemic variation in consumption of consumers with different attributes, while the variance σ_υ is pinned down by the remaining unexplained variation in consumption.

4.4 Estimation Results and Pricing Implications

4.4.1 Estimation Results

The estimated results from the discrete/continuous model of cellular plan choice based on the selected sample of 11,479 observations are summarized in Table 4.3. The parameter estimates measure the effect of the various independent variables on the dependent variables in the model (plan choice d_i and consumption of minutes q_{ijt}).

The parameter estimates on the plan-specific dummies are all highly significant: First, subscribers exhibit a strong status quo bias for their current plan choice. This result is well-known in the cellular industry that consumers do not switch cellular plans very often. Second, the estimates for the plan-specific dummy variables suggest that subscriber's plan choice does systematically deviate from the optimal plan, but the bias is not uniform.

The overall estimates suggest that subscribers significantly undersubscribe to cellular plans 3 and 4 that feature the larger allowances with parameter estimates equal to -9.452 and -8.925 respectively. This is quite sur-

Estimation Results (N=11,479)	Parameter Estimate	Standard Error
Tariff Choice Preferences[A] (γ)		
Plan 2	2.692***	0.346
Plan 3	-9.452***	0.278
Plan 4	-8.925***	0.238
Status Quo	17.196***	0.242
New Cellular Users		
Plan 2	3.121***	0.355
Plan 3	3.893***	0.816
Plan 4	3.238***	0.489
Demand Slope (β)	12.263***	2.633
Usage Uncertainty (σ_υ)	1.558***	0.011
Demand Intercept (α)		
Constant	5.342***	0.066
Seasonal Variation[A] (δ_T)		
February	-0.088	0.082
March	-0.237***	0.057
April	-0.156***	0.057
May	-0.028	0.068
June	-0.204***	0.067
July	-0.031	0.070
August	-0.041	0.071
September	-0.023	0.067
October	0.114*	0.066
November	0.054	0.063
December	0.210***	0.062
Demographics (δ_D)		0.062
Urbanicity[A]		
Urban	0.025	0.041
Suburban	0.015	0.034
Income[A]		
Rich	0.005	0.039
Upper Middle Class	-0.005	0.042
Poor	-0.186	0.040
Age[B]	-0.054	0.040
New Cellular Consumer	-0.016***	0.001

Dependent Variables: Cellular Plan Choice (d_i) and Consumption of Minutes (q_{ijt}).
[A] The base categories are Plan 1, January, *Rural* and *Middle Class*.
[B] Missing Values have been imputed by the sample mean.
Statistical significance of parameter estimates: *(p<0.1), **(p<0.05) and ***(p<0.01)

Table 4.3: Parameter Estimates

prising contrasted against the plan-choice matrix in Table 4.2. This finding can however be explained by the convexity of three-part tariffs using a simple example: Suppose that a subscriber consumes 180 minutes with probability $p = 0.8$ and 480 minutes otherwise and consider the subscriber's choice between Plan 1 and Plan 4. The expected monthly bill for Plan 1 is \$52.40 while the expected monthly bill for Plan 4 is \$50. A plan-choice matrix similar to Table 4.2 would erroneously indicate that 80% of such subscribers to Plan 4 do not choose their optimal cellular plan and could have saved money by choosing Plan 1 instead. The plan-specific dummy variable estimates in this structural model do not suggest that overall subscribers show a bias for plans that include a allowance that is "too large" given their demand profile, in fact, quite the opposite is true: Overall, subscribers tend to significantly oversubscribe to plans 1 and 2, the two plans with the smaller allowances.

The picture changes slightly when the focus shifts to new cellular users (first-time subscribers): New cellular subscribers tend to prefer plans with larger allowances compared to Plan 1 more *relative* to all subscribers, the parameter estimate is around 3 on the plan-specific dummy variable. Nonetheless, in absolute terms, first-time cellular users display a similar pattern with a general preference for plans 1 and 2 that include smaller allowances.[100] Hence, the results on tariff-specific preferences from this study do not support previous findings from two-part tariff environments of a broadly based and general "flat-rate bias" in tariff choice.

After customers have selected their preferred cellular plan, they choose

[100]The overall plan-specific effect for new cellular users is obtained by adding the plan's two parameter estimates.

their monthly consumption of minutes based on the subscribed cellular plan. The estimate of the demand slope parameter β is 12.263 with a standard error of 2.633. The result indicates that consumers do respond to the overage price, although only to a relatively minor extent. The overage price of 40 cents reduces subscribers usage by about 5 minutes if usage exceeds the allowance.

Demand uncertainty measured by the standard deviation of the usage shock σ_υ is estimated at 155 minutes, or over 80% of average usage of minutes in the sample. This is quite substantial and suggests that demand uncertainty is one of the main factors that drives consumption and choice behavior in the wireless telecommunications industry. However, the parameter measuring demand uncertainty (σ_υ) also contains unobserved consumer heterogeneity that the chosen model specification does not explicitly control for, such as additional demographic parameters (household size) or socioeconomic factors (education). Hence, the true underlying consumer demand uncertainty is likely to be lower than estimated.

Among demographic variables that allow to control for observed demand heterogeneity, consumption decreases statistically significant with age and new cellphone users tend to have lower demand although the parameter is not statistically significant. The parameter estimates suggest that income does not affect consumption in a statistically significant way across most income categories except for low-income consumers who have significantly lower consumption of cellular services. Furthermore, urban and suburban consumers consume more wireless services than rural consumers but neither parameter is statistically significant. The estimated parameters for the

seasonal dummy variables are in line with industry facts and suggest that consumption of cellular phone services tends to be lower in the first half of the year, with demand picking up considerably in the last quarter of the year, particularly in December.

4.4.2 Plan-Choice and Usage Elasticities

Plan-choice and usage elasticities with respect to fixed fee, allowance and the overage prices are obtained by estimation customer plan choice and consumption of minutes in response to increasing the relevant parameter by 1%.[101] Of particular interest are the changes of consumers' cellular plan choice and usage in response to a change in one of the three pricing characteristics of a three-part tariff. The resulting estimated elasticities are presented in Table 4.4. The columns of Table 4.4 present the plan-choice and usage elasticities with respect to the monthly fee, the allowance of minutes included in the cellular plan and the overage price for consumption of minutes in excess of the included plan allowance based on the parameter estimates of the model (Table 4.3).

[101] The elasticities are obtained by averaging over 50 simulations of the demand shock and all consumers.

	Plan-Choice Elasticity	**Usage Elasticity**
Monthly Fee (F)		
Plan 1	-0.258	-
Plan 2	-4.939	-
Plan 3	-0.306	-
Plan 4	-1.942	-
Overage Price (p)		
Plan 1	-0.844	-
Plan 2	-6.041	-
Plan 3	-0.251	-
Plan 4	-1.739	-
Overall		-0.005
Allowance (A)		
Plan 1	0.266	0.005
Plan 2	5.113	0.006
Plan 3	0.122	0.007
Plan 4	1.828	0.013

The elasticities represent averages across all consumers of the percentage changes in simulated own plan-choice probabilities and simulated usage corresponding to a 1% increase in the respective pricing parameter of each of the four offered cellular plans.

Table 4.4: Plan-Choice and Usage Elasticities

Plan-choice elasticities with respect to the monthly fixed fee amount to -0.258 for Plan 1 with the smallest allowance (200 minutes), -4.939 for Plan 2, -0.306 for cellular Plan 3, and -1.942 for cellular Plan 4 with the largest allowance of minutes included. In other words, a 1% increase of 30 cents in the fixed monthly fee for Plan 1 from \$30 to \$30.3 leads to a decrease in the (average) plan-choice probability of around 0.2% for Plan 1. To put this plan-choice elasticity in perspective, a 1% increase in the monthly fee of cellular Plan 1 implies a 6% reduction in the monthly fee differentials between Plan 1 and Plan 2 with the next larger allowance. The largest plan-choice elasticity is estimated for Plan 2, where a 1% increase in the fixed monthly fee leads to

a decrease in the (average) plan-choice probability of 4.9% for Plan 2. The large response for Plan 2 can be explained by the fact that Plan 2 is tightly wedged between plans 1 and 3 and consumers can respond to an increase in the monthly fee of Plan 2 by either downgrading to Plan 1 with a lower allowance or upgrading to Plan 3 with a larger allowance. In general tough, the plan-choice elasticities are relatively small in relation to the reduction in monthly fee differential between plans. The meager consumer response occurs because of subscribers' strong preference for the status quo, i.e. the current plan.

Plan-choice elasticities with respect to the overage price for consumption in excess of the plan allowance are -0.844 for Plan 1, -6.041 for Plan 2, -0.251 for Plan 3, and -1.739 for cellular Plan 4. In words, a 1% increase in the overage price from 40 cents per minute to 40.4 cents per minute leads to a reduction in the (average) plan-choice probability of around 1.7% for cellular Plan 4. Similar to the plan-choice elasticities estimated with respect to the monthly fee, the elasticity with respect to the overage price is largest for Plan 2, presumably due to closeness of alternative options in plans 1 and 3.

In addition to the analysis of two-part pricing, the analysis of cellular plans with a three-part tariff structure provides additional results on plan-choice elasticities with respect to the allowance of minutes that is included with the plan. The pattern of plan-choice elasticities with respect to the allowance is similar to the plan-choice elasticities with respect to the monthly fee and the overage price, but with opposite sign. The largest plan-choice elasticity with respect to an 1% increase in the plan allowance is estimated for Plan 2 where such a pricing change leads to an an increase in the plan-

choice probability of around 5.1%. To put this in perspective, a 1% increase in the allowance for Plan 2 is equivalent to an 6% reduction (increase) in the allowance difference to Plan 3 (Plan 1).

In terms of usage elasticities with respect to the plan allowance, the simulation results show that the usage response is increasing from Plan 1 to Plan 4, although it is minuscule in general. A 1% increase in the allowance of Plan 4 from 500 to 505 minutes raises usage by 0.013% or 0.043 minutes on average for consumers subscribing to Plan 4. Similarly, the usage elasticity with respect to the overage price (across all tariffs) is -0.005. This implies that a 1% increase in the overage price from 40 cents to 40.4 cents across all cellular plans leads to a decrease in usage of 0.005% or 0.009 minutes on average across all consumers.

All the policy simulation results are contingent on the particular price structure chosen by the cellular service provider. They also seem to be influenced quite strongly by the special status of Plan 2 within the set of four plans offered. As Figure 4.1 illustrates, Plan 2 was not available during the sign-up period for new subscribers. Plan 2 was only offered at a later time for consumers already subscribing to one of the other three offered plans. Hence, there are relatively few subscribers to Plan 2 in the selected sample and all of them must have switched cellular plans at least once during the data period. While controlling for the temporal availability of cellular plans ensures proper specification of the consumer choice set, in this particular case it also leads to parameter estimates for Plan 2 that are difficult to compare with the other three plans offered since the sample selection of subscribers is quite different.

In general, the estimated usage elasticities with respect to the allowance and the overage price are small and stand in stark contrast to the fairly significant plan-choice elasticities with respect to the allowance and the overage price parameters. The usage elasticity with respect to the overage price is estimated at -0.005, substantially below elasticities found on two-part tariff pricing of local telephone service.[102] The usage elasticity with respect to the allowance estimated only marginally higher.

There are multiple explanations for this result: First, the parameter β that captures the demand response to variation in the marginal price is identified by the usage choices of consumer with similar characteristics under different levels of the marginal price. However, given the plan pricing structure used by the cellular service provider, only two levels of the marginal price are observed, zero for consumption within the allowance and 40 cents for consumption exceeding the plan allowance. Consequently, the marginal price variation might be insufficient to properly identify the parameter β. Second, as illustrated in Figure 4.2 (Section 4.2), consumers usage seldomly exceeds the plan allowance, most usage is far below the allowance. Consequently, the allowance and the overage price are less likely to influence the ex-post usage decision. The linear demand specification coupled with a single (monthly) usage decision implies that raising the allowance or increasing the overage price can only affect the less than 20% of consumers that exceed their monthly plan allowance in a particular month. Chapter 2 of

[102] Typical estimates of usage elasticity with respect to marginal prices for local telephone service under two-part pricing range from -0.10 to -0.75 (Train, McFadden, and Ben-Akiva, 1987; Hobson and Spady, 1988; Kling and Van Der Ploeg, 1990), but can go as high as -1.70 to -2.50 (Narayanan, Chintagunta, and Miravete, 2007).

this thesis presented a theoretical model that takes into account the sequential decision-making process on usage throughout the billing period. The discussion in Section 2.5 emphasized the point that empirical models with a one-shot (monthly) usage decision are likely to underestimate the usage elasticities with respect to the plan allowance and the overage price. Incorporation within-period consumption patterns however requires information on the timing of cellular consumption over the billing period, information that was not available in the study dataset.

Despite the fact that the policy simulation results are contingent on the particular price structure chosen by the cellular service provider, the analysis of plan-choice and usage elasticities and how they relate to the three-part tariff structure often employed for cellular pricing can provide more general insights into consumer behavior under three-part tariffs. Previous telecommunications studies with two-part tariffs that consist of a fixed fee and a marginal price (but no allowance of minutes included) have shown that the pricing structure affects both plan-choice and usage (Train, McFadden, and Ben-Akiva, 1987; Train, Ben-Akiva, and Atherton, 1989). From the presented plan-choice and usage elasticities in this study, it becomes apparent that the allowance and usage price parameters of the three-part pricing structure primarily affect consumers' plan choice and much less so usage.

4.4.3 The Effect of Demand Uncertainty under Three-Part Tariffs

Three-part pricing is intrinsically linked to the uncertainty of consumer demand. The parameter estimates in Table 4.3 show that demand uncertainty

is substantial and it is of interest to further investigate the role of usage uncertainty in plan-choice behavior. Table 4.5 presents the results of a counterfactual experiment that simulates consumer plan choice by raising the standard deviation of the demand shock σ_υ by 1%:[103]

	Plan-Choice/Revenue Elasticity
Demand Uncertainty (σ_υ)	
Plan 1	-0.613
Plan 2	1.405
Plan 3	0.105
Plan 4	2.570
Firm Revenue	-0.002

The elasticities represent averages across all consumers of the percentage change in simulated plan-choice probabilities corresponding to a 1% increase of σ_υ.

Table 4.5: The Effect of Demand Uncertainty

The estimated plan-choice elasticities represent the net effect of consumers switching among cellular plans in response to higher demand uncertainty. The simulation results confirm the notion that higher demand uncertainty increases the likelihood that consumers choose a plan with a larger allowance. The plan-choice elasticity with respect to an increase in usage uncertainty of Plan 1 with the lowest allowance is negative while the three plans with larger allowances have positive estimated plan-choice elasticities with respect to demand uncertainty. An increase of the standard deviation of usage uncertainty by 1% leads to an increase in the plan-choice probability of 2.57% for the Plan 4 with the largest allowance. Hence, due

[103]Due to the exponential specification of the demand shock σ_υ, an increase in the standard deviation of the demand shock increases the expected mean usage. In order to focus exclusively on the effect of demand uncertainty, the plan-choice probabilities in Table 4.5 are simulated by changing the standard deviation of the demand shock σ_υ holding expected usage constant.

to the convex pricing structure, an increase in demand uncertainty leads consumers to respond by switching to cellular plans that include a larger monthly allowance of minutes.

While consumers respond to increased demand uncertainty by switching to plans with larger allowances and larger fixed monthly fees, whether firm revenue increases in step depends on the relation of lost overage charges from subscribers on plans with smaller allowances relative to the increased revenue from higher monthly fees from consumers subscribing to plans with larger allowances.

The results in Section 4.4 indicate that in general, consumers tend to prefer plans with smaller allowances as indicated by the plan-specific dummy variables. Given the small allowances, they tend to incur disproportionately large overage charges. Indeed, the estimated elasticity of the cellular provider's expected revenue with respect to an increase in the standard deviation of the demand shock σ_v is negative, -0.002. In essence, increased demand uncertainty leads subscribers to upgrade their plan to a plan with a larger allowance, and the reduced overage revenue for the firm is not completely compensated by the higher monthly fees.

This result underscores that plan-specific preferences are not only of interest to the researcher trying to understand plan-choice and consumer behavior, but it is of critical importance to cellular companies in helping them design an optimal menu and structure of pricing plans to offer.

4.5 Concluding Remarks

In many industries, companies segment consumers by offering them a menu of plan options to choose from. There has been an explosion of pricing plans in the wireless telecommunications industry, many of them are a form of three-part pricing scheme that includes a fixed monthly fee, an allowance of minutes included with the plan and an overage price for consumption of cellular services in excess of the monthly allowance. While there exists an extensive literature on consumer behavior under two-part tariffs, there has been little exploration of choice behavior in the presence of three-part tariffs. Consumer behavior can substantially differ between two-part and three-part tariff structures in part due to the effects of demand uncertainty.

This study contributes to the empirical literature on consumer behavior under three-part tariffs by estimating a detailed discrete/continuous demand model of cellular plan choice and consumption using a large consumer-level dataset from a major US cellular service provider. This data is combined with information on plan availability to account for temporal variation in consumers' choice sets resulting from introduction or discontinuance of cellular plans and the practice of grandfathering cellular plans for current plan subscribers.

The empirical findings of the model identify substantial and significant tariff-specific biases in consumer behavior. However, the identified consumer bias favors plans with smaller allowances and stands in contrast to the "flat-rate" bias previous studies in two-part tariff environments have found. In addition, the results suggest that the pricing parameters of three-part tariffs

affect consumer's plan choice to a much greater extent than their monthly consumption. Demand uncertainty is substantial among cellular consumers and is identified as of the main driving forces of plan-choice behavior in three-part tariff environments. Demand uncertainty and plan-specific preferences not only affect consumer behavior and plan-choice, it also has important implications for the tariff design by cellular companies: Consumers' demand uncertainty affects the profitability of certain cellular pricing schemes such as three-part tariffs studied in Chapter 2. But furthermore, firms could employ deceptive pricing strategies and design menus of cellular plans that specifically aim to profit from consumers' biases and plan-specific preferences.

While all the results in this study are contingent on the particular cellular plans offered by the large US cellular carrier, this empirical study is an attempt to gain a better understanding of consumer behavior (and firm pricing) in environments that are subject to large demand uncertainty. In such environments, one is likely to observe other innovative pricing strategies such a rollover pricing where consumers are allowed to carry over unused portion of their allowance into the next month or automatic plan adjustment where consumers are billed based on the ex-post least costly plan. Cellular firms try out (and cancel) new pricing strategies as they attempt to optimize their tariff structure and gain a competitive advantage in the market place. This is a great opportunity for future research to further explore the interaction of consumer behavior and innovative pricing strategies in the cellular telecommunications industry.

Bibliography

Armstrong, M. (1998): "Network Interconnection in Telecommunications," *Economic Journal*, 108(448), 545–564.

——— (2002): "The Theory of Access Pricing and Interconnection," in *Handbook of Telecommunications Economics*, vol. 1, pp. 295–384, North-Holland. Elsevier Science.

——— (2006): "Price Discrimination," MPRA Working Paper.

Armstrong, M., and J. Wright (2007): "Mobile Call Termination," Available at SSRN: http://ssrn.com/abstract=1014322.

Baron, D. P., and D. Besanko (1984): "Regulation and Information in a Continuing Relationship," *Information Economics and Policy*, 1(3), 267–302.

Behringer, S. (2006): "Asymmetric Equilibria and Non-Cooperative Access Pricing in Telecommunications," mimeo, Universität Frankfurt.

Berger, U. (2004): "Access Charges in the Presence of Call Externalities," *Contributions to Economic Analysis & Policy*, 3(1), 1–16.

——— (2005): "Bill-and-Keep vs. Cost-based Access Pricing Revisited," *Economics Letters*, 86(1), 107–112.

Birke, D., and P. G. Swann (2005): "Social Networks and Choice of Mobile Phone Operator," Nottingham University Business School, Industrial Economics Division Occasional Paper Series, No. 2005-14.

——— (2006): "Network Effects and the Choice of Mobile Phone Operator," *Journal of Evolutionary Economics*, 16(1), 65–84.

Buehler, S., R. Dewenter, and J. Haucap (2006): "Mobile Number Portability in Europe," *Telecommunications Policy*, 30 (7), 385–399.

Calzada, J., and T. M. Valletti (2005): "Network Competition and Entry Deterrence," CEPR Discussion Paper No. 5381.

Carter, M., and J. Wright (1999): "Interconnection in Network Industries," *Review of Industrial Organization*, 14(1), 1–25.

——— (2003): "Asymmetric Network Interconnection," *Review of Industrial Organization*, 22(1), 27–46.

Cherdron, M. (2001): "Interconnection, Termination-Based Price Discrimination and Network Competition in a Mature Telecommunications Market," Working Paper.

Competition Commission (2003): *Vodafone, O2, Orange and T-Mobile*. The Stationary Office Books.

Courty, P., and H. Li (2000): "Sequential Screening," *Review of Economic Studies*, 67(4), 697–717.

Crandall, R. W., and G. J. Sidak (2004): "Should Regulators Set Rates to Terminate Calls on Mobile Networks," *Yale Journal on Regulation*, 21(2), 261–314.

CWTA (2008): "Wireless Phone Subscribers in Canada," http://www.cwta.ca/CWTASite/english/facts_figures_downloads/SubscribersStats_en_2008_Q4.pdf.

DellaVigna, S., and U. Malmendier (2004): "Contract Design and Self-Control: Theory and Evidence," *Quarterly Journal of Economics*, 119(2), 353–402.

Dessein, W. (2004): "Network Competition with Heterogeneous Customers and Calling Patterns," *Infromation Economics and Policy*, 16(3), 323–345.

Economides, N., K. Seim, and B. V. Viard (2008): "Quantifying the Benefits of Entry into Local Phone Service," *RAND Journal of Economics*, 39(3), 699–730.

Ellis, R. P. (1986): "Rational Behavior in the Presence of Coverage Ceilings and Deductibles," *RAND Journal of Economics*, 17(2), 158–175.

Farrell, J., and P. Klemperer (2007): "Coordination and Lock-In: Competition with Switching Costs and Network Effects," in *Handbook of Industrial Organization, Volume 3*, ed. by M. Armstrong, and R. Porter. North-Holland, Amsterdam.

Gabaix, X., and D. Laibson (2006): "Shrouded Attributes, Consumer

Myopia, and Information Suppression in Competitive Markets," *Quarterly Journal of Economics*, 121(2), 505–540.

Gabrielsen, T. S., and S. Vagstad (2007): "Why Is On-Net Traffic Cheaper Than Off-Net Traffic? Access Markup as a Collusive Device," *European Economic Review*, forthcoming.

Gans, J. S., and S. P. King (2000): "Mobile Network Competition, Customer Ignorance and Fixed-to-Mobile Call Prices," *Information Economics and Policy*, 12 (4), 301–327.

——— (2001): "Using 'Bill and Keep' Interconnect Arrangements to Soften Network Competition," *Economics Letters*, 71(3), 413–420.

Grilo, I., O. Shy, and J.-F. Thisse (2001): "Price Competition When Consumer Behavior Is Characterized by Conformity or Vanity," *Journal of Public Economics*, 80(3), 385–408.

Grubb, M. D. (2007): "Selling to Overconfident Consumers," Working Paper.

Guo, L., and T. Erdem (2006): "Measuring Usage Flexibility in Wireless Tariff Choice," Working Paper.

Hahn, J.-H. (2004): "Network Competition and Interconnection with Heterogeneous Subscribers," *International Journal of Industrial Organization*, 22(5), 611–631.

Hanemann, M. W. (1984): "Discrete/Continuous Models of Consumer Demand," *Econometrica*, 52(3), 541–562.

Harbord, D., and M. Pagnozzi (2008): "On-Net/Off-Net Price Discrimination and Bill-and-Keep vs. Cost-Based Regulation of Mobile Termination Rates," Working Paper.

Harris, M., and R. M. Townsend (1981): "Resource Allocation Under Asymmetric Information," *Econometrica*, 49(1), 33–64.

Hausman, J. A. (1985): "The Econometrics of Nonlinear Budget Sets," *Econometrics*, 53(6), 1255–1282.

Hobson, M., and R. H. Spady (1988): "The Demand for Local Telephone Service Under Optimal Local Measured Service," Bellcore Discussion Paper No. 50.

Hoernig, S. (2007): "On-net and Off-net Pricing on Asymmetric Telecommunications Networks," *Information Economics and Policy*, 19(2), 171–188.

Huang, C. I. (2008): "Estimating Demand for Cellular Phone Service under Nonlinear Pricing," *Quantitative Marketing and Economics*, 6(4), 371–413.

Iyengar, R. (2004): "A Structural Demand Analysis for Wireless Services under Nonlinear Pricing Schemes," Working Paper.

Jensen, S. (2006): "Implementation of Competitive Nonlinear Pricing: Tariffs With Inclusive Consumption," *Review of Economic Design*, 10(1), 9–29.

Jeon, D.-s., J.-J. Laffont, and J. Tirole (2004): "On the 'Receiver-Pays' Principle," *RAND Journal of Economics*, 35(1), 85–110.

Keeler, E. B., J. P. Newhouse, and C. E. Phelps (1977): "Deductibles and the Demand for Medical Care Services: The Theory of a Consumer Facing a Variable Price Schedule Under Uncertainty," *Econometrica*, 45(3), 641–655.

Kling, J. P., and S. S. Van Der Ploeg (1990): "Estimating Local Call Elasticities with a Model of Stochastic Class of Service and Usage Choice," in *Telecommunications Demand Modelling: An Integrated View*, ed. by A. de Fontenay, M. H. Shugard, and D. S. Sibley, vol. 187 of *Contributions to Economic Analysis Series*, pp. 119–136. North Holland, Amsterdam.

Kridel, D. J., D. E. Lehman, and D. L. Weisman (1993): "Option Value, Telecommunication Demand, and Policy," *Information Economics and Policy*, 5(2), 125–144.

Laffont, J.-J., and D. Martimort (2002): *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, Princeton.

Laffont, J.-J., P. Rey, and J. Tirole (1998a): "Network Competition: I. Overview and Nondiscriminatory Pricing," *RAND Journal of Economics*, 29(1), 1–37.

Laffont, J. J., P. Rey, and J. Tirole (1998b): "Network Competition: II. Price Discrimination," *RAND Journal of Economics*, 29(1), 38–56.

Laffont, J.-J., and J. Tirole (1996): "Pollution Permits and Compliance Strategies," *Journal of Public Economics Letters*, 62(1-2), 85–125.

——— (2000): *Competition in Telecommunications*, Munich Lectures in Economics. MIT Press, Cambridge and London.

Lambrecht, A., K. Seim, and B. Skiera (2007): "Does Uncertainty Matter? Consumer Behavior Under Three-Part Tariffs," *Marketing Science*, 26(5), 698–710, Working Paper.

Lambrecht, A., and B. Skiera (2006): "Paying Too Much and Being Happy About It: Causes and Consequences of Tariff Choice-Biases," *Journal of Marketing Research*, 43(2), 212–223, Working Paper.

Maskin, E., and J. Riley (1984): "Monopoly with Incomplete Information," *RAND Journal of Economics*, 15(2), 171–196.

Miravete, E. J. (2002): "Estimating Demand for Local Telephone Service with Asymmetric Information and Optional Calling Plans," *Review of Economic Studies*, 69(4), 943–971.

——— (2003): "Choosing the Wrong Calling Plan? Ignorance and Learning," *American Economic Review*, 93(1), 297–310.

——— (2005): "The Welfare Performance of Sequential Pricing Mechanisms," *International Economic Review*, 46(4), 1321–1360.

——— (2009): "Competing with Menus of Tariff Options," *Journal of the European Economic Association*, 7, 188–205.

Mitchell, B. M., and I. Vogelsang (1991): *Telecommunication Pricing: Theory and Practice*. Cambridge University Press, New York.

Myerson, R. B. (1979): "Incentive Compadibility and the Bargaining Problem," *Econometrica*, 47(1), 61–73.

——— (1986): "Multistage Games with Communication," *Econometrica*, 54(2), 323–358.

Narayanan, S., P. K. Chintagunta, and E. J. Miravete (2007): "The Role of Self-Selection and Usage Uncertainty in the Demand for Local Telephone Service," *Quantitative Marketing and Economics*, 5(1), 1–34.

Nunes, J. C. (2000): "A Cognitive Model of People's Usage Estimation," *Journal of Marketing Research*, 37(4), 397–409.

OECD (2007): "OECD Communicatons Outlook 2007," Discussion paper, Organisation fro Economic Co-Operation and Development (OECD).

Ofcom (2006): "The Communications Market 2006," Discussion paper, Office of Communications, London.

Oftel (2001): "Review of the Charge Control on Calls to Mobiles," Discussion paper, Office of Telecommunications (UK).

Olmstead, S. M., M. Hanemann, and R. N. Stavins (2006): "Do Consumers React to the Shape of Supply? Water Demand Under Heterogeneous Price Structures," Working Paper.

Shi, M. (2003): "Social Network-Based Discriminatory Pricing Strategy," *Marketing Letters*, 14(4), 239–256.

Stennek, J., and T. P. Tangerås (2006): "Competition vs. Regulation in Mobile Telecommunications," RIIE Working Paper.

Tirole, J. (1988): *The Theory of Industrial Organization*. MIT Press, Cambridge, Massachusetts.

Train, K. E., M. Ben-Akiva, and T. Atherton (1989): "Consumption Patterns and Self-selecting Tariffs," *Review of Economic and Statistics*, 71(1), 62–73.

Train, K. E., D. L. McFadden, and M. Ben-Akiva (1987): "The Demand for Local Telephone Service: A Fully Discrete Model of Residential Calling Patterns and Service Choices," *RAND Journal of Economics*, 18(1), 109–123.

Wilson, R. (1993): *Nonlinear Pricing*. Oxford University Press, Oxford.

Woroch, G. A., F. R. Warren-Boulton, and K. C. Baseman (1998): "Exclusionary Behavior in the Market for Operating System Software: The Case of Microsoft," in *Opening Networks to Competition: The Regulation and Pricing of Access*, ed. by D. Gabel, and D. F. Weiman, Topics in Regulatory Economics and Policy Series, pp. 221–238. Kluwer Academic, Boston.

Wright, J. (2002): "Access Pricing Under Competition: An Application to Cellular Networks," *Journal of Industrial Economics*, 50 (3), 289–315.

Appendix A

Supplementary Material for Chapter 2

A.1 Proof of Lemma 2.1

(i) The threshold is

$$\widehat{p}_i = p_1 + \int_{p_1}^{\overline{v}} (v - p_1)\, g_i(v)\, dv - \int_{p_2}^{\overline{v}} (v - p_2)\, g_i(v)\, dv \quad \text{for } i = \theta_L, \theta_H.$$

If $p_1 > p_2$, then

$$\int_{p_1}^{\overline{v}} (v - p_1)\, g_i(v)\, dv < \int_{p_2}^{\overline{v}} (v - p_2)\, g_i(v)\, dv$$

and $\widehat{p}_i < p_1$. Similarly, $p_1 < p_2$ implies that $\widehat{p}_i > p_1$.

(ii) We have

$$\begin{aligned}\widehat{p}_{\theta_H} - \widehat{p}_{\theta_L} =& \int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_H}(v)\, dv - \int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_H}(v)\, dv \\ & - \int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_L}(v)\, dv + \int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_L}(v)\, dv \\ =& \int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_H}(v)\, dv - (1 - G_{\theta_H}(p_1))\, p_1 \\ & - \int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_H}(v)\, dv - \int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_L}(v)\, dv \\ & + \int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_L}(v)\, dv\end{aligned}$$

and first-order stochastic dominance implies that

$$\int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_H}(v)\, dv \geq \int_{p_1}^{\overline{v}} (v - p_1)\, g_{\theta_L}(v)\, dv$$

$$\int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_H}(v)\, dv \geq \int_{p_2}^{\overline{v}} (v - p_2)\, g_{\theta_L}(v)\, dv.$$

For $p_1 > p_2$, we have from part (i)

$$\int_{p_1}^{\overline{v}} (v - p_1)\, g_i(v)\, dv < \int_{p_2}^{\overline{v}} (v - p_2)\, g_i(v)\, dv \quad \text{for } i = \theta_L, \theta_H$$

and these four inequalities give $\widehat{p}_{\theta_H} - \widehat{p}_{\theta_L} > 0$. Similarly, for $p_1 < p_2$ we have $\widehat{p}_{\theta_H} - \widehat{p}_{\theta_L} < 0$.

A.2 Proof of Lemma 2.2

The individual rationality constraints can be rewritten as

$$\begin{aligned}
&-F_i - p_{1i}\left(1 - G_i\left(\widehat{p}_i\right) G_i\left(p_{1i}\right)\right) - p_{2i}\left(1 - G_i\left(\widehat{p}_i\right)\right)\left(1 - G_i\left(p_{2i}\right)\right) \\
&+\int_{\widehat{p}_i}^{\overline{v}} v dG_i\left(v\right) + G_i\left(\widehat{p}_i\right)\int_{p_{1i}}^{\overline{v}} v dG_i\left(v\right) + \left(1 - G_i\left(\widehat{p}_i\right)\right)\int_{p_i^2}^{\overline{v}} v dG_i\left(v\right) \quad \geqq \quad 0
\end{aligned}$$

or

$$\begin{aligned}
&-F_i - p_{1i} - p_{2i}\left(1 - G_i\left(\widehat{p}_i\right)\right) + \int_{\widehat{p}_i}^{\overline{v}} v dG_i\left(v\right) \\
&+G_i\left(\widehat{p}_i\right)\int_{\underline{v}}^{\overline{v}} \max\left\{p_{1i}, v\right\} dG_i\left(v\right) + \left(1 - G_i\left(\widehat{p}_i\right)\right)\int_{\underline{v}}^{\overline{v}} \max\left\{p_{2i}, v\right\} dG_i\left(v\right) \quad \geqq \quad 0
\end{aligned}$$

for $\forall i = \theta_L, \theta_H$. IC_{θ_H,θ_L} implies

$$
\begin{aligned}
& - F_{\theta_H} - p_{1\theta_H} - p_{2\theta_H}\left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \\
& + \int_{\widehat{p}_{\theta_H}}^{\overline{v}} v dG_{\theta_H}(v) + G_{\theta_H}\left(\widehat{p}_{\theta_H}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_H}, v\right\} dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_H}, v\right\} dG_{\theta_H}(v) \\
\geqq \quad & - F_{\theta_L} - p_{1\theta_L} - p_{2\theta_L}\left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right) \\
& + \int_{\widehat{p}_{\theta_H|\theta_L}}^{\overline{v}} v dG_{\theta_H}(v) + G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_L}, v\right\} dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_L}, v\right\} dG_{\theta_H}(v)
\end{aligned}
$$

Type θ_H's optimal cutoff in consumption subperiod 1 on the type-θ_L contract, $\widehat{p}_{\theta_H|\theta_L}$ must be weakly preferred by type θ_H to the cutoff $\widehat{p}'$ where $\widehat{p}'$ is defined by $G_{\theta_H}\left(\widehat{p}'\right) = G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)$. This cutoff implies that the probability of consumption in the first subperiod of the two types is the same. Therefore,

we must have

$$\begin{aligned}
& -F_{\theta_H} - p_{1\theta_H} - p_{2\theta_H}\left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \\
& + \int_{\widehat{p}_{\theta_H}}^{\overline{v}} v dG_{\theta_H}(v) + G_{\theta_H}\left(\widehat{p}_{\theta_H}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_H}, v\right\} dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_H}, v\right\} dG_{\theta_H}(v) \\
\geqq \quad & -F_{\theta_L} - p_{1\theta_L} - p_{2\theta_L}\left(1 - G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right) \\
& + \int_{\widehat{p}'}^{\overline{v}} v dG_{\theta_H}(v) + G_{\theta_L}\left(\widehat{p}_{\theta_L}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_L}, v\right\} dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_L}, v\right\} dG_{\theta_H}(v).
\end{aligned}$$

Since the function $\max\{p, v\}$ is increasing in v and G_{θ_H} first-order stochastically dominates G_{θ_L} we have

$$\begin{aligned}
\int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_L}, v\right\} dG_{\theta_H}(v) &\geqq \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_L}, v\right\} dG_{\theta_L}(v) \\
\int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_L}, v\right\} dG_{\theta_H}(v) &\geqq \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_L}, v\right\} dG_{\theta_L}(v)
\end{aligned}$$

and

$$\int_{\widehat{p}'}^{\overline{v}} v dG_{\theta_H}(v) \geq \int_{\widehat{p}_{\theta_L}}^{\overline{v}} v dG_{\theta_L}(v).$$

It follows that

$$\begin{aligned}
& - F_{\theta_H} - p_{1\theta_H} - p_{2\theta_H}\left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \\
& + \int_{\widehat{p}_{\theta_H}}^{\overline{v}} v dG_{\theta_H}(v) + G_{\theta_H}\left(\widehat{p}_{\theta_H}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_H}, v\right\} dG_{\theta_H}(v) \\
& + \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_H}, v\right\} dG_{\theta_H}(v) \\
\geqq \quad & - F_{\theta_L} - p_{1\theta_L} - p_{2\theta_L}\left(1 - G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right) \\
& + \int_{\widehat{p}_{\theta_L}}^{\overline{v}} v dG_{\theta_L}(v) + G_{\theta_L}\left(\widehat{p}_{\theta_L}\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{1\theta_L}, v\right\} dG_{\theta_L}(v) \\
& + \left(1 - G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right) \int_{\underline{v}}^{\overline{v}} \max\left\{p_{2\theta_L}, v\right\} dG_{\theta_L}(v) \\
\geqq \quad & 0
\end{aligned}$$

where the last inequality follows from IR_{θ_L}.

A.3 Proof of Proposition 2.1

It is sufficient to show that the ignored incentive compatibility constraint (IC_{θ_L,θ_H}) is satisfied given the profit-maximizing screening contract $\left\{p^*_{\theta_L}(\cdot), p^*_{\theta_H}(\cdot)\right\}$. Since IR_{θ_L} binds, IC_{θ_L,θ_H} holds as long as

$$2\int_c^{\bar{v}} (v-c)\, dG_{\theta_L}(v) - F_{\theta_H} \leq 0.$$

Since IC_{θ_H,θ_L} also binds at the optimal solution, this implies that IC_{θ_L,θ_H} holds as long as

$$R_{\theta_H}\left(p^*_{\theta_L}\right) - 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_H}(v) + 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_L}(v) \leq 0$$

where $R_{\theta_H}\left(p^*_{\theta_L}\right)$ is the information rent for the high-type agent. Let $R_{\theta_H}(c)$ be the information rent to the high-type agent when the contract for the low-type agent has marginal price equal to marginal cost c, that is $p_{\theta_L} = p_{1\theta_L} = p_{2\theta_L} = c$, and note that

$$R_{\theta_H}(c) - 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_H}(v) + 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_L}(v) = 0.$$

The profit-maximizing screening contract for the low-type agent is

$$p^*_{\theta_L}(p_{1\theta_L}, p_{2\theta_L}) = \arg\max_{p(\cdot)} \left[\lambda_{\theta_L} S_{\theta_L}(p_{\theta_L}) - \lambda_{\theta_H} R_{\theta_H}(p_{\theta_L})\right].$$

The surplus extracted from the low-type agent θ_L is maximized when $S_{\theta_L}(p_{\theta_L}) = S_{\theta_L}(c)$. Hence the profit-maximizing screening contract has $R_{\theta_H}\left(p^*_{\theta_L}\right) \leq$

$R_{\theta_H}(c)$ and IC_{θ_L,θ_H} holds:

$$R_{\theta_H}\left(p^*_{\theta_L}\right) - 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_H}(v) + 2\int_c^{\bar{v}} (v-c)\, dG_{\theta_L}(v) \leq 0.$$

A.4 Proof of Proposition 2.2

The "relaxed" maximization problem for the profit-maximizing two-part tariff is

$$\begin{aligned}\max_{p_{\theta_L},p_{\theta_H}} \quad & 2\lambda_{\theta_L}\int_{p_{\theta_L}}^{\overline{v}}(v-c)\,dG_{\theta_L}(v)\\ & -2\lambda_{\theta_H}\left(\int_{p_{\theta_L}}^{\overline{v}}(v-p_{\theta_L})\,dG_{\theta_H}(v)-\int_{p_{\theta_L}}^{\overline{v}}(v-p_{\theta_L})\,dG_{\theta_L}(v)\right)\\ & +2\lambda_{\theta_H}\int_{p_{\theta_H}}^{\overline{v}}(v-c)\,dG_{\theta_H}(v).\end{aligned}$$

Since p_{θ_H} is unrestricted, it is optimally set at $p_{\theta_H}=c$ and the first-order condition is

$$\frac{d\Pi}{dp_{\theta_L}}=-2\lambda_{\theta_L}(p_{\theta_L}-c)\,g_L(p_{\theta_L})-2\lambda_{\theta_H}\left[G_{\theta_H}(p_{\theta_L})-G_{\theta_L}(p_{\theta_L})\right]=0.$$

and the profit-maximizing two-part tariff is characterized by a markup equal to

$$p_{\theta_L}^{TPT}-c=\frac{\lambda_{\theta_H}\left[G_{\theta_L}\left(p_{\theta_L}^{TPT}\right)-G_{\theta_H}\left(p_{\theta_L}^{TPT}\right)\right]}{\lambda_{\theta_L}g_L\left(p_{\theta_L}^{TPT}\right)}.$$

The derivatives of the optimal threshold for type θ_i are $\frac{d\widehat{p}_{\theta_i}}{dp_{1\theta_i}}=G_{\theta_i}(p_{1\theta_i})$ and $\frac{d\widehat{p}_{\theta_i}}{dp_{2\theta_i}}=1-G_{\theta_i}(p_{2\theta_i})$. Hence, the derivatives of the surplus extracted

from type θ_L with respect to $p_{1\theta_L}$ and $p_{2\theta_L}$ are

$$\frac{dS_{\theta_L}(\cdot)}{dp_{1\theta_L}} = -\frac{d\widehat{p}_{\theta_L}}{dp_{1\theta_L}} g_{\theta_L}(\widehat{p}_{\theta_L}) \times \left((\widehat{p}_{\theta_L} - c) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L} - \int_{p_{2\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L} \right) \right)$$
$$- G_{\theta_L}(\widehat{p}_{\theta_L})(p_{1\theta_L} - c)\, g_{\theta_L}(p_{1\theta_L})$$

$$\frac{dS_{\theta_L}(\cdot)}{dp_{2\theta_L}} = -\frac{d\widehat{p}_{\theta_L}}{dp_{2\theta_L}} g_{\theta_L}(\widehat{p}_{\theta_L}) \times \left((\widehat{p}_{\theta_L} - c) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L} - \int_{p_{2\theta_L}}^{\overline{v}} (v-c)\, dG_{\theta_L} \right) \right)$$
$$- (1 - G_{\theta_L})(\widehat{p}_{\theta_L})(p_{2\theta_L} - c)\, g_{\theta_L}(p_{2\theta_L})$$

and evaluated at $p_{1\theta_L} = p_{2\theta_L} = p_{\theta_L}$

$$\frac{dS_{\theta_L}(\cdot)}{dp_{1\theta_L}} = -2G_{\theta_L}(p_{\theta_L})\, g_{\theta_L}(p_{\theta_L})(p_{\theta_L} - c)$$

$$\frac{dS_{\theta_L}(\cdot)}{dp_{2\theta_L}} = -2(1 - G_{\theta_L}(p_{\theta_L}))\, g_{\theta_L}(p_{\theta_L})(p_{\theta_L} - c).$$

The derivatives of the information rent that the principal has to leave to type $R_{\theta_H}(p_{\theta_L})$ are

$$\begin{aligned}
\frac{dR_{\theta_H}(p_{\theta_L})}{dp_{1\theta_L}} =& \frac{d\widehat{p}_{\theta_H|\theta_L}}{dp_{1\theta_L}} g_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right) \\
&\times \left(\left(\widehat{p}_{\theta_H|\theta_L} - p_{1\theta_L}\right) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_H} - \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_H} \right) \right) \\
&- \frac{d\widehat{p}_{\theta_L}}{dp_{1\theta_L}} g_{\theta_L}\left(\widehat{p}_{\theta_L}\right) \\
&\times \left(\left(\widehat{p}_{\theta_L} - p_{1\theta_L}\right) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v - p_{1\theta_L})\, dG_{\theta_L} - \int_{p_{2\theta_L}}^{\overline{v}} (v - p_{2\theta_L})\, dG_{\theta_L} \right) \right) \\
&- \left(1 + G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right)\left(1 - G_{\theta_H}\left(p_{1\theta_L}\right)\right) \\
&+ \left(1 + G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right)\left(1 - G_{\theta_L}\left(p_{1\theta_L}\right)\right)
\end{aligned}$$

$$\begin{aligned}
\frac{dR_{\theta_H}(p_{\theta_L})}{dp_{2\theta_L}} =& \frac{d\widehat{p}_{\theta_H|\theta_L}}{dp_{2\theta_L}} g_{\theta_L}\left(\widehat{p}_{\theta_H|\theta_L}\right) \\
&\times \left(\left(\widehat{p}_{\theta_H|\theta_L} - c\right) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v - c)\, dG_{\theta_L} - \int_{p_{2\theta_L}}^{\overline{v}} (v - c)\, dG_{\theta_L} \right) \right) \\
&- \frac{d\widehat{p}_{\theta_L}}{dp_{2\theta_L}} g_{\theta_L}\left(\widehat{p}_{\theta_L}\right) \\
&\times \left(\left(\widehat{p}_{\theta_L} - c\right) - \left(\int_{p_{1\theta_L}}^{\overline{v}} (v - c)\, dG_{\theta_L} - \int_{p_{2\theta_L}}^{\overline{v}} (v - c)\, dG_{\theta_L} \right) \right) \\
&- \left(1 - G_{\theta_H}\left(\widehat{p}_{\theta_H|\theta_L}\right)\right)\left(1 - G_{\theta_H}\left(p_{2\theta_L}\right)\right) \\
&+ \left(1 - G_{\theta_L}\left(\widehat{p}_{\theta_L}\right)\right)\left(1 - G_{\theta_L}\left(p_{2\theta_L}\right)\right)
\end{aligned}$$

and evaluated at $p_{1\theta_L} = p_{2\theta_L} = p_{\theta_L}$

$$\begin{aligned}\frac{dR_{\theta_H}(p_{\theta_L})}{dp_{1\theta_L}} =& (1 + G_{\theta_L}(p_{\theta_L}))(1 - G_{\theta_L}(p_{\theta_L})) \\ & - (1 + G_{\theta_H}(p_{\theta_L}))(1 - G_{\theta_H}(p_{\theta_L}))\end{aligned}$$

$$\frac{dR_{\theta_H}(p_{\theta_L})}{dp_{2\theta_L}} = (1 - G_{\theta_L}(p_{\theta_L}))^2 - (1 - G_{\theta_H}(p_{\theta_L}))^2.$$

The first-order conditions with respect to $p_{1\theta_L}$ and $p_{2\theta_L}$ once again evaluated at $p_{1\theta_L} = p_{2\theta_L} = p_{\theta_L}$ are

$$\begin{aligned}\left.\frac{d\Pi}{dp_{1\theta_L}}\right|_{p_{\theta_L}} =& -2\lambda_L G_{\theta_L}(p_{\theta_L})\, g_{\theta_L}(p_{\theta_L})\,(p_{\theta_L} - c) \\ & - \lambda_H (1 + G_{\theta_L}(p_{\theta_L}))(1 - G_{\theta_L}(p_{\theta_L})) \\ & - \lambda_H (1 + G_{\theta_H}(p_{\theta_L}))(1 - G_{\theta_H}(p_{\theta_L}))\end{aligned}$$

$$\begin{aligned}\left.\frac{d\Pi}{dp_{2\theta_L}}\right|_{p_{\theta_L}} =& -2\lambda_L (1 - G_{\theta_L}(p_{\theta_L}))\, g_{\theta_L}(p_{\theta_L})\,(p_{\theta_L} - c) \\ & - \lambda_H (1 - G_{\theta_L}(p_{\theta_L}))^2 - (1 - G_{\theta_H}(p_{\theta_L}))^2.\end{aligned}$$

Substituting in the markup from the profit-maximizing two-part tariff

$$p_{\theta_L}^{TPT} - c = \frac{\lambda_H (G_{\theta_L}(p) - G_{\theta_H}(p))}{\lambda_L g_{\theta_L}(p)}$$

gives

$$\left.\frac{d\Pi}{dp_{1\theta_L}}\right|_{p_{\theta_L}^{TPT}} = -\lambda_H \left(G_{\theta_L}\left(p_{\theta_L}^{TPT}\right) - G_{\theta_H}\left(p_{\theta_L}^{TPT}\right)\right)^2 < 0$$

$$\left.\frac{d\Pi}{dp_{2\theta_L}}\right|_{p_{\theta_L}^{TPT}} = \lambda_H \left(G_{\theta_L}\left(p_{\theta_L}^{TPT}\right) - G_{\theta_H}\left(p_{\theta_L}^{TPT}\right)\right)^2 > 0.$$

Appendix B

Supplementary Material for Chapter 3

B.1 Proof of Proposition 3.1

There does not exist an equilibrium in which one firm corners the market: Suppose that network 1 corners the market. It sets price at marginal cost, $p_1 = c$, and $F_1 - f \geq 0$, and $\pi_2 = 0$. But network 2 could charge $p_2 = c$, and $F_2 = F_1 + \epsilon$. For ϵ small enough, its profit would then be $\widetilde{\pi}_2 \backsimeq \frac{1}{2}(F_2 - f) \geq \frac{\epsilon}{2} > 0$, a contradiction.

Profit of network i is

$$\pi_i = \alpha_i (F_i - f) + \alpha_i (p_i - c) q(p_i) + \alpha_i (1 - \alpha_i^{\gamma}) mc (q(p_j) - q(p_i)).$$

Suppose first that networks maximize profits holding market share constant. That is, they choose a marginal price p_i while the fixed fee is adjusted to offset deviations of the market share. Holding market share constant, differentiating Eq. 3.1 with respect to p_i and using the fact that $v'(p_i) = -q(p_i)$ and $q(p_i) = p_i^{-\eta}$ gives

$$\frac{\partial F_i}{\partial p_i} = -p_i^{-\eta} - \frac{1}{2}\eta \left(\alpha_i^{\gamma} + \alpha_j^{\gamma} - 1\right) p_i^{-\eta}. \tag{B.1}$$

Maximizing the profit function with respect p_i holding market shares constant yields

$$\frac{\partial F_i}{\partial p_i} = -\left(p_i^{-\eta} - \eta\left(p_i - c - \left(1 - \alpha_i^{\gamma} mc\right)\right) p_i^{-(\eta+1)}\right) \tag{B.2}$$

and Eqs. B.1 and B.2 together imply

$$\frac{p_i - \left(1 + \left(1 - \alpha_i^{\gamma}\right) m\right) c}{p_i} = -\frac{1}{2}\left(\alpha_i^{\gamma} + \left(1 - \alpha_i\right)^{\gamma} - 1\right). \tag{B.3}$$

Hence, let $\kappa(\gamma) = \left(\frac{1}{2}\right)^{\gamma} - \frac{1}{2}$ and in the symmetric equilibrium we have

$$\frac{p^* - \left(1 + \left(1 - \left(\frac{1}{2}\right)^{\gamma}\right) m\right) c}{p^*} = -\kappa(\gamma) \quad \text{or} \quad p^* = \frac{\left(1 + \left(\frac{1}{2} - \kappa(\gamma)\right) m\right) c}{1 + \kappa(\gamma)}.$$

The marginal price is below c if

$$m < \frac{\kappa(\gamma)}{\frac{1}{2} - \kappa(\gamma)}$$

with the right-hand term decreasing in γ but always positive. Differentiating the profit function with respect to the fixed fee gives

$$\frac{\partial \pi_i}{\partial F_i} = \alpha_i + \frac{\partial \alpha_i}{\partial F_i}\left(F_i - f + \left(p_i - c\right) q\left(p_i\right)\right) + \left(1 - \left(1 + \gamma\right)\alpha_i^{\gamma}\right)\frac{\partial \alpha_i}{\partial F_i} mc\left(q\left(p_j\right) - q\left(p_i\right)\right). \tag{B.4}$$

From Eq. 3.1, $\frac{\partial \alpha_i}{\partial F_i} = -\sigma$ in the symmetric equilibrium and the fixed fee is

$$F^* = \frac{1}{2\sigma} + f - \left(p^* - c\right) q\left(p^*\right)$$

and network profit is

$$\pi^* = \frac{1}{4\sigma}. \tag{B.5}$$

The socially optimal marginal price $\tilde{p}$ maximizes

$$\max_{\tilde{p}} \int_0^1 (1+\beta)\, u\left(q\left(\tilde{p}\right)\right) - cq\left(\tilde{p}\right) d\beta \qquad \text{or} \qquad \tilde{p} = \frac{2}{3}c.$$

Optimal usage pricing

$$p_i^*\left(F_i, F_j\right) = \frac{2\left(1 + \left(1 - \alpha_i\left(F_i, F_j\right)^{\gamma}\right) m\right) c}{1 + \alpha_i\left(F_i, F_j\right)^{\gamma} + \left(1 - \alpha_i\left(F_i, F_j\right)\right)^{\gamma}}$$

combined with the first-order condition for F_i defines a reaction function $F_i = F_i^R\left(F_j\right)$ with slope

$$\frac{dF_i}{dF_j} = -\frac{\frac{\partial^2 \pi}{\partial F_i F_j} - \sigma\delta\left(p\right)\left(\frac{\partial^2 \pi}{\partial F_i p_i} - \frac{\partial^2 \pi}{\partial F_i p_j}\right)}{\frac{\partial^2 \pi}{\partial F_i^2} + \sigma\delta\left(p\right)\left(\frac{\partial^2 \pi}{\partial F_i p_i} - \frac{\partial^2 \pi}{\partial F_i p_j}\right)}$$

where $\frac{\partial p_i}{\partial F_i} = -\sigma \frac{\partial p_i}{\partial \alpha_i} = -\sigma\delta\left(\alpha_i\right)$ and $\frac{\partial p_i}{\partial F_i} = -\frac{\partial p_j}{\partial F_i}$. Eq. B.4 yields

$$\begin{aligned}\frac{dF_i}{dF_j} &= -\frac{\sigma + \sigma^2\gamma\left(1+\gamma\right)\alpha_i^{\gamma-1} mc\left(q\left(p_j\right) - q\left(p_i\right)\right) - \sigma\delta\left(p\right)\left(\frac{\partial^2 \pi}{\partial F_i p_i} - \frac{\partial^2 \pi}{\partial F_i p_j}\right)}{-2\sigma - \sigma^2\gamma\left(1+\gamma\right)\alpha_i^{\gamma-1} mc\left(q\left(p_j\right) - q\left(p_i\right)\right) + \sigma\delta\left(p\right)\left(\frac{\partial^2 \pi}{\partial F_i p_i} - \frac{\partial^2 \pi}{\partial F_i p_j}\right)} \\ &= \frac{D+\sigma}{D}.\end{aligned}$$

The last equality follows from $N = -D - \sigma$, where N is the numerator and D the denominator of the left-hand term. Hence, $\frac{dF_i}{dF_j}$ is positive and smaller than 1 if $D < -\sigma$. This condition is satisfied if σ is close to zero, since then

$$\begin{aligned}
D =& -2\sigma - \sigma^2\gamma\left(1+\gamma\right)\alpha_i^{\gamma-1}mc\left(q\left(p_j\right)-q\left(p_i\right)\right) \\
& -\sigma^2\gamma\delta\left(\alpha_i\right)\left(q\left(p_i\right)+\left(p_i-c\right)\frac{\partial q\left(p_i\right)}{\partial p_i}\right. \\
& \qquad \left. -\left(1-\left(1+\gamma\right)\alpha_i^{\gamma}\right)mc\left(\frac{\partial q\left(p_i\right)}{\partial p_i}+\frac{\partial q\left(p_j\right)}{\partial p_j}\right)\right) \\
\backsimeq & -2\sigma < -\sigma.
\end{aligned}$$

Given network j's strategy (F_j, p_j), network i's best response entails $p_i^*\left(F_i, F_j\right)$ and network i's profit if it chooses F_i is

$$\pi_i\left(F_i\right) = \alpha_i\left(F_i - f\right) + \alpha_i\left(p_i^* - c - \left(1-\alpha_i^{\gamma}\right)mc\right)q\left(p_i^*\right) + \alpha_i\left(1-\alpha_i^{\gamma}\right)mcq\left(p_j\right)$$

The first-order derivative of this function is

$$\begin{aligned}
\frac{d\pi_i\left(F_i\right)}{dF_i} =& \alpha_i - \sigma\left(F_i - f\right) - \sigma\left(p_i^* - c\right)q\left(p_i^*\right) \\
& + \alpha_i\frac{\partial p_i}{\partial F_i}\left(\frac{\partial q\left(p_i^*\right)}{\partial p_i}\left(p_i^* - c - \left(1-\alpha_i^{\gamma}\right)mc\right) + q\left(p_i^*\right)\right) \\
& - \sigma\left(1-\left(1+\gamma\right)\alpha_i^{\gamma}\right)mc\left(q\left(p_j\right)-q\left(p_i^*\right)\right)
\end{aligned}$$

and the second-order derivative is

$$\begin{aligned}
\frac{\partial^2\Pi_i}{\partial F_i^2} =& -2\sigma + \sigma^2\alpha_i\frac{\partial\delta\left(\alpha_i\right)}{\partial\alpha_i}\left(p_i^* - c - \left(1-\alpha_i^{\gamma}\right)mc\right)\frac{\partial^2 q\left(p_i^*\right)}{\partial p_i^2} \\
& + \sigma^2\gamma mc\left(1+\gamma\right)\alpha_i^{\gamma-1}\left(q\left(p_j\right)-q\left(p_i^*\right)\right) \\
& + \sigma^2\alpha_i\frac{\partial\delta\left(\alpha_i\right)}{\partial\alpha_i}\left(q\left(p_i^*\right) - \eta\left(p_i^*\right)^{-(\eta-1)}\left(p_i^* - c - \left(1-\alpha_i^{\gamma}\right)mc\right)\right) \\
& - 2\sigma^2\alpha_i\left(\delta\left(\alpha_i\right)\right)^2\eta\left(p_i^*\right)^{-(\eta-1)}.
\end{aligned}$$

Hence, for σ close to zero, network i's profit function is strictly concave, that is

$$\frac{\partial^2 \Pi_i}{\partial F_i^2} \backsimeq -2\sigma.$$

B.2 Proof of Proposition 3.2

A similar argument to the Proof of Proposition 3.1 illustrates that there does not exist an equilibrium in which one firm corners the market: Suppose that network 1 corners the market. It will set the marginal price equal to the welfare-maximizing price given the call externalities and not charge a termination differential, that is $p_1 = \widehat{p}_1 = \frac{2}{3}c$ with $F_1 - f \geq 0$ and $\pi_2 = 0$. But network 2 could charge $p_2 = \widehat{p}_2 = \frac{2}{3}c$, and $F_2 = F_1 + \epsilon$. For ϵ small enough, its profit would then be $\tilde{\pi}_2 \simeq \frac{1}{2}(F_2 - f) \geq \frac{\epsilon}{2} > 0$, a contradiction.

With termination-based price discrimination, profit of network i is

$$\begin{aligned}\pi_i =& \alpha_i (F_i - f) + \alpha_i^{1+\gamma} (p_i - c) q(p_i) \\ &+ \alpha_i (1 - \alpha_i^{\gamma}) \left((\widehat{p}_i - c) q(\widehat{p}_i) + mc \left(q(\widehat{p}_j) - q(\widehat{p}_i) \right) \right).\end{aligned}$$

Suppose again that networks first maximize profits by choosing optimal prices p_i and $\widehat{p}_i$ while holding market share constant (through adjustments in the fixed fee F_i). Differentiating Eq. 3.2 with respect to the on-net price p_i (holding market shares constant) gives

$$\frac{\partial F_i}{\partial p_i} = -\alpha_i^{\gamma} \left(1 + \frac{1}{2}\eta \right) p_i^{-\eta}. \tag{B.6}$$

Maximizing the profit function with respect to the on-net price p_i holding market share constant gives

$$\frac{\partial F_i}{\partial p_i} = -\alpha_i^{\gamma} \left(p_i^{-\eta} - \eta (p_i - c) p_i^{-(\eta+1)} \right) \tag{B.7}$$

and Eqs. B.6 and B.7 imply

$$\frac{p_i - c}{p_i} = -\frac{1}{2}. \tag{B.8}$$

Analogous, differentiating Eq. 3.2 with respect to the off-net price $\widehat{p}_i$ gives

$$\frac{\partial F_i}{\partial \widehat{p}_i} = -\widehat{p}_i^{-\eta}\left(1 - \alpha_i^{\gamma} - \frac{1}{2}\eta\left(1 - \alpha_j^{\gamma}\right)\right). \tag{B.9}$$

Maximizing the profit function with respect to the off-net price $\widehat{p}_i$ holding market share constant gives

$$\frac{\partial F_i}{\partial \widehat{p}_i} = -\left(1 - \alpha_i^{\gamma}\right)\left(\widehat{p}_i^{-\eta} - \eta\left(\widehat{p}_i - (1+m)\,c\right)\widehat{p}_i^{-(\eta+1)}\right) \tag{B.10}$$

and Eqs. B.9 and B.10 imply

$$\frac{\left(\widehat{p}_i - (1+m)\,c\right)}{\widehat{p}_i} = \frac{1}{2}\frac{\left(1 - \alpha_j^{\gamma}\right)}{\left(1 - \alpha_i^{\gamma}\right)}. \tag{B.11}$$

Hence, marginal prices in the symmetric equilibrium are

$$p^* = \frac{2}{3}c \qquad \text{and} \qquad \widehat{p}^* = 2\,(1+m)\,c.$$

Differentiating the profit function with respect to the fixed fee yields

$$\begin{aligned}\frac{\partial \pi_i}{\partial F_i} =& \alpha_i + \frac{\partial \alpha_i}{\partial F_i}\left(F_i - f + (1+\gamma)\,\alpha_i^{\gamma}\left(p_i - c\right)q\left(p_i\right)\right) \\ &+ \frac{\partial \alpha_i}{\partial F_i}\left(1 - (1+\gamma)\,\alpha_i^{\gamma}\right)\left[\left(\widehat{p}_i - c_i\right)q\left(\widehat{p}_i\right) + mc\left(q\left(\widehat{p}_j\right) - q\left(\widehat{p}_i\right)\right)\right]\end{aligned} \tag{B.12}$$

and from Eq. 3.2, we have in the symmetric equilibrium

$$\frac{\partial \alpha_i}{\partial F_i} = \frac{-\sigma}{1 - \gamma\sigma\left(\alpha_i^{\gamma-1} + \left(1 - \alpha_i\right)^{\gamma-1}\right)\left[w\left(p\right) - w\left(\widehat{p}\right)\right]},$$

where $w(p) = v(p) + \frac{1}{2}u(q(p))$ is the average variable surplus of consumers. Hence, the fixed fee in the symmetric equilibrium is

$$\begin{aligned} F^* =& \frac{1}{2\sigma} + f - \gamma\left(\frac{1}{2}\right)^{1+\gamma}(w(p) - w(\widehat{p})) - (1+\gamma)\left(\frac{1}{2}\right)^{\gamma}(p-c)\,q(p) \\ & - \left(1 - (1+\gamma)\left(\frac{1}{2}\right)^{\gamma}\right)(\widehat{p}-c)\,q(\widehat{p}) \end{aligned}$$

and network profit is

$$\pi^* = \frac{1}{4\sigma} - \gamma\left(\frac{1}{2}\right)^{1+\gamma}\left(\frac{1}{2}(w(p) - w(\widehat{p})) + (p^*-c)\,q(p^*) - (\widehat{p}^*-c)\,q(\widehat{p}^*)\right).$$

Notice that $\frac{\partial\alpha_i}{\partial F_i} = -\frac{\partial\alpha_i}{\partial F_j}$ and let

$$\lambda = \frac{1}{1 - \gamma\sigma\left(\alpha_i^{\gamma-1} + (1-\alpha_i)^{\gamma-1}\right)(w(p) - w(\widehat{p}))}$$

and therefore $\frac{\partial\alpha_i}{\partial F_i} = -\sigma\lambda$. Optimal usage pricing with

$$p_i^* = \frac{2}{3}c \quad \text{and} \quad \widehat{p}_i^*(F_i, F_j) = \frac{2\left(1 - (\alpha_i(F_i, F_j))^{\gamma}\right)}{2\left(1 - (\alpha_i(F_i, F_j))^{\gamma}\right) - \left(1 - (\alpha_j(F_i, F_j))^{\gamma}\right)}(1+m)\,c.$$

combined with the first-order condition for F_i defines a reaction function $F_i = F_i^R(F_j)$ with slope

$$\frac{dF_i}{dF_j} = -\frac{\frac{\partial^2\pi}{\partial F_i F_j} - \sigma\lambda\frac{\partial\widehat{p}_i}{\partial\alpha_i}\left(\frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_i} - \frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_j}\right)}{\frac{\partial^2\pi}{\partial F_i^2} + \sigma\lambda\frac{\partial\widehat{p}_i}{\partial\alpha_i}\left(\frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_i} - \frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_j}\right)}.$$

From Eq. B.12 we get $\frac{\partial^2\pi}{\partial F_i F_j} = -\frac{\partial^2\pi}{\partial F_i^2} - \frac{\partial\alpha_i}{\partial F_j} = -\frac{\partial^2\pi}{\partial F_i^2} - \sigma\lambda$ and therefore

$$\frac{dF_i}{dF_j} = \frac{\frac{\partial^2\pi}{\partial F_i^2} + \sigma\lambda + \sigma\lambda\frac{\partial\widehat{p}_i}{\partial\alpha_i}\left(\frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_i} - \frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_j}\right)}{\frac{\partial^2\pi}{\partial F_i^2} + \sigma\lambda\frac{\partial\widehat{p}_i}{\partial\alpha_i}\left(\frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_i} - \frac{\partial^2\pi}{\partial F_i\partial\widehat{p}_j}\right)} = \frac{D + \sigma\lambda}{D}.$$

Hence, $\frac{dF_i}{dF_j}$ is positive and smaller than 1 if $D < -\sigma\lambda$. Let

$$\psi = \gamma(\gamma-1)\left((\gamma-1)\alpha_i^{\gamma-2} - (1-\alpha_i)^{\gamma-2}\right)(w(p) - w(\hat{p}))$$

and notice that the terms $\frac{\partial \widehat{p}_i}{\partial \alpha_i}$ and ψ are independent of σ and λ is positive and approaches a constant as $\sigma \to 0$. Hence, the condition $D < -\sigma\lambda$ is satisfied if σ is close to zero since then

$$\begin{aligned}
D =& -2\sigma\lambda + \sigma^3\lambda^3\psi(F_i - f) + \sigma^2\lambda^2\left(\sigma\lambda\psi + \gamma\alpha_i^{\gamma-1}\right)(1+\gamma)(p_i - c)\,q(p_i) \\
&+ \sigma^2\lambda^2\left(\sigma\lambda\psi(1-(1+\gamma)\alpha_i^{\gamma}) - \gamma(1+\gamma)\alpha_i^{\gamma-1}\right) \\
&\qquad \times \left((\widehat{p}_i - c)\,q(\widehat{p}_i) + mc\,(q(\widehat{p}_j) - q(\widehat{p}_i))\right) \\
&- \sigma^2\lambda^2\frac{\partial \widehat{p}_i}{\partial \alpha_i}\left((1-(1+\gamma)\alpha_i^{\gamma})\,q(\widehat{p}_i) + (1-(1+\gamma)\alpha_i^{\gamma})(\widehat{p}_i - c)\frac{\partial q(\widehat{p}_i)}{\partial \widehat{p}_i}\right. \\
&\qquad \left. - (1-(1+\gamma)\alpha_i^{\gamma})\,mc\left(\frac{\partial q(\widehat{p}_i)}{\partial \widehat{p}_i} + \frac{\partial q(\widehat{p}_j)}{\partial \widehat{p}_j}\right)\right) \\
\backsimeq& -2\sigma\lambda < -\sigma\lambda.
\end{aligned}$$

Given network j's strategy (F_j, p_j), network i's best response entails p_i^* and $\widehat{p}_i^*(F_i, F_j)$ and network i's profit if it chooses F_i is

$$\begin{aligned}
\pi_i(F_i) =& \alpha_i(F_i - f) + \alpha_i^{1+\gamma}(p_i^* - c)\,q(p_i^*) + \alpha_i(1-\alpha_i^{\gamma})\big[(\widehat{p}_i^* - c)\,q(\widehat{p}_i^*) \\
&+ mc\,(q(\widehat{p}_j) - q(\widehat{p}_i^*))\big].
\end{aligned}$$

The first-order derivative of this function is

$$\begin{aligned}\frac{d\pi_i(F_i)}{dF_i} = \alpha_i &+ \frac{\partial \alpha_i}{\partial F_i}\left(F_i - f + (1+\gamma)\,\alpha_i^{\gamma}\,(p_i^* - c)\,q(p_i^*)\right)\\ &+ \frac{\partial \alpha_i}{\partial F_i}\left(1-(1+\gamma)\,\alpha_i^{\gamma}\right)\left((\widehat{p}_i^* - c)\,q(\widehat{p}_i^*) + mc\,(q(\widehat{p}_j) - q(\widehat{p}_i^*))\right)\\ &+ \alpha_i\left(1-\alpha_i^{\gamma}\right)\frac{\partial \alpha_i}{\partial F_i}\frac{\partial \widehat{p}_i^*}{\partial \alpha_i}\left(\frac{\partial q(\widehat{p}_i^*)}{\partial \widehat{p}_i}\left(\widehat{p}_i^* - (1-m)\,c\right) + q(\widehat{p}_i^*)\right)\end{aligned}$$

and the second-order derivative is

$$\begin{aligned}\frac{d^2\Pi_i}{dF_i^2} =& -2\sigma\lambda + \sigma^3\lambda^3\psi\,(F_i - f) + \sigma^2\lambda^2\left(\sigma\lambda + \gamma\alpha_i^{\gamma-1}\right)(1+\gamma)\,(p_i^* - c)\,q(p_i^*)\\ &+ \sigma^2\lambda^2\left(\left(1-(1+\gamma)\,\alpha_i^{\gamma}\right) - \sigma\lambda\gamma\,(1+\gamma)\,\alpha_i^{\gamma-1}\right)\\ &\qquad\times\left((\widehat{p}_i^* - c)\,q(\widehat{p}_i^*) + mc\,(q(\widehat{p}_j) - q(\widehat{p}_i^*))\right)\\ &+ \sigma^2\lambda^2\left(1-(1+\gamma)\,\alpha_i^{\gamma}\frac{\partial \widehat{p}_i^*}{\partial \alpha_i} + \alpha_i\left(1-\alpha_i^{\gamma}\right)\left(\sigma\lambda\psi\frac{\partial \widehat{p}_i^*}{\partial \alpha_i} + \frac{\partial \widehat{p}_i^*}{\partial \alpha_i}\right)\right)\\ &\qquad\times\left(\frac{\partial q(\widehat{p}_i^*)}{\partial \widehat{p}_i}\left(\widehat{p}_i^* - (1-m)\,c\right) + q(\widehat{p}_i^*)\right)\\ &+ \sigma^2\lambda^2\alpha_i\left(1-\alpha_i^{\gamma}\right)\left(\frac{\partial \widehat{p}_i^*}{\partial \alpha_i}\right)^2\left(\frac{\partial^2 q(\widehat{p}_i^*)}{\partial \widehat{p}_i^2}\left(\widehat{p}_i^* - (1-m)\,c\right) + 2\frac{\partial q(\widehat{p}_i^*)}{\partial \widehat{p}_i}\right).\end{aligned}$$

For σ close to zero, network i's profit function is strictly concave, that is,

$$\frac{\partial^2\Pi_i}{\partial F_i^2} \backsimeq -2\sigma\lambda < 0.$$

B.3 Proof of Proposition 3.3

No Termination-Based Price Discrimination

With the ex-ante symmetry $\frac{A}{\sigma}$, the market share for network 1 is

$$\alpha_1 = \frac{1}{2} + A + \sigma\left[v\left(p_1\right) - F_1 - \left(v\left(p_2\right) - F_2\right)\right] - \frac{1}{2}\left(1 - \alpha_1^{\gamma} - \alpha_2^{\gamma}\right)\left[u\left(q\left(p_1\right)\right) - u\left(q\left(p_2\right)\right)\right].$$

The two expressions (Eq. B.1 and B.2) derived for $\frac{\partial F_i}{\partial p_i}$ from the Proof of Proposition 3.1 are unaltered by the ex-ante symmetry and Eq. B.3 then yields

$$p_i = \frac{2\left(1 + \left(1 - \alpha_i^{\gamma}\right) m\right) c}{1 + \alpha_i^{\gamma} + \alpha_j^{\gamma}}$$

from which immediately follows that $p_i < p_j$ if and only if $\alpha_i > \frac{1}{2}$ and $|p_i - p_j|$is decreasing in γ.

Termination-Based Price Discrimination

With the ex-ante symmetry $\frac{A}{\sigma}$, the market share for network 1 is

$$\begin{aligned}\alpha_i =& \frac{1}{2} + A - \sigma\left(F_i - F_j\right) + \sigma\alpha_i^{\gamma}\left[v\left(p_i\right) + \frac{1}{2}u\left(q\left(p_i\right)\right)\right] \\ &+ \sigma\left(1 - \alpha_i^{\gamma}\right) v\left(\widehat{p}_i\right) - \frac{1}{2}\sigma\left(1 - \alpha_j^{\gamma}\right) u\left(q\left(\widehat{p}_i\right)\right) \\ &- \sigma\alpha_j^{\gamma}\left[v\left(p_j\right) + \frac{1}{2}u\left(q\left(p_j\right)\right)\right] - \sigma\left(1 - \alpha_j^{\gamma}\right) v\left(\widehat{p}_j\right) + \frac{1}{2}\sigma\left(1 - \alpha_i^{\gamma}\right) u\left(q\left(\widehat{p}_j\right)\right).\end{aligned}$$

Again, the expressions derived for $\frac{\partial F_i}{\partial p_i}$ and $\frac{\partial F_i}{\partial \widehat{p}_i}$ from the Proof of Proposition 3.2 are unaltered by the ex-ante symmetry and Eq. B.8 shows that network i's on-net price p_i is unaffected by its market share α_i, $p_i = \frac{2}{3}c$. On the other

hand, Eq. B.11 yields an off-net price for network i of

$$\widehat{p}_i = \frac{2\left(1-\alpha_i^{\gamma}\right)}{2\left(1-\alpha_i^{\gamma}\right)-\left(1-\alpha_j^{\gamma}\right)}\left(1+m\right)c$$

which implies that $\widehat{p}_i < \widehat{p}_j$ if and only if $\alpha_i < \frac{1}{2}$. Notice that large market share asymmetries result in a connectivity breakdown as the larger network l will charge prohibitively high off-net prices $\widehat{p}$. Indeed, as $\alpha_l \to \frac{2}{3}$, the off-net price $\widehat{p}_l$ goes to $+\infty$. The wedge between the termination differentials of the two networks is

$$\begin{aligned}\Delta &= \delta_l - \delta_s = \widehat{p}_l - \widehat{p}_s \\ &= \left(\frac{2\left(1-\alpha^{\gamma}\right)}{2\left(1-\alpha^{\gamma}\right)-\left(1-\left(1-\alpha\right)^{\gamma}\right)} - \frac{2\left(1-\left(1-\alpha\right)^{\gamma}\right)}{2\left(1-\left(1-\alpha\right)^{\gamma}\right)-\left(1-\alpha^{\gamma}\right)}\right)\left(1+m\right)c \\ &> 0\end{aligned}$$

where $\alpha_l = \alpha > \frac{1}{2}$. The wedge Δ is increasing in the market share differential, the termination markup m and cost c. Furthermore, the wedge Δ is increasing in the network bias (decreasing in γ) since the network bias reinforces market share asymmetries.

B.4 Proof of Proposition 3.4

In the absence of termination-based price discrimination, network profits are independent of the termination markup m, which follows straight from Eq. B.5. For the profit-maximizing price p^* to be social welfare maximizing, the termination markup m should lead to $p^* = \widetilde{p} = \frac{2}{3}c$, or

$$\widetilde{m} = -\frac{1+\kappa(\gamma)}{3(1-\kappa(\gamma))} \leq \frac{2}{3}.$$

The first-best social welfare-maximizing termination markup is always negative $\widetilde{m}$ but only feasible if $\widetilde{m} > \overline{m} = -\frac{c_0}{c}$.

On the other hand, in the symmetric equilibrium on-net/off-net differentials, industry profit is

$$\begin{aligned}\Pi^* = 2\pi^* =& \frac{1}{2\sigma} - \gamma\left(\frac{1}{2}\right)^{\gamma} \\ & \times \left(\frac{1}{2}\left[w(p_D) - w(\hat{p}_D)\right] + (p_D - c)\,q(p_D) - (\hat{p}_D - c)\,q(\hat{p}_D)\right) \\ =& \frac{1}{2\sigma} - \gamma\left(\frac{1}{2}\right)^{\gamma} c^{-\eta+1}\left[\left(\frac{2+\eta}{4(\eta-1)} - \frac{1}{2}\right)\left(\frac{2}{3}\right)^{-\eta+1}\right. \\ & \left. - \left(\frac{(2+\eta)(1+m)}{2(\eta-1)} + 1 + 2m\right)\left([2(1+m)]^{-\eta}\right)\right].\end{aligned}$$

Only the last term in the square brackets depends on the termination markup m. During preliminary access negotiation, network maximize

$$\max_m \left(\frac{(2+\eta)(1+m)}{2(\eta-1)} + 1 + 2m\right)\left([2(1+m)]^{-\eta}\right)$$

and the optimal termination markup

$$\widehat{m} = -\frac{3\eta - 2}{5\eta - 2}$$

is always negative, and a bill-and-keep system is optimal whenever $\widehat{m} < \bar{m}$ or

$$\eta > \frac{2\left(c - c_0\right)}{3c - 5c_0} = \frac{2\left(1 - \frac{c_0}{c}\right)}{3 - 5\frac{c_0}{c}}.$$

With termination-based price discrimination, the social welfare maximizing first-best outcome can never be achieved for any feasible markup since $\widehat{p}^* = 2\left(1 + m\right)c$.