Mobile Call Termination

Armstrong, Mark and Wright, Julian

Department of Economics, University College London

March 2007

Mobile Call Termination*

Mark Armstrong[†] and Julian Wright[‡]

September 2007

Abstract

Motivated by recent UK experience, we study the problem of mobile call termination. This is an intriguing policy story, in which regulation has been imposed on what appears to be a competitive industry. We introduce a framework which integrates two existing literatures: one analyzing fixed-to-mobile call termination (where the predicted market failure involves the termination charge being set at the monopoly level), and one analyzing mobile-to-mobile network interconnection (where the predicted termination charge is below the efficient level). Our unified framework allows us to consider the impact of wholesale arbitrage and demand-side substitution. With these features, we find the unregulated termination charge lies *between* the efficient and the monopoly benchmarks. There remains a rationale for regulation, albeit reduced relative to the earlier literature.

1 Introduction

There is an important, and perhaps growing, set of markets in which monopoly prices emerge even when competition is intense. That is to say, while industry profit is not excessive overall, there is an inefficient balance of prices: too high for some services, too low for others. Familiar examples involve consumer "lock-in" of various kinds, including markets with switching costs.[1] In these markets the typical pattern of prices involves "bargains-then-ripoffs", so that firms attract new consumers with generous deals up-front, consumers pay high (perhaps monopoly) prices once locked in. If competition is vigorous, the monopoly profits from locked-in consumers are transferred to new consumers, and the lifetime profitability of a consumer is approximately zero. In a sense, a new consumer's "future self" is exploited by the "present self".

A related, but distinct, set of markets exhibit what might be termed "competitive bottlenecks". Here, firms compete to attract one group of consumers. For natural technological or

*We are grateful to Stefan Behringer, David Harbord, Bruce Lyons, Paul Muysert, David Sappington, Tommaso Valletti, John Vickers, Ingo Vogelsang and Helen Weeds for very helpful comments. Armstrong is grateful for the support of the Economic and Social Research Council (UK).

[†]University College London, email mark.armstrong@ucl.ac.uk.

[‡]National University of Singapore, email jwright@nus.edu.sg.

[1]See Farrell and Klemperer (2007) for an overview.

geographical reasons, each consumer in this group wishes to deal with just one firm. A second group of consumers wishes to interact with the first group and, because each consumer in the first group deals exclusively with one firm, that firm can charge the second group high prices (or pay low input prices) for access to its captive customers. If competition is vigorous, the monopoly profits generated by the first group are passed back to these consumers in the form of subsidised service. Here, the first group may be said to exploit the second group. Examples of this kind of market include: newspaper advertising (where most readers tend to read a single newspaper due to time constraints, and so a newspaper can charge high fees to advertisers for access to its captive readers) and supermarkets (where a consumer tends to visit just one shop over the relevant time horizon, and so the supermarket can pay low wholesale prices to its suppliers for access to these captive shoppers). At the end of the paper we will discuss another possible application of this approach, to the contentious issue of "net neutrality" on the internet.[2]

However, perhaps the leading example of a competitive bottleneck involves call termination on mobile telephone networks. Call termination refers to the service whereby a network completes—or "terminates"—a call made to one of its subscribers by a caller on another network. (We focus on mobile, as opposed to fixed-line, networks since the former are currently much more competitive than the latter in most countries, and the issue of regulation there is more controversial.) Mobile networks compete to capture subscribers, but in the absence of regulation they may charge other networks excessively to talk to their subscribers. Concerns about mobile call termination being a bottleneck, and the associated high charges for call mobile subscribers, have lead to regulation of termination charges around the world, including in Australia, France, Germany, Italy, Japan and the UK.[3]

There are two broad types of call termination on mobile telephone networks: termination of calls made from other mobile networks (termed mobile-to-mobile, or MTM, termination in what follows), and termination of calls made by callers on the fixed-line telephone network (fixed-to-mobile, or FTM, termination).[4] In the literature to date, FTM and MTM mobile termination have been largely treated separately, with distinct market failures being highlighted. One aim of this paper is to integrate these two strands.

Broadly speaking, FTM call termination viewed in isolation is likely to involve unilateral monopoly pricing if left unchecked. The vast majority of mobile subscribers join just one mobile network, and so callers on the fixed telephone network must route calls through a subscriber's chosen network. No matter how competitive the market for mobile subscribers, a mobile network holds a monopoly over—and can set high charges for—delivering calls to its subscribers. That is to say, FTM call termination gives rise to a competitive bottleneck.[5] In this paper we examine the robustness of the competitive bottleneck problem and the associated need for continued regulation in this market.

[2] See Armstrong (2006) and Armstrong and Wright (2007) for general analyses of competitive bottlenecks.

[3] In some countries, notably the United States, mobile termination charges are indirectly regulated (at low levels) through reciprocity requirements with the fixed-line networks.

[4] In addition to voice traffic, an increasingly important form of communication is text (or SMS) messaging. Currently, text messages are almost entirely sent between mobile phones, and so SMS termination fits well into the framework for (voice) MTM call termination (although in most countries it is not yet regulated).

[5] The formal analysis this competitive bottleneck in the context of mobile telephony was developed in Armstrong (2002, section 3.1) and Wright (2002).

A similar bottleneck might be expected with MTM call termination. Indeed, as we will establish in section 4.3, when networks choose their MTM termination charges unilaterally (for instance, because networks must choose a single termination charge for both FTM and MTM traffic) this bottleneck is present to some extent. However, for historical reasons, MTM termination charges have traditionally been modelled as being chosen jointly, rather than unilaterally, by networks. As such, the study of two-way interconnection has led to a different focus: whether mobile networks can use a negotiated termination charge to relax competition for subscribers.[6] They can do this by setting MTM termination charges below cost. This causes off-net calls (i.e., calls from one mobile network to another mobile network) to be cheaper than on-net calls (calls from one mobile network to the same network), so that consumers prefer to join the smaller network. Each network's incentive to attract subscribers is thereby reduced. Thus, when networks can coordinate on the choice of a MTM termination charge, they may have an incentive to choose too *low* a charge. This low termination charge acts—somewhat counter-intuitively—to harm mobile subscribers, who prefer the intense competition which accompanies a higher charge. (By contrast, a high unregulated FTM termination charge harms fixed-line callers and benefits MTM subscribers.)

In section 3 we present a basic model in which both FTM and MTM calls are present, and which captures the contrasting features of FTM and MTM call termination. We show that if there are a large number of mobile networks then the market failure associated with MTM termination is negligible, in contrast to the situation with FTM termination. In section 4 we explore the extent to which these results are robust to realistic generalizations. In part this is motivated by the observation that, to the best of our knowledge, no regulator has yet taken seriously the concern that firms may set MTM termination charges which are too low. Rather, like FTM termination, policy has acted to prevent operators setting MTM termination charges which are too high.

We find, in section 4.3, that this puzzle can be reconciled by taking into account the practical constraint of wholesale arbitrage. By this we mean that a mobile operator cannot maintain a high FTM termination charge together with a low MTM termination charge, since the fixed network could then "transit" its calls via another mobile operator and so end up paying the lower MTM rate (plus a small transit charge). As a result, a mobile operator is forced to set (approximately) a uniform termination charge for FTM and MTM traffic. With the two charges locked together, the temptation to extract termination profits dominates the incentive to set a low termination charge in order to relax network competition. Nevertheless, a network's incentive to set monopoly termination charges is mitigated. Setting the uniform termination charge too high strengthens network effects, making the firms tougher rivals. To avoid this effect, networks will keep their termination charges below the monopoly level.

Additional effects arise once we take into account that fixed-line callers are often also mobile subscribers, so that the two groups are not disjoint as assumed in the standard competitive bottleneck story. This means callers can make MTM calls instead of FTM calls if the former are cheaper. In section 4.4 we find such demand-side substitution between FTM and MTM calls sharpens our results, and brings the equilibrium unregulated charge still closer to the efficient level. As a result, the welfare gains from regulation are smaller than the previous literature may have indicated, although the predicted termination charge

[6] See Laffont, Rey and Tirole (1998b) and Gans and King (2001).

without regulation remains excessive.

Section 4 also explores the impact of call and network externalities. We find in section 4.1 that call externalities do not reduce the welfare costs of unregulated termination charges, since they act to reduce both the unregulated level *and* the efficient level of charges. By contrast, in section 4.2 we show that if the population of mobile subscribers expands when the deals on offer improve, it is efficient to set the termination charge above cost. If the termination charge is raised above cost, this will generate (i) FTM termination profit which is passed onto subscribers and (ii) positive network effects which intensify competition between networks. Both of these effects will induce more people to subscribe. Thus, the potential for network expansion reduces still further the welfare gains available from regulation.

Before turning to this analysis, though, we review some relevant aspects of the UK mobile sector and its regulation. This will be used to motivate our particular theoretical framework and to discuss the policy implications of our theoretical results throughout the paper.

2 UK Experience

Since their inception, regulatory price caps on mobile termination in the UK have faced considerable resistance from mobile operators (as well as support from fixed-line operators). They have also resulted in a large amount of paperwork. Since 1997, there have been several thousand pages of publicly released government reports and submissions on mobile call termination in the UK. Using materials from these reports, in this section we give a brief overview of the most relevant aspects of the UK mobile market.

The regulation of mobile termination was initiated in 1998 by Oftel (then the UK telecommunications sector regulator), which tried to control the FTM termination charges set by the two largest mobile networks *Cellnet* (the precursor to the current O_2) and *Vodafone*. The regulation was challenged by these mobile operators, leading to a competition enquiry—see MMC (1999). This enquiry did not investigate MTM termination charges, nor did it investigate FTM termination charges levied by the two newer networks, *Orange* and *T-Mobile*, which had entered only recently at that point. The enquiry concluded that the two established networks' FTM termination charges were too high in relation to cost, and based on its recommendations, Oftel regulated their FTM termination charges with a price cap. The imminent expiry of this price cap led to a 2002 Competition Commission enquiry—see Competition Commission (2003). The Commission upheld Oftel's proposed regulation, which covered all four mobile operators and both FTM and MTM (voice) termination charges. (However, SMS termination was not covered in the enquiry, nor by subsequent regulation to date.) Shortly before the 2002 enquiry, a fifth network (*H3G)* entered the market, although this immature network was excluded from that enquiry.

Subsequent reviews by Ofcom (the current UK telecommunications regulator), in 2004 and more recently in March 2007, extended these regulations. As of 2007, all five networks are subject to regulated price caps for call termination, with reductions in these caps applying from 2007 through to 2011, at which point they will be reviewed again (see Ofcom (2007)). Table 1 gives an indication of the history of average mobile termination charges in recent years (for all UK networks).

Table 1: Average mobile termination charges (in pence per minute)[7]

	2001	2002	2003	2004	2005
all UK operators:	11.1	10.7	9.9	7.9	5.9

Thus, termination charges approximately halved over five years, in large part due to tightened regulation. However, the networks do not all set the same average level of charges, as shown in Table 2 which shows termination charges set by the five operators in 2006. In 2006, the newest entrant, *H3G*, faced softer regulation than the established networks, and took advantage of this to set termination charges which were substantially higher than its rivals.

Table 2: The impact of asymmetric regulation on termination charges (March 2006)[8]

	Daytime	Evening	Weekends
O_2	6.4	6.3	3.1
Orange	7.6	5.4	4.3
T-Mobile	8.1	4.0	4.0
Vodafone	8.5	3.4	2.8
H3G	15.6	10.8	2.5

The mobile industry in the UK currently consists of four mature and roughly equal sized mobile networks, each with around 15 million subscribers, and a recent entrant (*H3G*) with fewer subscribers:

Table 3: Subscriber numbers in 2001 and 2005[9]

	Vodafone	O_2	*Orange*	*T-Mobile*	*H3G*
Active subscriber numbers in 2001 (m)	11.0	11.1	12.4	10.3	n/a
Active subscriber numbers in 2005 (m)	14.8	17.0	14.9	15.3	3.5

This compares to a total of only around 8 million subscribers at the time of initial regulation in 1998—see Competition Commission (2003, Figure 6.1). Total annual retail revenue in 2007 is about £13 billion, and mobile call termination generates annual revenue of around £2.5 billion (Ofcom, 2007, p7). As mobile penetration has grown, and as the cost of making off-net MTM calls has fallen (see below), the importance of MTM calls in generating termination revenue has risen. Whereas at the time of the Competition Commission's enquiry in 2002, nearly three quarters of mobile termination revenue was from FTM traffic, now it is

[7] See Figure 3.38 in Ofcom (2006a).

[8] Taken from Figure 1 in Ofcom (2006b).

[9] See Figure 3.40 in Ofcom (2006a). The current total number of subscribers exceeds the official UK population. This can be explained by the fact that some people have two or more phone subscriptions; for example, one for business and one for personal use; or because they continue to hold an old subscription which they no longer actively use. This is consistent with evidence from Figure 3.46 in Ofcom (2006a) that 10% of UK households do not have access to a mobile phone, similar to the proportion without access to a fixed-line phone.

only about one third of the total.[10] We will see in section 4.3 that the volume of FTM traffic relative to MTM traffic will play an important role in determining the equilibrium termination charge, and a greater share of MTM traffic will tend to reduce the market failures associated with unregulated termination charges, all else equal.

Importantly, and in contrast to the situation in many countries, all mobile networks in the UK are separately owned from the significant fixed networks. This means potential concerns that arise in other jurisdictions about vertical price squeezes and foreclosure are not likely to be an issue in the UK.

The regulated termination charges in Table 1 were calculated using two kinds of markup over estimates of marginal termination costs. The first markup is designed to tax fixed-line callers to subsidise mobile network use in order to stimulate mobile network expansion. As will be explained further in section 4.2, this is consistent with the positive network externalities generated by additional subscribers, although the quantification of this network externality which Ofcom (2007, p349) most recently estimated to be only 0.3 pence per minute, in part reflects Ofcom's assumption that any subsidy can be targeted at marginal subscribers. The second markup reflects an intended contribution to a mobile network's fixed and common costs. In section 3.3, we argue that including fixed and common costs in this context is likely to be a flawed policy.

The price caps for FTM and MTM termination have been set equal to each other, as were the actual FTM and MTM termination charges set by networks. This contrasts with our initial analysis in section 3 below, where in the absence of regulation we argue that networks may wish to set lower charges for terminating MTM calls (although often the welfare-maximizing charges are the same for the two kinds of calls.) It is worth noting there was no regulatory constraint that prevented networks setting lower MTM charges. We will return to what might explain this observation in section 3.3.

Table 4: Average call charges, pence per minute/message[11]

	Off-net MTM calls	On-net MTM calls	FTM calls	Text messages
2001	26.2	5.9	14.4	8.1
2005	11.3	4.2	11.5	6.3

In Table 4 we give some idea of average per-minute retail prices for on-net and off-net MTM calls, as well as FTM calls and text messages. The decline in off-net MTM and FTM call charges evident here is no doubt partly due to the fall in termination charges documented in Table 1. However, the decline in off-net MTM call charges has been particularly dramatic and this likely reflects that a growing number of call plans include some free off-net MTM calls. Despite the narrowing of the price differentials between off-net and on-net calls, though,

[10] See Figure 3.46 in Competition Commission (2003) and Ofcom (2007, p7).

[11] Cost of FTM calls is taken from Figure 3.22 in Ofcom (2006a). It is a complicated, and largely arbitrary, task to give precise estimates for the prices of the various types of calls and messages originating on mobile networks. This is because mobile networks each offer a wide variety of tariffs, with different monthly rentals (where applicable) corresponding to different volumes of inclusive call minutes and text messages. The numbers in Table 3 other than those for FTM calls are taken from Figure 3.39 in Ofcom (2006a), although the method of calculation is not clear from that document.

the difference remains striking. Due in part to this price differential, Table 5 shows the volumes of off-net and on-net calls to be rather unbalanced. With equal off-net and on-net charges and four roughly symmetric networks, one might expect that off-net traffic would be approximately three times greater than on-net traffic, rather than only twice as much as in 2001 or only some 35% more as was the case by 2005.

Table 5: Shares of types of mobile calls[12]

	Off-net MTM	On-net MTM	mobile to fixed
% in 2001	31.0	14.9	54.1
% in 2005	34.8	25.8	39.4

In addition to the price differential documented in Table 4, there are at least two other reasons why call volumes are biased towards on-net calls. First, "closed user groups", i.e., groups of subscribers who predominantly make calls within their own group, may be present. Often, such groups have their network subscription decision made centrally (e.g., by their employer's procurement office) and to a single network. To the extent these groups are widespread, this will boost the share of on-net calls in the market. Second, there may be some substitution between MTM and FTM calls. A mobile subscriber, when she is in the home or office, has a choice between calling another mobile subscriber by means of either her fixed line or mobile phone. In many cases, she will use the cheaper alternative. With the charges in Table 4, this implies she will often want to make an on-net MTM call if the recipient is on the same mobile network, although less so for off-net MTM calls. This will amplify the bias towards on-net call volumes. We discuss this issue further in section 4.4.

Since 2003, Ofcom has determined that the mobile *retail* market was effectively competitive.[13] In contrast, Oftel, UK regulators have consistently ruled that each mobile network has a monopoly of call termination on its own network. The idea is that there is no practical way for people to call someone on the go without calling the person's mobile phone and having the call terminated by the mobile network to which that person has subscribed. This is not to suggest that there are absolutely no substitution possibilities. One objective of this paper is to explore the extent to which this conclusion remains robust to such possibilities, including the fact that if FTM calls are too expensive, people may substitute by making cheaper MTM calls instead (see section 4.4).[14]

[12] Data from Figure 3.50 in Ofcom (2006a). We exclude international calls, premium rate services and various "other services".

[13] Oftel, *Mobile Access and Call Origination Market: Identification and Analysis of Market and Determination of Market Power*, 3 October 2003.

[14] Another possibility is that subscribers may sign up to multiple mobile networks. This could potentially resolve the bottleneck problem if a subscriber's callers can choose to call him on the network with the lower termination charge. However, despite relatively high off-net MTM and FTM pricing, so far this does not seem to have been a very significant constraint on pricing. According to Ofcom (2007, para. 3.29), in 2006 only 7% of mobile subscribers have more than one mobile phone, and this was typically for reasons other than pricing, such as to separate business and personal calls.

3 A Basic Model of the Mobile Market

The principal purpose of the basic model in this section is to contrast the pricing of MTM termination with the pricing of FTM termination. In this model we assume that firms can freely set differential termination charges for FTM and MTM traffic, and we assume away call and network externalities. (Section 4 will provide various extensions to this basic model.) In this framework, provided FTM retail call charges are regulated to be equal to cost, we will see that welfare is maximized when both the FTM and MTM termination charges are equal to cost. Without regulation, we will see that mobile networks in this model will wish to set an excessive FTM termination charge (in fact, the monopoly charge), while they wish to set too low a MTM termination charge.

To model MTM calls, a standard framework of two-way interconnection between symmetric networks is adopted, based on the model of Laffont, Rey, and Tirole (1998b) and Gans and King (2001) extended to more than two firms.[15] There are $K \geq 2$ mobile networks which offer differentiated services. Each network charges a termination charge for completing MTM calls originating on the other networks. The networks are assumed to negotiate a common MTM termination charge, denoted a.[16]

Mobile subscribers are assumed to be identical in terms of their demand for calls to other mobile subscribers.[17] With this simplification, if subscriber j faces a per-minute charge p for calling subscriber k, j will choose to make (an average of) $q(p)$ minutes of calls to k. Thus, each subscriber is equally likely to wish to call any other subscriber. Let $v(p)$ be consumer surplus associated with the demand function $q(\cdot)$, so that $v'(p) \equiv -q(p)$.

Added to this framework is a model of FTM termination described in section 3.1 of Armstrong (2002) and Wright (2002). There is a fixed-line network, from which a demand for FTM calls is generated. As in the UK, we assume this fixed sector is separately owned from the mobile sector. Each mobile network unilaterally chooses a termination charge for completing FTM calls, and network i's FTM termination charge is denoted A_i. (Where possible, we use upper case notation for calls from the fixed network and lower case notation for calls from the mobile networks.) We assume that in the first stage firms negotiate a reciprocal MTM interconnection price a, and subsequently, in a second stage, they set their FTM termination charges A_i together with their retail tariffs to mobile customers.[18]

If the retail price for FTM calls to mobile network i is P_i per minute, suppose that there are $Q(P_i)$ FTM minutes of calls to each subscriber on network i.[19] We assume the fixed

[15] Armstrong (1998), Laffont, Rey, and Tirole (1998a) and Carter and Wright (1999) analyze the situation where networks do not set differential on-net and off-net call charges.

[16] Throughout the paper it is assumed the parameters are such that an equilibrium exists. Laffont, Rey, and Tirole (1998b) discuss the conditions for such an equilibrium to exist in the retail pricing stage, which requires that the MTM termination charge does not differ too much from cost and/or that the mobile networks are sufficiently differentiated. More generally, for equilibrium to exist at all stages of the game we require that networks be sufficiently differentiated.

[17] Dessein (2003), Hahn (2004) and Valletti and Houpis (2005) develop models where subscribers differ in their demand for calls.

[18] The results obtained do not change when the FTM termination charge is instead set in stage two, with retail prices set in a subsequent third stage. However, the case with simultaneous setting of the FTM termination charge and other retail prices simplifies the exposition.

[19] One might also consider mobile-to-fixed calls. However, in the simple frameworks presented here, these

network can set different call charges to different mobile networks to reflect the networks' different FTM termination charges.[20] In general, we expect the price P_i to be an increasing function of the FTM termination charge A_i, and write $P_i = P(A_i)$. For instance, it may be that

$$P(A) = C + A \ , \tag{1}$$

where C is the fixed network's marginal cost of originating a call. In this case, the FTM call charge is equal to the fixed network's total cost of making such calls. Such pricing could arise as a result of the regulation of the fixed network or competition between fixed networks.[21] Let $V(\cdot)$ be the consumer surplus function associated with the demand function $Q(\cdot)$, so that $V'(P) \equiv -Q(P)$. Define

$$F(A) \equiv (A - c_T)Q(P(A)) \tag{2}$$

to be a mobile network's profit, per subscriber, from providing termination services for the fixed network when its FTM termination charge is A.

Each mobile firm is assumed to incur a marginal cost c_O of originating a call and a marginal cost c_T of terminating a call, so the actual marginal cost of a MTM call is $c_O + c_T$. In addition, there is a fixed cost f of serving each mobile subscriber, which includes the subscriber's handset, billing costs, and so on. For now, assume that FTM and MTM calls are independent markets, and that the call charge in one market does not affect the demand in the other market. Figure 1 depicts our stylized model of the mobile industry.

In this basic model, we assume there is an exogenously fixed number of mobile subscribers, which we normalise to 1. As depicted in Figure 1, denote firm i's on-net MTM call charge by $\hat{p}_i$ and its off-net MTM call charge by p_i. In addition, the firm charges a fixed (rental) charge r_i for subscribing. If firm i's market share is s_i, its subscribers make a fraction s_i of their calls on-net and the remaining $1 - s_i$ calls off-net. Then a subscriber's utility if she joins that network is

$$u_i = s_i v\left(\hat{p}_i\right) + (1 - s_i) v\left(p_i\right) - r_i \ . \tag{3}$$

We assume a Hotelling-type specification for subscriber choice, and the market share of network i given the list of utilities $\{u_1, ..., u_K\}$ available from the networks is

$$s_i = \frac{1}{K} + \frac{u_i - \bar{u}_i}{2t} \ , \tag{4}$$

calls play no significant role in the analysis and are ignored. (A mobile network would just set the price for such calls equal to its cost of providing such calls in the models we present.) Similarly, fixed-to-fixed calls play no role in the analysis and are ignored.

[20] An alternative assumption is that the fixed network cannot "price discriminate" in this way, perhaps because callers do not always know which mobile network they are calling. In a 2005 market survey of 2,158 users, only 13% of FTM callers reported knowing roughly how much it cost for FTM calls to the people they most regularly contact (Ofcom, 2007, p31). Without effective price discrimination, the market failures identified in the following analysis are typically amplified. For instance, a small mobile network's termination charge has only an insignificant impact on the fixed network's average cost of providing FTM calls, and so the network has a unilateral incentive to set an extremely high termination charge, since such a charge does not cause demand for calls from the fixed network to its subscribers to fall significantly. Indeed, FTM call charges will be *above* the monopoly level (see Gans and King (2000) and Wright (2002)).

[21] Indeed the FTM termination charge and the FTM retail price have declined together over the period 2001-2005 (Ofcom, 2006b, Figure 3.38) although perhaps not one-for-one as assumed in (1): recently Ofcom (2007, para. 3.22) concluded about two-thirds of the reductions to termination charges had been passed through directly to the FTM call charge.

where

$$\bar{u}_i = \frac{1}{K-1} \sum_{j \neq i} u_j$$

is the average of the rivals' utility. Here, t is the "transport cost" parameter which represents the degree of product differentiation in the market for mobile subscribers.

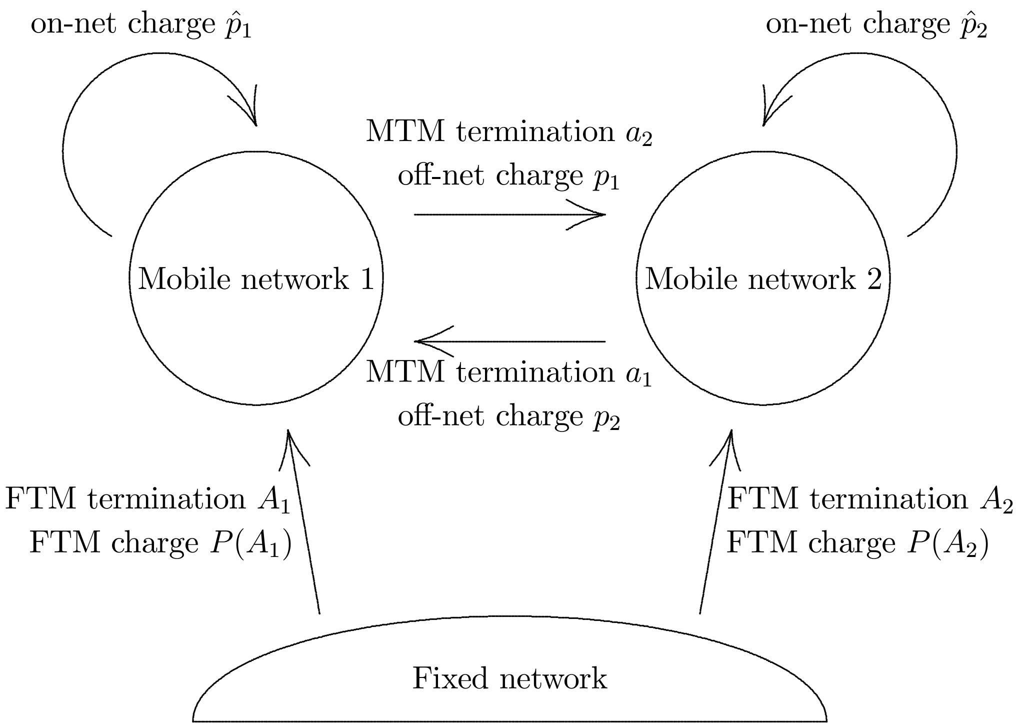

Figure 1: Call Termination on Mobile Networks

To calculate the equilibrium charges given an initial choice for the MTM termination charge a, suppose that each of $K-1$ networks chooses the same charges $(\hat{p}, p, r, A)$, and the remaining network i chooses its own charges to be $(\hat{p}_i, p_i, r_i, A_i)$. Then network i's profit is

$$\pi_i = s_i \times \Bigg[\underbrace{r_i - f + s_i \left(\hat{p}_i - c_O - c_T\right) q(\hat{p}_i) + (1-s_i)(p_i - c_O - a) q(p_i)}_{\text{retail services}} + \underbrace{(1-s_i)\left(a - c_T\right) q(p)}_{\text{MTM termination}} + \underbrace{F\left(A_i\right)}_{\text{FTM termination}} \Bigg]. \tag{5}$$

This consists of the retail profit from supplying service to its subscribers, the profit from providing termination for rival mobile networks, and the profit from providing termination for the fixed network.

3.1 Fixed-to-mobile call termination

First, consider network i's incentive to choose its FTM termination charge, A_i. The profit expression (5) shows that each network's unregulated FTM termination charge will be chosen to maximise its profits from FTM call termination, $F(\cdot)$. This is a dominant strategy for each network, regardless of choices for retail tariffs and the MTM termination charge. By setting A_i to maximize F, each firm will be able to subsidize subscribers to the maximum extent, thereby increasing market share without having to lower profit per subscriber. We denote the resulting unregulated termination charge by A_M, to indicate it is the termination charge that would be chosen by a monopoly mobile network. As discussed in the introduction, this situation in which the equilibrium termination charge with competition is equal to the monopoly charge is an example of what is sometimes known as a *competitive bottleneck*. The result does not depend on the competitiveness of the market for subscribers (as measured by t or the number of firms K). Thus, it is perfectly possible that one side of the mobile market (the market for mobile subscribers) is highly competitive, yet the other part (FTM call termination) is essentially a series of monopolies.

During the 2002 Competition Commission enquiry, *Vodafone* stated that the unregulated level was in range 17 ppm to 20 ppm. At this time, the cap on O_2 and *Vodafone*'s termination charge was 9.3 ppm, so *Vodafone* was suggesting that the FTM termination charge would roughly double without the charge control (Competition Commission, 2003, paras. 2.440–2.445). Ofcom (2007, para. 7.49), estimates that the current unregulated FTM termination charge would be about 24 ppm, compared with its current regulated charge of about 5 ppm, so the regulator believed that there would be a five-fold increase if regulation were abandoned.

Welfare (as measured by the sum of consumer surplus and profit) in the FTM segment when the termination charge is A is

$$\underbrace{V(P(A))}_{\text{consumer surplus}} \quad + \quad \underbrace{F(A)}_{\text{mobile termination profit}} \quad + \quad \underbrace{[P(A) - C - A]\, Q(P(A))}_{\text{retail profit by fixed network}} , \tag{6}$$

which simplifies to

$$V(P(A)) + [P(A) - (C + c_T)]\, Q(P(A)) .$$

As one would expect, this is maximized by setting the FTM call charge equal to the cost of such calls:

$$P(A_W) = C + c_T ,$$

where A_W denotes the welfare-maximising FTM termination charge. This ensures fixed-line callers face a FTM price equal to the true marginal cost of their calls, namely $C + c_T$. When the FTM call charge is equal to the fixed network's cost, so that (1) holds, welfare is maximized with a FTM termination charge equal to cost:

$$A_W = c_T .$$

On the other hand, if $P(A) > C + c_T$ then welfare is maximized by setting $A < c_T$ to counteract the markup present in the FTM retail charge.

3.2 Mobile-to-mobile call termination

Suppose that the mobile networks each set the FTM termination charge A. (As discussed, without regulation we expect $A = A_M$. However, regulation may impose $A = A_W$, or some other charge, in which case it is still useful to analyze the remaining parts of the mobile market.)

First, we derive the equilibrium call charges given A and a. In a symmetric equilibrium, each network will have market share $s_i = 1/K$. Therefore, from (3) and (4), firm i's market share is unchanged if it modifies its charges $(\hat{p}_i, p_i, r_i)$ in such a way that

$$\frac{1}{K}v(\hat{p}_i) + \frac{K-1}{K}v(p_i) - r_i = \text{constant} . \tag{7}$$

From (5), in a symmetric situation the network's profit in terms of $(\hat{p}_i, p_i, r_i)$ is

$$\pi_i = \frac{1}{K}\left[r_i + \frac{1}{K}(\hat{p}_i - c_O - c_T)q(\hat{p}_i) + \frac{K-1}{K}(p_i - c_O - a)q(p_i) + \text{constant}\right] .$$

It follows from (7) that $\hat{p}_i$ is chosen to maximize $v(p) + (p - c_O - c_T)q(p)$ and p_i is chosen to maximize $v(p) + (p - c_O - a)q(p)$. Therefore, in equilibrium each network will set the on-net call charge $\hat{p}$ and off-net call charge p given by

$$\hat{p} = c_O + c_T \ ; \ p = c_O + a \ , \tag{8}$$

so that call charges are equal to the respective marginal costs of making calls.

Having determined the equilibrium call charges, we complete the analysis of retail tariff decisions by considering the choice of rental charge, r. Analogously to (2), write

$$M(a) \equiv (a - c_T)q(c_O + a) \tag{9}$$

for the profit from MTM termination when the MTM termination charge is a. From (5) and the call charges in (8), network i's profit is

$$\pi_i = s_i \times [r_i - f + (1 - s_i)M(a) + F(A)] \ . \tag{10}$$

Expression (4) implies that firm i's market share s_i satisfies

$$s_i = \frac{1}{K} + \frac{1}{2t}\left[r - r_i + \frac{Ks_i - 1}{K-1}(\hat{v} - v)\right] ,$$

where $\hat{v} = v(c_O + c_T)$ and $v = (c_O + a)$. Solving this explicitly in terms of s_i implies that

$$s_i = \frac{1}{K} - \frac{r_i - r}{2t - \frac{K}{K-1}(\hat{v} - v)} \ . \tag{11}$$

Finally, substituting (11) into (10), maximizing with respect to r_i and setting $r_i = r$ shows that the equilibrium rental charge is given by

$$r = f + \frac{2t}{K} - F - \frac{K-2}{K}M - \frac{\hat{v} - v}{K-1} \ , \tag{12}$$

where the dependence of F on A and M on a has been suppressed.

Subscriber utility, which from (3) is equal to

$$u = \frac{1}{K}\hat{v} + \frac{K-1}{K}v - r,$$

is easily seen from (12) to be increasing in F. More subtly, subscriber utility is also increasing in the MTM termination charge, at least in the region around $a = c_T$. To see this, note that given (12),

$$u = \frac{K-1}{K}v + \frac{K-2}{K}M - \frac{v}{K-1} + \text{ constant},$$

which is increasing in a at $a = c_T$.

The final variable to determine is the choice of MTM termination charge, a. Substituting (12) into (10) shows that industry profit in the mobile sector, denoted Π, is

$$\Pi = \frac{1}{K}\left(2t + M - \frac{K}{K-1}(\hat{v} - v)\right) . \tag{13}$$

In particular, the FTM termination profit F has no impact on equilibrium profits in the mobile sector, and firms collectively should not object to regulatory policy which fixes A at any particular level.

By contrast, mobile networks are not indifferent to the choice of MTM termination charge. Recall that the off-net call charge p in (8) depends on a. Without regulation, the industry will choose a to maximize (13), so that

$$a = c_T + \frac{1}{K-1}\frac{q(p)}{q'(p)} < c_T \ , \tag{14}$$

which is below the cost of termination. Since the socially optimal MTM call charge, for both on-net and off-net calls, is equal to the cost $c_O + c_T$, expression (8) implies that the efficient MTM termination charge is $a = c_T$, just as with FTM termination. In particular, optimal regulatory policy treats the two termination charges symmetrically. We deduce that unregulated firms in this model will choose a MTM termination charge which is too low relative to the efficient level, in contrast to incentives concerning FTM termination.

3.3 Discussion

FTM termination: The result derived in expression (13), that profits from FTM termination are not retained by firms, is striking. For any FTM termination charge A, symmetric competition between networks to attract subscribers will ensure industry profit is unchanged. If firms can obtain high profits from terminating calls from the fixed network, this makes firms compete harder for subscribers—each subscriber brings a profit $F(A)$—and this acts to pass these profits onto the subscribers themselves. Correspondingly, if regulation squeezes out profits in FTM call termination, this will lead to price rises for mobile subscribers. Specifically, expression (12) suggests that the rental charge will rise by a factor equal to the reduction in per-subscriber FTM termination profits. This result is often termed the "waterbed" effect. In this basic model, there is a 100% waterbed effect, in that reduced profits from one source are completely clawed back from subscribers so that the overall profit impact is zero (consistent with empirical evidence provided by Andersson and Hansen (2007) and Genakos and Valletti (2007)). Alternative ways of modelling competition for subscribers might lead to less than 100% or greater than 100% waterbed effects.

Recall from section 2 that UK policy has been to allow mobile networks to recover a portion of their fixed costs via a surcharge on FTM termination charges. While such a procedure is fairly standard in setting access prices for a regulated fixed-line network, it is less clear that this policy should be applied to competitive mobile networks. If a 100% waterbed effect operates, setting higher FTM termination charges will not provide operators with any additional contribution towards their fixed and common costs, at least in the long-run, and it will simply act as an (inefficient) transfer of surplus between fixed and mobile subscribers. In such an environment, setting high termination charges purely to allow for fixed and common cost recovery would be a flawed policy.

In this model, the competitiveness of the retail mobile market has little bearing on either the firms' equilibrium choice of FTM termination charge or on the welfare costs of unregulated FTM termination charges. In our model, competitiveness is represented by the differentiation term t or the number of networks K, and these parameters do not enter into the expression for welfare in (6). It is a dominant strategy for firms to set the FTM termination charge at the monopoly level A_M, regardless of competitive conditions. Indeed, a firm would be at a competitive disadvantage if it chose a lower charge while its rivals set the monopoly charge, since its rivals would then have greater ability to fund subsidies to attract subscribers. For this reason, it would be misleading to describe the firm' monopoly pricing behaviour as "abusing their dominant position" in terminating calls on their own networks.

A feature of the waterbed effect is that unregulated monopoly profits from FTM call termination are passed onto mobile subscribers in the form of low rental charges (e.g. free or subsidized handsets). To the extent this occurs, the market failure associated with FTM termination does not lead to excessive profits by mobile networks, but rather to a sub-optimal pattern of prices. Of course, this observation should not affect the welfare analysis: if high margins on FTM call termination lead to negative margins on services to mobile subscribers, there is allocative inefficiency regardless of whether overall profits in the mobile sector are excessive or not. Relatedly, mobile firms "advanced the argument that, because most people had a mobile phone, what they lost in high termination charges they gained in low access and outbound call charges".[22] However, even if all fixed-line subscribers had a mobile phone, this argument is not correct: since high termination charges lead to allocative inefficiency, the total "size of the cake" is shrunk, and the gain from handset subsidies is *smaller* than the losses caused by high FTM call charges.[23] Nevertheless, as we will show in section 4.4, if FTM calls can substitute for MTM calls, the degree of overlap between the set of fixed-line and mobile subscribers does become relevant.

In this basic model, firms should not object to regulatory intervention to bring each firm's FTM termination charge down from the unregulated level A_M to the socially optimal

[22]See Competition Commission (2003, paras. 2.390–2.400).

[23]Ofcom (2007, para. A19.37), estimates that consumers will be better off by £3.2 billion in present value terms as a result of the current regulation in the period 2007-2011 compared to the unregulated alternative. (Ofcom's analysis assumed that competition was sufficiently strong so that excess profits were eliminated, and so "consumer" and "total welfare" standards coincide.)

level A_W. Nevertheless, mobile networks *do* typically object to such a reduction in the FTM termination charge, which suggests that this basic model fails to capture important aspects of the real-world market. (Section 4.2 considers an extension where mobile firms do collectively prefer a higher to a lower FTM termination charge.) In addition, even if networks make the same equilibrium profits with different choices of the termination charge, this does not imply that a *transition* from one equilibrium to another will be costless to firms. For instance, if networks had long-term contracts with subscribers, written at a time when termination revenues were relatively high, then if policy reduces termination revenues, firms may incur losses on these subscribers until contracts expire.

There is no reason why only large (or incumbent) mobile networks should have their FTM termination charges regulated, while small (or new) networks are free to set whatever charge they wish. A small network will set a monopoly FTM termination charge in exactly the same way as a large firm. Indeed, if incumbent firms are regulated while new firms are not, new firms have a powerful advantage over incumbent firms, in that they have greater ability to fund subsidies to attract subscribers. This asymmetric treatment of firms in the industry is likely to lead to a distorted pattern of supply. A recent illustration of what happens if some firms are regulated and some are not was presented in Table 2 above, where, on weekdays at least, *H3G* took advantage of its (then) less regulated position to set charges which were roughly double those levied by its more regulated rivals.

An implicit assumption in the basic model is that fixed-line callers alone determine the number and length of FTM calls. If in practice the number and length of FTM calls is jointly determined by the caller and receiver, then high FTM termination charges could lead mobile networks to *pay* subscribers for receiving FTM calls (rather than merely reducing rental charges), so as to stimulate FTM call volumes. For instance, in 2006 *H3G* announced that it would pay its subscribers 5 pence per minute for receiving calls, a marketing tactic which surely was motivated by its highly profitable termination charges (see Table 2). However, there are problems with the use of reception payments: in Italy, when mobile operators tried paying subscribers for calls received, they found people were calling their own mobile phones from office lines to obtain these rebates.

This last point is also helpful in understanding the difference between the "caller-pays" regime as used in the UK, in which FTM termination charges and call charges are relatively high, and a "receiver-pays" regime as used in the United States, whereby the price of a FTM call is low but mobile subscribers often incur charges for receiving calls. In the United States, reciprocity requirements imposed by the FCC mean that FTM termination charges are set equal to those for mobile-to-fixed termination, which are quite likely below the cost of mobile termination. With little termination revenue, mobile operators may choose to charge mobile subscribers directly to recover their costs and/or to induce their subscribers to discourage incoming traffic. For instance, with a charge for receiving calls, a subscriber may keep his or her phone switched off to eliminate incoming calls, something which is often socially inefficient. The key point is not that the problem of high FTM termination charges is solved by imposing a receiver-pays regime, but rather that a regulatory requirement to set low FTM termination charges may induce networks to charge their subscribers for receiving calls.

MTM termination: Why do firms wish to set a below-cost MTM termination charge—as in expression (14)—if free to do so? Unless firms set a low termination charge, call charges (as in (8) above) will be such that it is more expensive to call off-net than on-net. In such a situation, subscribers will, all else equal, prefer to join a larger network since they can then make a larger fraction of their calls at the cheaper rate. In other words, the market will exhibit (positive) network effects. As is well known, in such markets competition is particularly fierce and profits are low.[24] Firms can overturn this effect by setting a low termination charge, which results in off-net call charges which are below on-net charges. In this case, subscribers will, all else equal, prefer to join a smaller network, which acts to relax competition.

In contrast to FTM termination, though, in our model the market failure which results from an unregulated MTM termination charge is reduced when there are more mobile firms in the market. Expression (14) shows that when K is large, networks will voluntarily jointly choose a termination charge which approximates the socially optimal termination charge. We deduce that when there are a reasonable number of mobile firms in the market, the private incentives to choose the MTM termination charge are approximately in line with overall welfare.[25]

Our analysis shows that mobile subscribers benefit both from high FTM and high MTM termination charges. This observation will play an important role when we discuss market expansion issues in section 4.2. However, the reason is quite different in the two cases. Mobile subscribers benefit from high FTM charges since there is a waterbed effect at work, and profits from this source are simply passed onto mobile subscribers. High FTM termination charges are an (inefficient) means of transferring surplus from fixed callers to mobile recipients, and the profitability of mobile networks is not affected. High MTM termination charges act to intensify competition between mobile operators due to strong network effects. When the MTM termination charge is raised, this imposes a direct cost on subscribers since they must pay more for off-net calls, but this is outweighed by the lower rental charge they pay. High MTM termination charges have no impact on fixed callers, and so a high MTM termination charge acts as a means by which to transfer surplus from mobile networks to their subscribers.

As mentioned in the introduction, we are aware of no regulator which has been concerned that firms will set MTM termination charges which are unduly low. Probably the main reason for this is that due to arbitrage possibilities, mobile networks are likely to be constrained to set the same charge for terminating FTM and MTM traffic, so that $a_i = A_i$. In such cases, it is natural to suppose that unregulated firms set this termination charge unilaterally, and we analyze this model in section 4.3.

Another possible reason why existing operators prefer high MTM termination charges is that high charges may deter entry. By setting above-cost MTM termination charges, the incumbent networks can induce network effects which make entry less attractive for the newcomer. With high MTM termination charges, off-net calls will be more expensive, which

[24] See Farrell and Klemperer (2007) for an overview.

[25] Section 1.1 in Calzada and Valletti (2007) also analyzes a model with an arbitrary number of networks, but using a Logit specification for market shares instead of our Hotelling formulation (4). Although not emphasized in their paper, they appear also to find that as the number of firms increases, the industry's choice of MTM termination charge converges to the efficient level—see the proof of their Corollary 2.

particularly hurts a small network since the bulk of its subscribers' calls will be off-net. Call externalities will reinforce this effect, since when the established firms have high off-net prices, subscribers of a new (smaller) network will also receive relatively few calls.[26,27]

4 Externalities and Arbitrage

4.1 Call externalities

Here, we allow mobile subscribers to obtain benefits from receiving calls. The setting used is similar to Berger (2005), extended to take into account of FTM calls and the presence of more than two mobile firms. In so doing, we examine whether a mobile network's incentive to inflate FTM termination charges above efficient levels is constrained if subscribers care about how many calls they receive.

Suppose a mobile subscriber obtains a linear benefit $BQ + bq$ if she receives Q calls from the fixed network and q calls from the mobile networks. Here B and b measure the (possibly different) strengths of the respective call externalities. Our linear specification for these call externalities differs from the previous literature, which assumed that people obtained a diminishing marginal utility from receiving more calls. As well as simplifying the calculations, we believe that the linear specification better captures the fact that people generally do not control which calls to them are made (so each call can be taken to be a random draw from the recipient's willingness-to-pay for incoming calls). By contrast, it makes sense to have diminishing marginal utility for outbound calls, since subscribers tend to make calls in order of willingness-to-pay.[28]

With call externalities, the utility from subscribing to network i changes from (3) to

$$u_i = s_i[v(\hat{p}_i) + bq(\hat{p}_i)] + (1 - s_i)[v(p_i) + bq(p)] + BQ(P(A_i)) - r_i \ ,$$

where p is the off-net call charge on the rival networks. With this modification, the analysis presented in section 3 proceeds in the same manner. The impact of call externalities on the choice of the FTM termination charge is relatively straightforward. Without call externalities, unregulated firms would like to choose A to maximize $F(A)$, but with the externality

[26] Hoernig (2007) analyses the impact of on-net and off-net price differentials on the profitability of small networks. He shows that larger firms will choose greater differentials than smaller firms. Calzada and Valletti (2007) formally demonstrate that incumbent networks will sometimes use high MTM termination charges to deter entry.

[27] At least two further explanations have been proposed in the literature. First, Cherdron (2002) considers a setting where calling patterns are biased towards peers (closed user groups). Setting above cost MTM termination charges is a way to endogenously differentiate the networks, so that consumers prefer to stick to the network that their peers join, thereby reducing competition between operators. Second, Höffler (2006) argues in a dynamic model that higher MTM termination charges can sustain tacit collusion between firms for a wider range of discount rates.

[28] Jeon, Laffont, and Tirole (2004) and Berger (2005) model the call externality utility function as being proportional to a subscriber's utility function for making outbound calls. In these papers, as the benefits of receiving calls tend towards the benefits of making calls, the equilibrium off-net call charge becomes arbitrarily high and there is a "connectivity breakdown". In our framework, as the charge for off-net calls rises, the harm to a network's own subscribers eventually dominates the harm done to a rival's subscribers, and so there is never any such breakdown.

B, firms instead choose A to maximize

$$BQ(P(A)) + F(A) \ . \tag{15}$$

Of course, the unregulated FTM termination charge will be lower than without call externalities. If a network sets a high FTM termination charge, this will reduce the volume of calls received by its subscribers from the fixed network, and hence reduce the rental fee it can charge its subscribers if it wishes to maintain market share. In effect, the call externality B has exactly the same effect on a network's choice of FTM termination charge as a reduction in its termination cost c_T by B. Thus, as has been emphasized by the mobile networks, call externalities do indeed mitigate a network's incentive to set high FTM termination charges.

Welfare in the FTM segment when the FTM termination charge is A is

$$\underbrace{V(P(A))}_{\text{consumer surplus}} + \underbrace{BQ(P(A))}_{\text{call externality}} + \underbrace{F(A)}_{\text{mobile termination profit}} + \underbrace{[P(A) - C - A]\, Q(P(A))}_{\text{retail profit by fixed network}} ,$$

which is maximised by setting the FTM call charge equal to the "adjusted" cost of FTM calls:

$$P(\hat{A}_W) = C + c_T - B \ . \tag{16}$$

This ensures fixed-line callers face a FTM price equal to the true marginal cost of their calls, adjusted to take account of the impact of their calls on mobile recipients. When (1) holds, welfare is maximized with a FTM termination charge equal to the adjusted cost

$$\hat{A}_W = c_T - B \ . \tag{17}$$

In sum, when mobile subscribers obtain benefits from receiving calls from the fixed network, this causes the unregulated equilibrium termination charge *and* the welfare-maximizing termination charge to fall. As such, the presence of such externalities does not diminish the potential need to regulate this charge.

Consider next the impact of call externalities on the MTM termination charge. One can show that when the MTM termination charge is a, each network sets its call charges to be

$$\hat{p} = c_O + c_T - b \ ; \ p = c_O + a + \frac{1}{K-1} b \tag{18}$$

instead of (8). Thus, the on-net call charge is equal to a network's cost for the call, $c_O + c_T$, adjusted downwards to reflect the call externality its subscribers enjoy from being called more often from other people on the same network. This on-net call charge is not affected by a, and is socially efficient. The off-net call charge p is equal to a network's cost for the call, $c_O + a$, adjusted *upwards* to reflect the fact that when its subscribers make fewer calls to subscribers on rival networks, call externalities imply that its rivals' abilities to compete are harmed and this benefits the original network. This represents an anti-competitive motive to set high off-net call charges.

Notice that, even if the MTM termination charge is set equal to cost ($a = c_T$), the prices in (18) differ for on-net and off-net calls. Thus, this model predicts that observed differences in these call charges are by no means solely due to above-cost MTM termination charges. For instance, the 2005 numbers from Tables 1 and 4 above suggest that roughly $a = 6$, $p = 11$ and $\hat{p} = 4$ at that time. The difference between off-net and on-net calls predicted by the model equals $a + bK/(K-1) - c_T$, so setting this equal to the observed difference of 7 pence implies that $c_T = 4b/3 - 1$ with four mature firms. This suggests we require $b > 0.75$ to be consistent with positive termination costs. Using $c_T = 5$, roughly what Ofcom estimated the cost of mobile termination to be in 2005 (see Ofcom (2004), p64), suggests a receiver benefit of 4.5 pence per minute is required to explain the observed price differential. Thus, according to the model, it could either be that the difference between on-net and off-net call prices is explained mostly by above-cost MTM termination charges (this happens if the relevant marginal costs are actually close to zero in which case receiver benefits can be as low as 0.75 pence per minute) or that the difference between on-net and off-net call prices is explained mostly by large receiver benefits (as is the case if Ofcom's estimates of costs are accurate).

One can show that the MTM termination charge which maximizes the networks' joint profit is modified from (14) to be

$$a = c_T - \frac{K+1}{K-1}b + \frac{1}{K-1}\frac{q(p)}{q'(p)} .$$

Again, this is too low relative to the efficient MTM termination charge, since firms have an incentive to use the MTM termination charge to relax competition. However, as in section 3, with many mobile networks firms will agree to choose the approximately efficient MTM termination charge (i.e., $a \approx c_T - b$).

4.2 Network externalities and market expansion

A potentially important feature of the mobile market is that the number of mobile subscribers is not constant. Indeed, section 2 reports that the market grew dramatically in the years from 1998 to 2006. It is plausible that the UK market may now be near saturation point, with only 10% of households without access to a mobile phone, the same percentage that do not have access to a fixed-line phone (Ofcom, 2006a, p157). However, it is also possible that there remain some marginal subscribers, and that subsidizing these subscribers to have a mobile phone generates externalities to others. In a recent survey of 621 subscribers, Ofcom (2006c) reports that 12% knew at least one person who had subscribed to be mobile network for the first time in the previous year, and that on average they made 10 calls and received 9 calls per month from these new subscribers. In this section, we investigate how the analysis presented in section 3 is modified when the market expands as subscriber surplus rises.

The technical details of the analysis are relegated to the appendix, where we present a duopoly model without call externalities.[29,30] Our unified framework with both FTM and MTM calls allows us to take a first step towards understanding whether it is the FTM

[29] The presence of call externalities will amplify the impact of network externalities, since users will receive more calls when there are more mobile subscribers.

[30] The mobile-to-mobile part of the following discussion is somewhat related to section 5 of Dessein (2003), although he does not allow for off-net/on-net call charge differentials. Like us, he shows that the unregulated MTM termination charge is below cost and that the efficient MTM termination charge is above cost.

or the MTM termination charge, or both, which should contribute to subsidies for mobile subscription.

The first point is that the possibility of market expansion does not mitigate a mobile network's incentive to choose a high FTM termination charge, and their incentive is still to extract as much profit as possible from fixed callers. However, we show that market expansion possibilities suggest that the efficient FTM termination charge is above cost (unlike in the basic model). The reason is that high FTM termination profits will feed through into low tariffs for mobile subscribers via the waterbed effect, which will in turn feed through into more people deciding to subscribe to mobile networks, which will in turn benefit all users (fixed and mobile) since they have more people to call.

Nevertheless, the efficient FTM charge still lies below the unregulated level. The unregulated FTM charge is the charge which maximizes a mobile firm's profit from providing call termination to the fixed network. Starting from this profit-maximizing charge, a small reduction in the charge has a second-order impact on the profit from call termination, and hence only a second-order impact on the number of mobile subscribers. However, it has a first-order impact on the price for FTM calls. Therefore, welfare rises with a reduction in the FTM charge from the unregulated level. In sum, there remains a (reduced) rationale for regulatory control of the FTM termination charge.[31]

In addition, the waterbed effect discussed in section 3.3 is no longer 100% when market expansion is possible. In our model, profits from FTM termination are retained to some extent by mobile firms, and not fully passed onto subscribers. That is to say, the mobile industry as a whole now has an incentive to lobby against proposed regulation to bring FTM termination charges down to efficient levels. (The basic model in section 3 did not shed any light on why it was that mobile networks were opposed to industry-wide regulatory intervention to bring down termination charges.)

In their most recent analysis, Ofcom estimated that the "externality surcharge" needed to reflect the benefits of market expansion was 0.30 pence per minute, or around 5% of their estimated costs of call termination (Ofcom, 2007, p349). Their chosen methodology, following the earlier analysis of the Competition Commission (2003), was to assume that mobile networks could, at least in part, target a subsidy to new or marginal subscribers rather than to the body of existing subscribers. It is not clear exactly how this could be done in practice. In the model in the appendix, we assume that the subsidy is applied equally to all mobile subscribers, which means a substantially higher level of subsidy will be needed. Ofcom then estimates the subsidy which, when targeted, "brings about at least as much external benefit as the amount of the subsidy". Needless to say, such a calculation is sensitive to how one estimates the various elasticities and externalities.

What about the choice of MTM charge when market expansion is possible? In the appendix, we show that firms would like jointly to choose a termination charge which is below cost, just as in the basic model of section 3. However, we also show that the efficient MTM termination charge is now *above* cost. The reason is that mobile subscriber surplus is increased by a high MTM termination charge, as discussed in section 3.3: a high MTM

[31] The finding that the welfare-maximizing FTM termination charge is above cost but below the unregulated level has been noted in the existing literature (Armstrong (2002), Wright (2002), and Valletti and Houpis (2005)), but previous models have unrealistically assumed away MTM calls.

charge induces a direct cost to subscribers since they must pay more to call off-net, but in equilibrium this is outweighed by the smaller rental charge they pay.

In sum, if networks can feasibly set different termination charges for FTM and MTM traffic, without regulation networks would choose too high an FTM charge and too low a MTM charge. The socially efficient charges, though, are now both above the cost of supplying termination (and are typically set at different levels, if feasible).[32]

4.3 Uniform FTM and MTM termination charges

Up until this point, we have assumed that mobile networks negotiate a reciprocal MTM termination charge, which could freely be set at a different level from the FTM charge. However, in practice FTM and MTM charges are not set at different levels. For instance, as mentioned in the introduction, one reason why mobile networks may be forced to set a uniform charge is the ability of networks to arbitrage between significant differences between FTM and MTM termination charges at the wholesale level.[33] In addition, a network whose FTM termination charge is regulated may be unwilling to choose a lower charge for MTM termination if it suspects the regulator may use that information subsequently to tighten FTM regulation.

Nor is it clear that unregulated networks negotiate over their MTM (or uniform) termination charges with rival networks, rather than setting them unilaterally. For instance, in the 2002 Competition Commission enquiry *Vodafone* suggested that MTM termination should not be explicitly regulated, but rather that firms should be required to enter into bilateral negotiations over this termination charge (implying that this was not already happening).[34] Motivated by UK experience before regulation was introduced in 1998, in this section we assume that, in the absence of regulation, each network unilaterally sets a uniform termination charge for both FTM and MTM calls.[35]

We use the same benchmark model as in section 3. In particular, we ignore call externalities and market expansion possibilities.[36] For simplicity, suppose there are just two mobile networks. Denote network i's uniform termination charge (for both FTM and MTM traffic) by a_i. Then following the same logic as in expression (8), firm i's call charges are

$$\hat{p}_i = c_O + c_T \ ; \ p_i = c_O + a_j \ , \tag{19}$$

[32] If, as in the next section, the FTM and MTM termination charges are constrained to be equal, it follows that the efficient, uniform charge will also be above cost.

[33] In our discussions with industry representatives in various countries this was the explanation given for why mobile operators set a common termination charge for both types of calls.

[34] See Competition Commission (2003, para. 2.473).

[35] If instead the uniform termination charge was collectively set through negotiation, the analysis of section 3 (regarding the MTM charge) would apply, so that the uniform termination charge would be set below the efficient level. The reason is that, due to the waterbed effect, when firms choose the termination charge collectively, profits from FTM termination have no impact on their profits. Therefore, their incentives are exactly as if there was only MTM traffic.

[36] The analysis is related to section 3 in Gans and King (2001), section 6 in Laffont, Rey, and Tirole (1998a) and Behringer (2006), all of whom consider the unilateral setting of termination charges for MTM calls alone.

and, as in (10), its profit is

$$\pi_i = s_i \times [r_i - f + (1 - s_i)M(a_i) + F(a_i)] \ . \tag{20}$$

From (4), firm i's market share satisfies

$$s_i = \frac{1}{2} + \frac{s_i\hat{v} + (1 - s_i)v_j - r_i - [(1 - s_i)\hat{v} + s_i v_i - r_j]}{2t}$$

and so

$$s_i = \frac{1}{2} + \frac{r_j - r_i + \frac{1}{2}(v_j - v_i)}{2t + v_1 + v_2 - 2\hat{v}} \ . \tag{21}$$

(Here, $\hat{v} = v(c_O + c_T)$, $v_i \equiv v(c_0 + a_i)$ and $v_j \equiv v(c_O + a_j)$.) Differentiating (20) with respect to r_i and setting equal to zero yields

$$r_i = f - \frac{s_i}{\frac{\partial s_i}{\partial r_i}} - (1 - 2s_i)M_i - F_i = f + s_i(2t + v_1 + v_2 - 2\hat{v}) - (1 - 2s_i)M_i - F_i \ , \tag{22}$$

where we have written $M_i = M(a_i)$ and $F_i = F(a_i)$, and the second equality follows from (21). Substituting this value for r_i into (20) shows that firm i's profit is

$$\pi_i = s_i^2(2t + v_1 + v_2 - 2\hat{v} + M_i) \ . \tag{23}$$

Moreover, substituting r_i as given in (22) and the corresponding expression for r_j into (21) shows that

$$s_i = \frac{1}{2} + \frac{F_i - F_j + \frac{1}{2}(v_j - v_i)}{6t + 3(v_1 + v_2) - 6\hat{v} + 2(M_1 + M_2)} \ , \tag{24}$$

which gives the equilibrium market shares purely in terms of the two networks' termination charges, a_1 and a_2. Substituting (24) into (23), differentiating with respect to a_i and imposing symmetry ($a_1 = a_2 = a$) implies the first-order condition for the uniform, unilaterally chosen termination charge a is

$$\left.\frac{\partial \pi_i}{\partial a_i}\right|_{a_i = a_j = a} = \frac{1}{4}\left(M'(a) - q(c_O + a)\right)$$

$$+ \left[F'(a) + \frac{1}{2}q(c_O + a)\right]\left(\frac{2(t + v(c_O + a) - \hat{v}) + M(a)}{6(t + v(c_O + a) - \hat{v}) + 4M(a)}\right) = 0 \ . \tag{25}$$

We wish to compare the equilibrium termination charge given in (25) with the "monopoly" charge. Since the MTM demand function $q(\cdot)$ and the FTM demand function $Q(\cdot)$ can be different, we define the monopoly charge to be the common charge which maximizes total termination profit per subscriber; i.e., holding fixed the number of subscribers on each network. Given that there are two networks, half of the MTM calls received by each subscriber are off-net calls. Each subscriber therefore generates termination revenue of $F + \frac{1}{2}M$, and the monopoly termination charge, a_M, satisfies

$$F'(a_M) + \frac{1}{2}M'(a_M) = 0 \ . \tag{26}$$

As before, we require the networks to be sufficiently differentiated, so that network effects are not explosive. From (21), a simple sufficient condition to ensure this is that

$$2t > 2v(c_O + c_T) - v(c_O + a_M) \ . \tag{27}$$

This rules out any deviation by network i (even to the point that $v(c_O + a_i) = 0$), when the other network sets its termination charge no higher than the monopoly level a_M, as in the proposed equilibrium we find. Without such a condition, there is a possibility that network i could profitably deviate in this way. By setting a very high termination charge, network i could make off-net calls from the other network prohibitively expensive, so that under certain consumer expectations about which network people will join, everyone will subscribe to network i.

Equation (25) can be rewritten as

$$\frac{F' + \frac{1}{2}M'}{F' + \frac{1}{2}q} = \frac{t + v - \hat{v} + M}{3\,(t + v - \hat{v}) + 2M} \ , \tag{28}$$

where the dependence of F', M', M, q and v on a in (28) has been suppressed. It is straightforward to check that the left-hand side of (28) is greater than the right-hand side when $a = c_T$. In contrast, at the monopoly termination charge a_M the left-hand side equals zero (by definition), while the right-hand side is positive from (27). Therefore, there is at least one $a \in (c_T, a_M)$ where (28) holds. Moreover, under relatively mild regularity conditions, the left-hand side is decreasing and the right-hand side is increasing in a in the range (c_T, a_M).[37] Therefore, there is a unique $a \in (c_T, a_M)$ which satisfies (28).

In sum, the equilibrium termination charge is below the monopoly level but above the efficient level. When networks must set a uniform termination charge for FTM and MTM traffic, this model predicts unregulated firms will set a uniform charge which is too high relative to the efficient benchmark, and this justifies regulatory concerns that the charge is too high (rather than too low as our basic model predicted in section 3).

How close is the equilibrium charge to the efficient level? Consider for instance a numerical example in which costs are normalised to zero ($C = c_O = c_T = 0$), and the two demand functions are equal and linear, so that $q = Q = \frac{1}{2} - a$. The monopoly termination charge is $a_M = \frac{1}{4}$ (which is broadly in line with Ofcom's (2007, para. 7.49) estimate of the current unregulated FTM termination charge), while the efficient charge is zero. Condition (27) requires that $t > \frac{7}{64}$. If we set $t = \frac{1}{2}$, then the solution to (28) is approximately $a = 0.223$, which is some 10% below the monopoly charge. More generally, one can check that the right-hand side of (28) is decreasing in t when $a > c_T$. Therefore, the unique $a \in (c_T, a_M)$ which solves (28) is increasing in t. That is to say, if the retail market is more competitive, in the sense that services are closer substitutes, then the equilibrium uniform termination charge will be closer to the efficient level. This contrasts with our basic model in which the two charges were set independently: there, competitive conditions at the retail level played no role in the determination of the two unregulated termination charges.

[37] One can check that the right-hand side is increasing in a over this range provided that $M(\cdot)$ is concave in a. The left-hand side is decreasing in a over this range if the stronger condition $M'' < q'$ holds. (A sufficient condition for this second inequality to hold is that $q(\cdot)$ be weakly concave.)

Another useful observation is that if the relative importance of FTM traffic declines, in the sense that $F'(\cdot)$ in (28) is reduced over the relevant range for a, the equilibrium termination charge falls too. (The left-hand side of (28) increases with F'.) Recall from section 2 that the relative importance of FTM traffic has fallen in recent years, and so this discussion suggests that firms now have a reduced incentive to set very high termination charges compared to the period when regulation was first introduced. This is intuitive: if most termination traffic comes from the fixed network, the incentive to raise monopoly revenue from this source will dominate other incentives. In the limit where the MTM calls are negligible (so that $q = v = M = 0$), formula (28) reduces to $F' = 0$, i.e., the termination charge is chosen purely to extract monopoly profits from FTM calls.

Broadly speaking, when network i raises its uniform termination charge there are four effects: (i) its profit from supplying call termination to its mobile rivals and to the fixed network increases (just as with FTM termination in the basic model); (ii) its rival is forced to raise its off-net call charge, which places it at a competitive disadvantage compared to firm i; (iii) firm i's subscribers will receive fewer calls from the rival's subscribers and the fixed network due in part to (ii) (just as in the FTM case), which harms i's ability to compete for subscribers when call externalities are important, and (iv) it amplifies the differential pricing between on-net and off-net calls in the market which, as we argued in section 3.3, tends to intensify competition for subscribers. The equilibrium choice of termination charge will trade off these four effects. Effects (i) and (ii) together suggest that a firm will want to set its termination charge *above* the monopoly level, since (ii) gives a reason to boost the charge in addition to extracting termination profits. But (iii) and (iv) put downward pressure on termination charges.

In this model we ignored call externalities (so (iii) plays no role) and showed that the net incentive is to set the termination charge above the efficient level (equal to cost) but below the monopoly level. Thus, the temptation to extract termination profits from the fixed network and mobile rivals dominates the incentive to set a low charge in order to relax competition between networks. In contrast to the case where the termination charge was jointly chosen by the mobile networks and independently from the FTM charge (where the danger was that firms would choose too low a MTM charge), here the more intuitive danger is that the charge will be too high. Nevertheless, the fact that the FTM and MTM termination charges are locked together does mitigate a network's incentive to set monopoly termination charges. In other words, the competitive bottleneck result emphasized in section 3 is softened. The key insight here is that mobile operators will be constrained in their choice of a common FTM and MTM termination charge by the fact that setting it too high strengthens network effects, making the firms tougher rivals. To avoid this effect, a network will keep its unilaterally-chosen termination charge below the monopoly level.

One can perform the same analysis with $K \geq 2$ firms, with subscription decisions given as in (4).[38] This analysis shows that in the limit as the number of firms becomes large, the equilibrium uniform termination charge satisfies

$$F'(a) + M'(a) = 0 \ .$$

This is exactly the monopoly termination charge when there are many networks, since the

[38] Details available from the authors on request.

fraction of MTM calls which are on-net is negligible and so the monopoly charge maximizes $F(a) + M(a)$. Thus, as the number of rivals becomes large, the equilibrium charge converges to the monopoly level. This limit result suggests that effects (ii) and (iv) vanish when the market becomes more fragmented, and we are left only with (i). In this case, there is no softening of the competitive bottleneck effect with respect to FTM traffic. Moreover, the monopoly result now applies to the termination of MTM traffic as well.

Finally, if the analysis in this section is combined with the analysis of market expansion in section 4.2, it is theoretically possible for the equilibrium mobile termination charge to be too *low* compared to the welfare-maximizing level. An increase in the termination charge can directly raise the number of mobile subscribers, so that if network externalities are large enough, this could raise welfare. This does not, however, seem likely to apply to the current UK where mobile penetration is about the same as fixed-line penetration. The detailed comparison between the efficient and the unregulated termination charges when (i) market expansion is possible and (ii) networks set their termination charges unilaterally awaits further analysis.

4.4 Substitution between FTM and MTM calls

The previous literature, including this paper up to now, has largely treated FTM callers and MTM callers as though they are two separate groups. This need not be the case. Mobile subscribers who have access to a fixed line can choose between the two types of calls. For instance, a survey by Ofcom (2007, p34) found that 18% of callers reported using their mobile rather than fixed line when at home over the previous month. In this section we consider how this possibility changes firms' incentives to set termination charges.

As in the previous section, assume there are two networks, there are no call externalities, and networks set a uniform charge for terminating FTM and MTM traffic. Suppose there is full penetration of both the fixed and the mobile networks, so everyone potentially has access to both a fixed and a mobile phone to make calls.[39] When a user has access to a fixed phone, we assume that making calls to a mobile subscriber using a mobile phone is a *perfect* substitute for making the call with a fixed-line phone. Thus, in that situation the caller will use whichever method is the cheaper. Since people may have different demands for calling when they have access to a fixed-line phone (for example, when they are at home) from when they do not (for example, when they are on-the-go), we allow what we previous termed the FTM demand function $Q(\cdot)$ to denote demand in the former situation and the MTM demand function $q(\cdot)$ to denote demand in the latter situation. (The fact that people only have access to a fixed-line phone a fraction of the time is encompassed by this general specification.) To make the main points as cleanly as possible, we suppose that the cost of making calls using the two methods is the same, i.e., $C = c_O$.

Given these assumptions, a subscriber's utility from network i is modified from (3) to be

$$u_i = s_i v(\hat{p}_i) + (1 - s_i) v(p_i) - r_i + s_i V(\min\{P_i, \hat{p}_i\}) + (1 - s_i) V(\min\{P_j, p_i\}) \ . \tag{29}$$

[39] An alternative kind of substitution between fixed and mobile networks may take place at the subscription level rather than the per-call level, so that some people might give up their fixed line altogether and become a mobile-only user. We do not consider this possibility here (see Hansen (2006) for an interesting analysis).

The last two terms of (29) are new compared to (3), and represent the utility of a subscriber being able to make calls either with her fixed-line phone or her mobile phone whenever she has a choice. Previously, we ignored the possibility of such substitution and so implicitly assumed subscribers in these situations could only make these calls using their fixed-line phone. In that case, the corresponding utility $s_i V(P_i) + (1 - s_i) V(P_j)$ was independent of the network a subscriber joins, and so did not affect network i's market share and had no bearing on the equilibrium analysis. Allowing for substitution, though, implies network i's market share in (29) will depend on whether it allows subscribers to call other mobile subscribers more cheaply using its mobile network than they can do using their fixed-line phone.

Suppose that FTM calls satisfy (1), so that

$$P_i = c_O + a_i \ .$$

Suppose that in the first stage the two networks have chosen their, possibly unequal, termination charges a_1 and a_2, and are considering their choice of retail tariff. When choosing its call charges, a network must decide whether or not to undercut the prevailing FTM call charge, which will determine which phone a subscriber uses when she has a choice.

Consider the situation in which one of its subscribers wishes to make a call to another subscriber of network i, and where the caller has access to her fixed-line phone. If the caller uses the fixed-line phone, she pays $P_i = c_O + a_i$ and enjoys surplus $V(c_O + a_i)$, while the network obtains termination profit $(a_i - c_T)Q(c_O + a_i)$. Thus, the joint surplus available to the two parties is

$$V(c_O + a_i) + (a_i - c_T)Q(c_O + a_i) \ .$$

On the other hand, if the network undercuts the FTM call charge by choosing $\hat{p}_i \leq P_i$, the subscriber enjoys surplus $V(\hat{p}_i)$, the firm obtains profit $(\hat{p}_i - c_O - c_T)Q(\hat{p}_i)$ (where this profit now comes from supplying calls rather than call termination), and the available surplus is

$$V(\hat{p}_i) + (\hat{p}_i - c_O - c_T)Q(\hat{p}_i) \ .$$

If a_i is above cost c_T, the latter strategy yields the higher joint surplus, and indeed this second joint surplus is maximized by setting $\hat{p}_i = c_O + c_T$. We deduce that when $a_i \geq c_T$ firm i will set $\hat{p}_i = c_O + c_T$, with the result that its subscribers will always use their mobile phone to make calls to others on the same network, even when they have a choice of phone. On the other hand, if $a_i < c_T$, total surplus is reduced if the network strictly undercuts the (low) FTM call charge, and the network might as well set $\hat{p}_i = c_O + c_T$ in order to achieve the maximum total surplus in those situations in which the subscriber cannot use her fixed-line phone.

Consider next the off-net call charge, p_i. If the subscriber uses her fixed-line phone to call someone on the rival network, her surplus is $V(c_O + a_j)$ and network i makes nothing. If network i undercuts this FTM call charge, so that $p_i \leq P_j$, then the surplus of the subscriber is $V(p_i)$ and the network's profit is $(p_i - c_O - a_j)Q(p_i)$. The joint surplus following the latter strategy is maximized by setting $p_i = P_j$, in which case the joint surplus with the two strategies is identical. Therefore, the firm and subscriber are indifferent between the two strategies. As such, it is optimal for the network to set $p_i = c_O + a_j$ (since that also maximizes the total surplus in those situations where the subscriber can only use her mobile phone). In sum, for all choices of termination charges a_1 and a_2, it is optimal for each network to set its call charges to reflect its cost of making calls, so that (19) holds.

In those situations where $a_i \geq c_T$, it follows that subscriber utility in (29) becomes

$$u_i = s_i \left(v\left(c_O + c_T\right) + V\left(c_O + c_T\right) \right) + \left(1 - s_i\right) \left(v\left(c_O + a_j\right) + V\left(c_O + a_j\right) \right) - r_i,$$

so that from (4) firm i's market share satisfies

$$s_i = \frac{1}{2} + \frac{r_j - r_i + \frac{1}{2}\left(v_j + V_j - v_i - V_i\right)}{2t + v_1 + V_1 + v_2 + V_2 - 2(\hat{v} + \hat{V})} ,$$

where, as with v and $\hat{v}$, we define $V_i \equiv V\left(c_O + a_i\right)$, $V_j \equiv V\left(c_O + a_j\right)$ and $\hat{V} \equiv V\left(c_O + c_T\right)$.[40] In particular, relative to section 4.3, networks effects are more pronounced, and the condition (27) there now needs to be tightened to $2t > 2\hat{v} + 2\hat{V} - v(a_M) - V(a_M)$.

Network i's profit is

$$\pi_i = s_i \times \left[r_i - f + \left(1 - s_i\right) \left(M(a_i) + F\left(a_i\right) \right) \right].$$

To understand this expression, notice that (i) network i's subscribers will never use a fixed-line phone to call others on the same network, (ii) when the network's subscribers call someone on the other network, they are indifferent about using their mobile or fixed-line phone (when they have a choice), and the network makes no profit in either case, and (iii) when a subscriber on the rival network calls a network i subscriber and has the option to use their fixed-line phone to make the call, they are indifferent about which type of phone to use, and network i makes the same profit $F(a_i)$ in either event (and where they do not have this option they will use their mobile phone and network i makes profit $M(a_i)$).

Following the same steps as used in section 4.3 shows that instead of (25) the equilibrium uniform termination charge here satisfies

$$\begin{aligned} \left. \frac{\partial \pi_i}{\partial a_i} \right|_{a_i = a_j = a} &= \frac{1}{4}\left(M' + F' - q - Q\right) \\ &+ \left[\frac{q + Q}{2} \right] \left(\frac{2(t + v + V - \hat{v} - \hat{V}) + M + F}{6(t + v + V - \hat{v} - \hat{V}) + 4\left(M + F\right)} \right) = 0 , \end{aligned}$$

which can be rewritten as

$$\frac{F' + M'}{Q + q} = \frac{t + v + V - \hat{v} - \hat{V} + F + M}{3(t + v + V - \hat{v} - \hat{V}) + 2(F + M)} . \tag{30}$$

We are interested in whether this form of substitution reduces or enhances a network's incentive to set an excessive termination charge, i.e., whether the solution to (30) is below or

[40] It can be checked that firm i will not wish to set $a_i < c_T$. Such a policy would induce subscribers to replace on-net calls with FTM calls whenever possible, for which firm i will incur a termination loss. Moreover, even ignoring FTM calls, firm i will always wish to set an above-cost termination charge when it is set unilaterally.

above that in (28). The comparison is most transparent if we specialize the demand for calls somewhat. Suppose that the demand for calls is the same whether the caller is at home or on-the-go, and that each caller is at home a fraction λ of the time. In this case, $Q = \frac{\lambda}{1-\lambda}q$, and so $V = \frac{\lambda}{1-\lambda}v$ and $F = \frac{\lambda}{1-\lambda}M$. Then unlike (28), the equilibrium termination charge in (30) does not depend on λ, since all that matters is the *sum* of demands $Q + q$. In addition, expression (28) coincides with (30) when $\lambda = 0$, i.e., when there are no FTM calls. Then one can check that the left-hand side of (28) is increasing in λ in this example, while the right-hand side is decreasing with λ. It follows that the equilibrium termination which solves (28) is increasing with λ. We may deduce that the charge a which solves (30) is below the charge a which solves (28). Moreover, the larger λ is, the greater is the difference between the two.

Thus, we can conclude that demand side substitution between FTM and MTM calls lowers the equilibrium termination charge towards marginal cost. For instance, using the linear demand example of section 4.3 and assuming that $\lambda = \frac{1}{2}$, one can show that the equilibrium termination charge which solves (30) is approximately $a = 0.194$, which is some 22% below the monopoly charge (rather than the 10% which applied when there was no substitution).

In sum, the fact that consumers substitute on-net MTM calls for FTM calls whenever the former's price is lower helps to mitigate a mobile network's incentive to set high mobile termination charges but it does not eliminate it altogether. The intuitive reason for why this substitution reduces a network's incentive to set high termination charges stems from two sources. First, the volume of FTM termination traffic falls since all calls made to people on the same mobile network are made with a mobile phone, even when a fixed-line phone is available. As discussed in section 4.3, when the volume of FTM traffic falls, this will reduce the equilibrium termination charge since firms put more weight on avoiding the intense competition caused by network effects induced by high termination charges. Second, for a given termination charge, network effects are more important when substitution is possible. This is because FTM-type calls (those with demand function Q) also have on-net and off-net charge differentials now. This reinforces the first effect, since firms place still more weight on the danger of intensifying competition via high termination charges.

This discussion has assumed that the costs of making the two types of call are equal. It would be more realistic to assume that FTM calls are more efficient than MTM calls when callers have the choice of phone, so that $C < c_O$. However, unless the cost differential is very large, the previous analysis remains valid. A network still has an incentive to set an above-cost termination charge, and it still wishes to under-cut the FTM call charge by setting lower on-net MTM call charges. That is to say, a mobile firm will encourage its subscribers to use the less efficient MTM mode of communication whenever subscribers have a choice. For instance, with the call charges in Table 3 for 2001, it is plausible that mobile subscribers were making too many on-net calls relative to FTM calls. In other words, high mobile termination charges distort competition between FTM and on-net MTM calls, and this danger provides an additional benefit of regulating termination charges to equal cost.

Finally, when there is the possibility of FTM and MTM call substitution, high FTM call charges and the availability of cheap on-net MTM calls will give people an additional reason to subscribe to a mobile network, which is to avoid high FTM prices for on-net calls. This introduces another avenue by which high mobile termination charges can induce marginal subscribers to join (in addition to the two effects identified in section 4.2). This enhances the possibility that if network externalities are large enough, equilibrium termination charges will be too low.

5 Conclusions

The starting point for our analysis (section 3) was a basic framework in which we combined a model of FTM calls with a model of MTM calls, so we could simultaneously study FTM and MTM termination. Here, we assumed that these two charges could freely be set at different levels. In this setting, a competitive bottleneck exists for FTM termination, so FTM charges are set too high (at the monopoly level), while reciprocal MTM termination charges are set too low so as to relax retail competition between mobile operators. This theory gives insight into the UK experience with respect to high FTM termination charges, especially when we allow subscribers to get utility from receiving calls (section 4.1) and for network externalities (section 4.2). However, in terms of MTM termination, like the existing literature, the model incorrectly predicts that authorities should be concerned that unregulated charges will be set too low.

Section 4.3 resolves this puzzle by noting that wholesale arbitrage implies that a mobile network cannot sustain FTM termination charges significantly above the level of its MTM termination charges. Taking this into account, we provide a new analysis in which networks set a single mobile termination charge. To avoid intensifying retail competition through network effects, mobile networks limit their termination charges below the monopoly level, although termination charges remain above cost. The dichotomy that existed in the previous literature is resolved, with unregulated termination charges lying between the two earlier extremes. Section 4.4 shows that demand side substitution, in which mobile subscribers use their mobile phones to call others to avoid the high price of FTM calls, strengthens these constraints on unregulated termination charges. Assuming that FTM and MTM termination charges are equal, such substitution only arises with respect to on-net, not off-net, MTM calls. On-net MTM calls are made by a network's own subscribers, so by setting a low price for on-net MTM calls, a network can insulate such callers from the effects of its choice of high termination charges. As a result, a network's FTM termination profit become less important and the intensification of retail competition through network effects becomes more important in the choice of termination charges. Thus, with both technological arbitrage and demand side substitution, we find the incentive to set mobile termination charges above cost is mitigated, although not eliminated altogether.

These constraints on mobile termination charges rest on the principle that each mobile operator can reduce network effects by reducing its termination charge, thereby helping to soften retail competition. There are, however, several situations in which this principle may not be a very effective constraint on termination charges. First, in an extension to our model of section 4.3, we found that as the number of mobile competitors becomes large, equilibrium termination charges tend towards the monopoly level. Thus, the competitive bottleneck problem re-emerges. Second, if one network enjoys a larger subscriber base than the others, then it may prefer to strengthen network effects rather than weaken them, thereby making it harder for the remaining operators to compete. A similar consideration suggests that the threat of entry will also weaken the constraints we study on the unregulated setting of termination charges. This issue remains largely unexplored.

A quite different approach to overcoming the bottleneck present with mobile termination is to try to adjust the bargaining position of fixed-line and mobile operators so that mobile operators no longer hold all the bargaining power.[41] Key to this (without relying on the threat of regulation) would be to relax the fixed-line operator's obligation to interconnect with any mobile operator at the latter's chosen termination charge. For instance, if there were no such obligation, then with a single fixed network and multiple mobile operators, the natural outcome is that the fixed network holds most of the bargaining power. The mobile networks might then compete in a winner-take-all fashion for the right to deliver the fixed operator's FTM calls (which they will want to do since their subscribers value receiving calls), resulting in low mobile termination charges. How this might work when off-net MTM calls are also taken into account (for instance, taking into account the technological arbitrage and demand side substitution that we consider) remains an open question. It is worth noting, however, that for such a market-based mechanism to replace a regulatory approach (or one based on the threat of regulation) would require authorities be willing to let a breakdown in interconnection occur. It is clear in the UK they are not willing to do so at the present time (Ofcom, 2007, paras. 5.148-5.162).

In the coming years, regulation of the termination of voice calls in mobile networks may also be expanded to include other services. Ofcom has announced that it plans to review the market for SMS termination (Ofcom, *Wholesale SMS Termination Market Review*, 13 September 2006). Most of the issues discussed in this paper also apply to such text messages, which are an increasingly important and profitable part of the mobile market.[42] Similarly, regulation of wholesale termination charges for international roaming has also come under the regulatory spotlight recently.

We close by noting our paper also has implications for the internet, where a debate has emerged over what is termed "net neutrality". Net neutrality might be said to fail if network providers take control over who receives particular content or on what terms. This debate has arisen, in part, due to statements of broadband providers in the United States that they intend to start charging content providers to access their networks. If they do so, the competitive bottleneck story in this paper may become relevant. Consumers (who seek content) typically only sign up to one Internet Service Provider (ISP). However, content providers typically want to reach all consumers regardless of their chosen ISP. This means ISPs will not compete for content providers with respect to providing access to their customers. The result will be a distorted price structure, with content providers charged too much and too little content being supplied. As with mobile termination, regulations limiting how much ISPs can charge other networks to access their customers might then need to be considered. We plan to explore this issue in future work.

[41] Binmore and Harbord (2005) consider the bargaining process between a large fixed network and a small mobile entrant.

[42] SMS termination charges are already controlled in France and Israel, for instance.

References

Andersson, K., and B. Hansen (2007): "Network Competition: Emprical Evidence on Termination Rates and Profitability," mimeo.

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, Volume 1*, ed. by M. Cave, S. Majumdar, and I. Vogelsang. North-Holland, Amsterdam.

——— (2006): "Competition in Two-Sided Markets," *Rand Journal of Economics*, 37(3), 668–691.

Armstrong, M., and J. Wright (2007): "Two-Sided Markets, Competitive Bottlenecks and Exclusive Contracts," *Economic Theory*, 32(2), 353–380.

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

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

Binmore, K., and D. Harbord (2005): "Bargaining Over Fixed-To-Mobile Termination Rates: Countervailing Buyer Power as a Constraint on Monopoly Power," *Journal of Competition Law and Economics*, 1(3), 449–472.

Calzada, J., and T. Valletti (2007): "Network Competition and Entry Deterrence," *Economic Journal*, forthcoming.

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

Cherdron, M. (2002): "Interconnection, Termination-Based Price Discrimination, and Network Competition in a Mature Telecommunications Market," mimeo, Mannheim.

Competition-Commission (2003): *Vodafone, O2, Orange and T-Mobile: Reports on References Under Section 13 of Telecommunications Act 1984 on Charges Made by Vodafone, Orange, O2 and T-Mobile for Terminating Calls Made by Fixed and Mobile Networks.* HMSO, London.

Dessein, W. (2003): "Network Competition in Nonlinear Pricing," *Rand Journal of Economics*, 34(4), 593–611.

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.

Gans, J., and S. King (2000): "Mobile Network Competition, Customer Ignorance and Fixed-to-Mobile Call Prices," *Information Economics and Policy*, 12(4), 301–328.

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

Genakos, C., and T. Valletti (2007): "Testing the 'Waterbed' effect in Mobile Telephony," mimeo.

Hahn, J.-H. (2004): "Network Competition and Interconnection with Heterogeneous Subscribers," *International Journal of Industrial Organization*, 22(1), 611–631.

Hansen, B. (2006): "Termination Rates and Fixed Mobile Substitution," mimeo, Norwegian School of Management.

Höffler, F. (2006): "Mobile Termination and Collusion, Revisited," mimeo, Bonn.

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

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

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

——— (1998b): "Network Competition: II. Price Discrimination," *Rand Journal of Economics*, 29(1), 38–56.

MMC (1999): *Cellnet and Vodafone: A Report on a Reference Under Section 13 of Telecommunications Act 1984 on the Charges Made by Cellnet and Vodafone for Terminating Calls from Fixed-Line Networks*. HMSO, London.

Ofcom (2004): *Wholesale Mobile Voice Call Termination: Statement.* Office of Communications, London.

——— (2006a): *The Communications Market 2006.* Office of Communications, London.

——— (2006b): *Mobile Call Termination: Market Review.* Office of Communications, London.

——— (2006c): *Mobile Call Termination: Report of Market Research Findings.* Office of Communications, London.

——— (2007): *Mobile Call Termination: Statement.* Office of Communications, London.

Valletti, T., and G. Houpis (2005): "Mobile Termination: What is the 'Right' Charge?," *Journal of Regulatory Economics*, 28(3), 235–258.

Wright, J. (2002): "Access Pricing under Competition: An Application to Cellular Networks," *Journal of Industrial Economics*, 50(3), 289–316.

Technical Appendix: Network Externalities and Market Expansion

Suppose there are two networks and call externalities are absent. If networks 1 and 2 offer subscribers the respective utilities u_1 and u_2, suppose that network i attracts $n_i = \phi(u_i, u_j)$ subscribers. Unlike the specification in (4), the total number of mobile subscribers, denoted $N \equiv n_1 + n_2$, is not constant. Suppose network i offers the two-part tariff with on-net call charge $\hat{p}_i$, off-net call charge p_i and fixed rental charge r_i. Since mobile subscribers have more people they can call when the mobile market expands, subscriber utility at firm i is modified from (3) to be

$$u_i = v_0 + n_i v(\hat{p}_i) + n_j v(p_i) - r_i \ . \tag{31}$$

Here, v_0 is a subscriber's utility from other mobile services, most notably from calls to the fixed and international network. (This parameter plays no role when the market size is constant and was ignored elsewhere in the paper.) Firm i's profit is modified from (5) to be

$$\pi_i = n_i \times \left[r_i - f + n_i \left(\hat{p}_i - c_O - c_T\right) q(\hat{p}_i) + n_j (p_i - c_O - a) q(p_i) + n_j \left(a - c_T\right) q(p_j) + F\left(A_i\right)\right] \ .$$

It is immediate that each firm will set its FTM termination charge A_i to maximise its profits from termination, $F(\cdot)$. The potential for market expansion has no impact on a mobile network's incentive to set a high FTM termination charge, since even with a fixed market size the termination charge was set at its profit-maximizing level.

When the MTM termination charge is a, equilibrium call charges still reflect calling costs, so that (8) holds. Therefore, firm i's profit is

$$\pi_i = n_i \left[r_i - f + n_j M(a) + F(A)\right] \ . \tag{32}$$

In order to make further progress, it seems necessary to specify subscriber demand ϕ explicitly. To that end, suppose that (4) is modified in the following way: if the two mobile networks offer utilities u_1 and u_2 then firm i attracts

$$n_i = \frac{1}{2} + \frac{u_i - u_j}{2t} + \lambda u_i \tag{33}$$

subscribers. (This model of consumer demand is sometimes known as the "Hotelling model with hinterlands".) Here, $\lambda \geq 0$ represents the magnitude of the market expansion possibilities. In order to ensure market expansion is non-explosive λ cannot be too large, and the following condition is assumed:

$$2\lambda\hat{v}\left(1 + \lambda t\right) < 1 \ , \tag{34}$$

where, as in section 3, $\hat{v} = v(c_O + c_T)$.

Combining (31) with (33) shows that

$$N = \frac{1 + \lambda(2v_0 - r_1 - r_2)}{1 - \lambda(\hat{v} + v)} \ ; \quad n_i = \frac{N}{2} + \frac{\frac{1}{2}(1 + \lambda t)(r_j - r_i)}{t - (1 + \lambda t)(\hat{v} - v)} \ , \tag{35}$$

where $v = v(c_O + a)$ as before. (From (34), the denominators in (35) are positive so long as a is not too far away from cost c_T.) Finally, write

$$\gamma_1 = -\frac{\partial n_i}{\partial r_i} = \frac{\frac{1}{2}\lambda}{1 - \lambda(\hat{v} + v)} + \frac{\frac{1}{2}(1 + \lambda t)}{t - (1 + \lambda t)(\hat{v} - v)} > 0$$

$$\gamma_2 = \frac{\partial n_j}{\partial r_i} = -\frac{\frac{1}{2}\lambda}{1-\lambda(\hat{v}+v)} + \frac{\frac{1}{2}(1+\lambda t)}{t-(1+\lambda t)(\hat{v}-v)} < \gamma_1$$

for the own and cross-price elasticities of subscriber numbers with respect to the rental charge. Here γ_1 is positive, γ_2 is smaller than γ_1, and γ_2 is positive when a is close to c_T and (34) holds. Clearly γ_1 and γ_2 depend on a via v.

From (32) the equilibrium rental charge is given by

$$r = f - \frac{N}{2}M - F + \frac{N}{2}\Lambda \ , \tag{36}$$

where

$$\Lambda(a) \equiv \frac{1+\gamma_2 M}{\gamma_1} \ .$$

Substituting the value for r in (36) into the formula for profit in (32) shows that industry profit is

$$\Pi = \frac{1}{2}N^2\Lambda \ , \tag{37}$$

while substituting the value for r in (36) into expression (35) shows that the market size given a and A is[43]

$$N = \frac{1+2\lambda(F+v_0-f)}{1-\lambda(M+\hat{v}+v-\Lambda)} \ . \tag{38}$$

Since N in (38) is increasing with F, it follows from (37) that the mobile sector's profits are higher with a higher FTM termination profit, $F(A)$, and the waterbed effect emphasized in section 3.3 is no longer complete.

The impact on profit of the MTM termination charge is more complex, since a affects the elasticity terms γ_1 and γ_2 in Λ. For simplicity, consider the impact of a small change in a away from marginal cost c_T. One can show that

$$\Lambda'(c_T) = -q(c_O+a)\left[\frac{1-2\hat{v}\lambda-2t\hat{v}\lambda^2}{1+2t\lambda-2\hat{v}\lambda-2t\hat{v}\lambda^2}\right] < 0 \ ,$$

where $\Lambda' < 0$ follows from (34). From (38) we see that

$$N_a \equiv \left.\frac{\partial N}{\partial a}\right|_{a=c_T} = -\frac{\lambda N\Lambda'}{1+\frac{\lambda}{\gamma_1}-2\lambda\hat{v}} > 0 \ ,$$

and so raising the MTM termination charge above cost induces network expansion. From (37) we have

$$\left.\frac{\partial \Pi}{\partial a}\right|_{a=c_T} = NN_a\Lambda + \frac{1}{2}N^2\Lambda' = N^2\Lambda'\left[\frac{1}{2} - \frac{\frac{\lambda}{\gamma_1}}{1+\frac{\lambda}{\gamma_1}-2\lambda\hat{v}}\right] < 0 \ ,$$

[43] From (34), the denominator in the above is positive for a close to c_T. The numerator is positive when a is close to c_T provided we make the additional assumption that fixed cost is not too great in the sense $f < v_0+\hat{v}$. This inequality essentially requires that mobile subscribers gain some positive utility at marginal-cost prices (when $N \approx 1$).

where the inequality follows (after some manipulation) from (34). Therefore, mobile networks wish to set a below-cost MTM termination charge in order to relax competition for subscribers (just as in the basic model).

What are the socially efficient termination charges in this framework? When subscriber utility is $u_1 = u_2 = u$, aggregate consumer surplus of mobile subscribers is $\Phi(u) = \lambda u^2 + u$, where $u = v_0 + \frac{1}{2}N(\hat{v} + v) - r$. (So $\Phi'(u) \equiv N$.) In equilibrium, r is given by (36). The consumer surplus of fixed-line subscribers from calls to mobile subscribers is NV, where $V = V(P(A))$ is the consumer surplus on the fixed network for each mobile subscriber. Assume that FTM calls are charged at cost, so that expression (1) holds.[44] Then total welfare is obtained by summing mobile sector profit, mobile subscriber surplus and fixed caller surplus, which is

$$W = NV + \frac{1}{2}N^2\Lambda + \Phi(v_0 + \frac{1}{2}N(v + \hat{v}) - (f - \frac{N}{2}M - F + \frac{N}{2}\Lambda)) \ . \tag{39}$$

When there are no market expansion possibilities (i.e., $\lambda = 0$), it is optimal to set both the FTM and MTM termination charges equal to cost c_T as in our benchmark model. When $\lambda > 0$ it is socially optimal to set both termination charges above cost. To see this for the FTM termination charge, differentiate (39) with respect to A to obtain

$$\begin{aligned} \left.\frac{\partial W}{\partial A}\right|_{A=a=c_T} &= -NQ + N_A\hat{V} + N\Lambda N_A + N\left\{\hat{v}N_A - \frac{\Lambda}{2}N_A + Q\right\} \\ &= N_A[\hat{V} + N\hat{v} + \frac{1}{2}N\Lambda] > 0 \ , \end{aligned} \tag{40}$$

where $\hat{V} = V(C + c_T)$ and $N_A > 0$ is the derivative of N in (38) with respect to A when $A = a = c_T$. Thus, raising the FTM termination charge above cost induces a network expansion, and this benefits both fixed-line callers and existing mobile subscribers (and also the mobile networks).

To see the impact of raising the MTM termination charge, differentiate (39) to obtain:

$$\begin{aligned} \left.\frac{\partial W}{\partial a}\right|_{A=a=c_T} &= N_a\hat{V} + [NN_a\Lambda + \frac{1}{2}N^2\Lambda'] + N\left\{N_a\hat{v} - \frac{1}{2}N_a\Lambda - \frac{1}{2}N\Lambda'\right\} \\ &= N_a[\hat{V} + N\hat{v} + \frac{1}{2}N\Lambda] > 0 \ , \end{aligned}$$

which is the same form as (40). Thus, setting the MTM termination charge above cost benefits both mobile and fixed-line subscribers and overall welfare, but it harms the mobile networks.

[44] Assume also that the mobile-to-fixed termination charge is regulated to be equal to the fixed network's cost. Otherwise, the extra profits on the fixed network caused by terminating more traffic when the mobile market expands would need to be considered too.