A Model-based Study of the Impact of Managed Services and the Spawning of Applications in Broadband Networks

Debasis Mitra and Qiong Wang
Bell Labs, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, NJ 07974
{mitra,qwang}@research.bell-labs.com

***Abstract*— Whether to allow a broadband service provider to supplement basic best-effort service by high quality managed service at a premium price has been hotly debated, especially in the context of network neutrality. Offering managed service provides delay-sensitive customers with a better option of guaranteed quality of service, but may allow the service provider to starve the best-effort service of bandwidth so as to force best-effort users to subscribe to more expensive managed service. We study this issue through models in which the library of applications available to all broadband users is key to usage and decisions made by users and service provider, and, importantly, for investigating how the innovations process leading to new network applications, affects, and is affected by, the service provider's bandwidth allocation. First, we hold the number of applications fixed and potential users weigh utility and cost to decide whether to subscribe to the broadband network, and if so, whether to use best-effort service applications that are free but associated with congestion-dependent delay or to pay a per-use fee to use managed service. Likewise, the service provider provisions bandwidth for the two services and sets the per-use fee for managed service to maximize its profit. Next, we model innovations that spawn new applications to originate from use of best-effort service. Cognizant of the separation of time constants in application generation and bandwidth provisioning, we assume that the service provider undertakes a sequence of myopic optimizations, in each of which it takes the number of applications as fixed by the preceding decision. We find that the innovations process serves as a stabilizer that keeps bandwidth allocation in balance: with fewer applications, the provider will sponsor more best-effort service and more usage leads to more new applications. When the number of applications becomes excessively large, the provider is induced to cut back on bandwidth for best-effort service to provide more managed service.**

I. Introduction

We consider a population of N potential subscribers to broadband services. The broadband service provider offers subscribers the use of a library of applications supported by best-effort as well as managed, i.e., premium, services. There is a fixed fee that users pay to the service provider to subscribe to broadband service. The determination of this fee is regulated and outside the control of the broadband service provider. Best-effort service is provided freely to every broadband service subscriber. In contrast, there is a fee for each use of a managed service application. In return, the subscriber receives guaranteed quality of service, for instance, without any delay that may be caused by congestion. In this case, the broadband service provider sets the per-use fee for managed service. It also provisions the required bandwidth for quality of service in the use of each managed service application. The service provider undertakes an optimization exercise to maximize its profit in determining the fee and the bandwidth allocation. Correspondingly, users have decisions to make, first, whether to subscribe to the broadband service, and, second, whether to use individual managed service applications. Users too make their decisions based on the outcome of optimizations to maximize their individual benefits that take into account utility and costs.

The broadband service provider controls the allocation of the given, fixed network bandwidth B to the two services. It guarantees bandwidth for each subscriber's individual managed service application. The bandwidth allocation to the two services is constrained, i.e., the sum of bandwidth allocated to best-effort and managed services cannot exceed B. Note that the delay perceived by best-effort service users is based on sharing of the allocated bandwidth by all best-effort users, and this delay is a factor in the self-optimization performed by users and consequently in the determination of best-effort service usage.

Whether a regulator should allow service providers to offer managed service is a focal point in the debate on network neutrality. The central issue is price discrimination, which is a double-edged sword that can be

used to promote or damage customer interest and social efficiency [8]. For users of heterogeneous applications, offering a menu of different quality of service levels to meet different requirements of individual applications can improve welfare for every party involved, provided that a suitable pricing schedule is implemented in conjunction [1]. For users of the same application, offering different services facilitates Ramsey pricing, which, by charging a lower price to those with higher demand elasticity, allocates the common cost of the network infrastructure in a socially efficient way [2]. However, it is questionable that such end outcome will emerge when broadband service is offered by a profit-maximizing monopoly provider. Of particular concern is the "damaged goods" approach, i.e., the provider intentionally reduces quality of service of best-effort service, or even refuse to offer the service at all [4], to force customers to choose more expensive managed service [8]. Profit is maximized, but network resource that could be used to serve more customers is wasted, which ultimately translates into a loss of consumer surplus and social welfare.

Beyond the immediate consequences of price discrimination, there are longer-term factors that may enhance or weaken the desirability of allowing managed service. Some argue that service differentiation gives the provider incentive to invest more in network infrastructure ([6], [7]). Others consider the broad impact of offering managed service in a two-sided market, and examine if it will shift the burden of paying for the network investment from end users to content providers [3], or will be used by service providers to extract excess revenue from content providers [6] and grant unfair advantage to its own content business [9]. Moreover, debates also take place regarding whether offering different services and charging non-zero prices will prompt or obstruct innovations that spawn new applications and contents on the Internet ([5], [10]). The last subject is of particular interest of this paper.

We study the relationship between offering managed service and innovations in the creation of new applications in the network, albeit in a distinct setting. Rather than discussing the impact of allowing managed service on the innovation process, we develop a modeling framework to examine how innovations in the broadband networks affect the service provider's incentive of offering different services, especially with regard to its bandwidth allocation. Important aspects of the overall model here are the mechanisms that affect innovations, and usage of best-effort and managed service applications. The intuition here is two-fold. First, increased best-effort service usage promotes the spawning, i.e., birth, of new applications. Second, increased lack of quality of service (also called (dis)quality here and typified by delay) discourages use of best-effort and promotes increased use of managed service applications. Note the implication that over time reduced usage of best-effort service leads to a lower rate of spawning of new applications.

We find that innovations of new applications play an important role in balancing service offering when managed service is allowed into the broadband network, even when the service provider only myopically optimizes its profit. With few applications, the marginal benefit of allocating more bandwidth to the managed service is small, inducing the provider to offer more best-effort service, and the reverse is true if the network hosts an excessive number of applications. Since new applications are spawned from the usage of best-effort service, the end result is that innovations of application serve as a stablizer that prevent best-effort service from total collapsing.

The paper is organized as follows. Section 2 considers customers' self-optimizations regarding whether to subscribe to the broadband service; and, if so, which of two services to use and by what amount. Section 3 discusses service provider's profit optimization with respect to the bandwidth allocation to the two services and the usage price for managed service applications. Section 4 considers the emergence of new applications over time that are spawned by the use of best-effort service. The section discusses steady state behavior of the bandwidth provisioned by the service provider under myopic profit-maximization in response to the varying number of applications. The analysis identifies self-stabilizing mechanisms that allow the co-existence of both managed and best-effort services.

II. Self Optimizations

A. *User self-optimization: best-effort service*

We assume that all individuals have the same utility from using an application. Let ω be an individual's utility per unit of time of using an application, which is common for all applications. For each application, let λ be the rate, i.e., frequency, of using an application and $\bar{\lambda}$ be a nominal usage rate. The utility u of using the application increases with frequency and we will use the following representative function:

$$u(\lambda) = \omega \ln\left(1 + \frac{\lambda}{\bar{\lambda}}\right).$$

Customers differ in their opportunity cost of the delay in using network applications. We combine customers' personalized cost of per-unit delay and preferences in a

single index θ. In general, θ follows some distribution. Here we assume the distribution is uniform over interval $\left[\underline{\theta}, \bar{\theta}\right]$. Also, for notational convenience, we normalize the population N to unity.

The (dis)quality of service of the best-effort service is captured by variable D, a QoS-related performance measure, such as the mean delay. Delay D is experienced by best-effort users and avoided by managed-service users. For given D, every individual with index θ chooses her usage level of best-effort service to maximize surplus, which is defined as the excess of utility over delay cost

$$v^{BE}(\lambda, \theta) = u(\lambda) - \theta\lambda D, \tag{1}$$

which decreases as θ increases. Let $[\underline{\theta}, \theta_b]$ be the range of all customers who use the best-effort service. For $\theta \in [\underline{\theta}, \theta_b]$, the optimal usage level is

$$\lambda^{BE*}(\theta) = \frac{\omega}{\theta D} - \bar{\lambda} \tag{2}$$

and the optimal surplus is

$$v^{BE*}(\theta) = \omega\left(\ln\frac{\omega}{\bar{\lambda}\theta D} - 1 + \frac{\bar{\lambda}\theta D}{\omega}\right).$$

Let B^{BE} be total bandwidth allocated in the broadband network by the provider to the best effort service. Let A be the number of applications, which is assumed to be fixed here. We assume that (dis)quality of service is determined by the formula for the mean delay in $M/M/1$ systems (however our main insights do not depend on this specific form of the delay function),

$$D\left(B^{BE}, \theta_b\right) = \left(B^{BE} - \Lambda(\theta_b)\right)^{-1} \tag{3}$$

where

$$\begin{aligned}\Lambda(\theta_b) &= A\int_{\underline{\theta}}^{\theta_b} \lambda^{BE*}(\theta)\, d\theta \\ &= A\left(\frac{\omega}{D}\ln\left(\frac{\theta_b}{\underline{\theta}}\right) - \bar{\lambda}(\theta_b - \underline{\theta})\right)\end{aligned} \tag{4}$$

Observe that $\lambda^*(\theta) > 0$ for all $\theta \le \theta_b$. Thus $\Lambda \ge 0$. For given θ_b, D can be determined from (3) and (4) as

$$D\left(B^{BE}, \theta_b\right) = \frac{\omega\ln(\theta_b/\underline{\theta}) + 1/A}{B^{BE}/A + \bar{\lambda}(\theta_b - \underline{\theta})}. \tag{5}$$

The above expression shows immediately that (dis)quality D decreases as bandwidth for the best-effort service B^{BE} increases. Moreover, by (2), $\lambda^{BE*}(\theta_b) > 0$ implies that $\theta_b < \omega/\left(\bar{\lambda}D\right)$, so

$$\begin{aligned}\frac{\partial D}{\partial \theta_b} &= \left[B^{BE}/A + \bar{\lambda}(\theta_b - \underline{\theta})\right]^{-1}\left(\frac{\omega}{\theta_b} - \bar{\lambda}D\right) \\ &> 0,\end{aligned} \tag{6}$$

i.e., (dis)quality increases with the number of users of best-effort service.

The total surplus that a type θ customer derives from the best effort service is obtained by aggregating $v^{BE*}(\theta)$ over all applications,

$$S^{BE}(\theta, D) = \omega A\left[\ln\frac{\omega}{\bar{\lambda}\theta D} - 1 + \frac{\bar{\lambda}\theta D}{\omega}\right]. \tag{7}$$

B. *User self-optimization: managed service*

Now consider the managed service. The broadband service provider provisions average bandwidth b^{MS} for the usage of each application, such that users are not affected by congestion-related (dis)quality. The broadband provider charges managed-service customers a per-use fee p. Responding to this price, a customer of managed service chooses the usage level of an application to maximize her surplus

$$v^{MS}(\lambda) = \omega\ln\left(1 + \frac{\lambda}{\bar{\lambda}}\right) - p\lambda. \tag{8}$$

Note the differences from the analogous expression in (1) for best-effort service: the additional cost of the per-use fee and the compensating benefit from the elimination of congestion-dependent delay D. This yields the optimal usage

$$\lambda^{MS*}(p) = \frac{\omega}{p} - \bar{\lambda} \tag{9}$$

and optimal surplus

$$v^{MS*}(p) = \omega\left(\ln\left(\frac{\omega}{\bar{\lambda}p}\right) - 1 + \frac{\bar{\lambda}p}{\omega}\right).$$

The total surplus that a customer derives from using the managed service is obtained by aggregating over all applications,

$$S^{MS}(p) = \omega A\left[\ln\frac{\omega}{\bar{\lambda}p} - 1 + \bar{\lambda}\frac{p}{\omega}\right]. \tag{10}$$

This expression should be compared to the analogous expression in (7) for best-effort service. Note that the total surpluses for the two services are equal when $p = \theta D$. Also, observe that customers' tolerance to delay, which is parameterized by θ, does not play a role in determining the usage level and surplus of the managed service because quality of service is guaranteed.

C. *User self-optimization: subscription to broadband service*

Suppose customers of both managed and best-effort services are charged a subscription fee for broadband service at the rate s. The fee is regulated and outside the control of the service provider.

To maximize surplus, a customer of type θ chooses the best-effort service if and only if

$$S^{BE}(\theta, D) \ge \max\left(s, S^{MS}(p)\right), \tag{11}$$

chooses the managed service if and only if

$$S^{MS}(p) \geq \max\left(s, S^{BE}(\theta, D)\right), \tag{12}$$

and chooses not to subscribe to broadband service if and only if

$$\max\left(S^{MS}(p), S^{BE}(\theta, D)\right) < s. \tag{13}$$

Let θ_b be the customers index that satisfy

$$\theta_b = \max\left\{\frac{p}{D\left(B^{BE}, \theta_b\right)}, \underline{\theta}\right\}.$$

From (5) and (6), $D\left(B^{BE}, \theta_b\right)$ increases in θ_b and $D\left(B^{BE}, \underline{\theta}\right) = 1/B^{BE}$. Therefore, given p, if B^{BE} exceeds the minimum value of $\underline{\theta}/p$, then there exists a unique value of $\theta_b > \underline{\theta}$ and consequently there will be customers subscribing to best-effort service. Otherwise, $\theta_b = \underline{\theta}$, the service provider may as well set $B^{BE} = 0$. In the discussion below, we show that the optimal p^* is independent of B^{BE}. Hence for there to be best-effort service subscribers, $B^{BE} \geq \underline{\theta}/p^*$.

From the customer choice model (11), all broadband subscribers of type $\theta \in [\underline{\theta}, \theta_b]$ prefer the best-effort service to the managed service. All other customers, type $\theta > \theta_b$ prefer to subscribe to managed service if and only if

$$S^{MS}(p) \geq s.$$

The service provider may not have enough bandwidth to serve all these customers, in which case it has to ration the managed service. Without loss of generality, we assume customers of types $\theta \in (\theta_b, \theta_s]$ will get the managed service, where θ_s ($\theta_b < \theta_s \leq \bar{\theta}$) is determined by the bandwidth constraint.

III. Service Provider's Profit Maximization

A. Profit maximization for a given number of applications

The broadband provider offers both best-effort and managed services. The provider sets the usage fee for managed service p and bandwidth for the best-effort service B^{BE} to maximize its profit $\Pi\left(p, B^{BE}\right)$, which includes contributions from the usage fee for the managed service applications and the subscription fee from customers of both services.

$$\Pi\left(p, B^{BE}\right) \equiv pA\int_{\theta_b}^{\theta_s} \lambda^{MS*}(p)\, d\theta + s\left(\theta_s - \underline{\theta}\right),$$

which simplifies to

$$\Pi\left(p, B^{BE}\right) = \left(\theta_s - \theta_b\right)\left(pA\lambda^{MS*}(p) + s\right) + s\left(\theta_b - \underline{\theta}\right) \tag{14}$$

The provider's choice of p and B^{BE} are subject to following constraints: the bandwidth constraint

$$b^{MS}A\int_{\theta_b}^{\theta_s} \lambda^{MS*}(p)\, d\theta + B^{BE} \leq B,$$

which simplifies to

$$b^{MS}A\lambda^{MS*}(p)\left(\theta_s - \theta_b\right) + B^{BE} \leq B; \tag{15}$$

the constraint that defines θ_b

$$\theta_b D\left(B^{BE}, \theta_b\right) = p; \tag{16}$$

and the constraint that the surplus of managed service customers has to exceed the subscription fee s,

$$S^{MS}(p) \geq s. \tag{17}$$

Furthermore we require that

$$\frac{\bar{\lambda}\underline{\theta}}{\omega} \leq B^{BE} \leq B.$$

The lower bound above establishes the minimum bandwidth requirement for customers who are least sensitive to delay ($\theta = \underline{\theta}$) to use best-effort service ($\lambda^{BE*}(\underline{\theta}) \geq 0$) if it is the only service offered. Recall that $\omega > \bar{\lambda}p^*$, so the bound is also a necessary condition for satisfying the aforementioned limit $B^{BE} > \underline{\theta}/p^*$ under which some subscribers will choose best-effort over managed service.

Define

$$f(p) \equiv p + \frac{s}{A\left(\omega/p - \bar{\lambda}\right)} = p\left(1 + \frac{s}{A\left(\omega - p\bar{\lambda}\right)}\right). \tag{18}$$

Recall that p is the per-use fee of managed service applications and $\omega/p - \bar{\lambda}$ is the usage rate per application (see (9). Thus $f(p)$ represents the profit generated from each use of a managed service application, including the distribution of the fixed rate of the subscription fee s over all such uses. Inserting the bandwidth constraint (15) into the profit function (14), we arrive at

$$\Pi\left(p, B^{BE}\right) = \frac{B - B^{BE} - \Delta}{b^{MS}} f(p) + s\left(\theta_b - \underline{\theta}\right), \tag{19}$$

where $\Delta \in [0, B - B^{BE})$ is the slack variable of constraint (15). It is immediate from the expression that to maximize $\Pi\left(p, B^{BE}\right)$, $\Delta = 0$. Since $f(p)$ increases in p, and θ_b increases in p by (6) and (16), it is optimal to set p to p^*, the maximum value that satisfies (17) at equality. Using (10) for $S^{MS}(p)$,

$$p^* = \frac{\omega}{\bar{\lambda}}x \tag{20}$$

where x is the unique solution of

$$-\ln x - 1 + x = \frac{s}{\omega A},\ x \in (0, 1). \tag{21}$$

In summary,

Proposition 1: The profit maximizing service provider sets the per-use fee of managed service applications p to p^*, where p^* is uniquely defined by (20) and (21).

One may easily verify that p^* increases in both A and ω and decreases in s, i.e., the service provider should charge subscribers of the managed service a higher usage fee if there are more applications and/or each application delivers higher utility. On the other hand, to attract them to subscribe, the usage fee needs to be reduced in the presence of a higher subscription fee.

With p^* fixed, the service provider's profit maximizing problem in (19) reduces to the choice of B^{BE} that maximizes

$$\Pi\left(B^{BE}\right) = \frac{B - B^{BE}}{b^{MS}} f\left(p^*\right) + s\left(\theta_b\left(B^{BE}\right) - \underline{\theta}\right), \tag{22}$$

where θ^{BE} is related to B^{BE} through (16).

Lemma 2: $\Pi\left(B^{BE}\right)$ is strictly concave in B^{BE}.

The proof is in the Appendix.

The lemma allows us to identify the optimal solution as follows. Observe that when the provider offers only managed service, $B^{BE} = 0$, in which case

$$\Pi\left(0\right) = \frac{B}{b^{MS}} f\left(p^*\right).$$

When the provider offers only best-effort service, $B^{BE} = B$.

When the provider offers both services, the bandwidth required for best-effort service should be within the range of $(\underline{\theta}/p^*, B)$, and the optimal value B^{BE*} solves

$$\frac{d\Pi\left(B^{BE}\right)}{dB^{BE}} = -\frac{f\left(p^*\right)}{b^{MS}} + s\frac{d\theta_b}{dB^{BE}} = 0. \tag{23}$$

This derivative naturally separates into two parts. The first part, $f\left(p^*\right)/b^{MS}$, represents the incremental revenue that can be generated by allocating more bandwidth to the managed service, while the second part, $sd\theta_b/dB^{BE}$, represents the marginal revenue of provisioning bandwidth to the best-effort service. From the concavity result established in Lemma 2, (23) has a unique solution if and only if

$$\begin{aligned} s\frac{d\theta_b}{dB^{BE}} &> \frac{f\left(p^*\right)}{b^{MS}} \text{ at } B^{BE} = \frac{\bar{\lambda}\underline{\theta}}{\omega} \\ \text{and } s\frac{d\theta_b}{dB^{BE}} &< \frac{f\left(p^*\right)}{b^{MS}} \text{ at } B^{BE} = B \end{aligned} \tag{24}$$

Otherwise, $\Pi\left(B^{BE}\right)$ is optimized by an extreme solution in which only one service is offered by the service provider.

Proposition 3: The service provider maximizes its profit by a) provisioning bandwidth B^{BE*} and $B - B^{BE*}$ for best-effort and managed services respectively if and only if (24) and

$$\Pi(B^{BE*}) > \Pi\left(0\right);$$

b) by provisioning all bandwidth B for the best-effort service only if and only if

$$s\frac{d\theta_b}{dB^{BE}} > \frac{f\left(p^*\right)}{b^{MS}} \text{ at } B^{BE} = B \text{ and } \Pi\left(B\right) > \Pi\left(0\right),$$

and, c) by provisioning all bandwidth B for the managed service only in all other cases.

IV. The Equilibrium Number of Applications

A. Dynamics of the number of applications

The discussion above assumes a fixed number of applications A. However, a broadband network is constantly changing, where new applications are born while old ones fade away. New applications are typically spawned and trialed in the domain of best-effort service and then penetrate into managed service when the need for quality guarantees arises. We model the dynamics of the number of applications thus,

$$\frac{dA}{dt} = \gamma\Lambda\left(\theta_b\right) - \mu A. \tag{25}$$

Applications enter the system at a birth rate that is proportional to the use of best effort service ($\Lambda\left(\theta_b\right)$) and fall into disuse at a constant rate μ. The equilibrium is reached when

$$\frac{dA}{dt} = 0, \text{ i.e., } \Lambda\left(\theta_b\right) = \frac{\mu}{\gamma}A. \tag{26}$$

To maximize long-run profit, a far-sighted service provider should keep (25) in mind when provisioning the bandwidth B^{BE} for best-effort service. That is, the provider should take into account the impact that the decision has on $\Lambda\left(\theta_b\right)$ and thus on the equilibrium number of applications, A. This is because the number of applications has significant impact on the profit.

However, changes in the number of applications normally occur much more slowly than the time for a provider to update its provisioning decision. Hence it is reasonable, even more realistic, to assume that the service provider optimizes the provisioning decision myopically, i.e., it takes A as given input in each optimization problem (in a sequence of optimizations), rather than as a function of the decision variable in the global optimization.

Hence under myopic optimization, the service provider responds to the current number of applications by allocating bandwidth to maximize its profit. This decision affects the use of the best-effort service, causing the number of applications to change. Reacting to that

change, the provider updates its provisioning decision, and the process evolves into an equilibrium where (26) holds under the optimized θ_b. An important question to ask is whether the process leads to an extreme case at the equilibrium, i.e., whether the provider offers only one service. The lemma below suggests the answer is no.

Lemma 4: Let

$$\begin{aligned} g(A) &= \frac{f(p(A))}{b^{MS}},\ h(A) = s\frac{d\theta_b(A)}{dB^{BE}}, \\ \text{and } M(A) &= g(A) - h(A) \end{aligned}$$

where $p(A)$ is the solution to (20) and $\theta_b(A)$ is given by (16) with B^{BE} given and $p = p(A)$. Then for any given B^{BE},

$$\lim_{A\to+\infty} \frac{dM(A)}{dA} > 0 \tag{27}$$

and

$$\lim_{A\to 0} \frac{dM(A)}{dA} > 0 \tag{28}$$

The proof is in the Appendix.

Recall that $M(A)$ is the marginal benefit of taking additional bandwidth away from the best-effort service and using it to offer managed services. In the above lemma, condition (27) shows that in the presence of a large number of applications, any further increase of new applications will cause the marginal benefit to increase, inducing the service provider to shift bandwidth to managed service, slowing down the growth of new applications. On the other hand, when the network has only a small number of applications, see (28), any decrease in the number of new applications results in the loss of the marginal benefit for provisioning bandwidth to managed service, making it more profitable for the service provider to offer more bandwidth to best-effort service. In this sense, the birth and death of new applications works as an automatic stabilizer that not only keeps the number of applications from exploding or collapsing, but also prevent the service provider from going extreme when allocating bandwidth to the two services.

Observe that for the stabilizer to function, there is no need for purposeful management of new applications. The service provider does not even have to be aware of the long-run profit implications of the number of applications. All is assumed is myopic profit maximization. When there is a large number of applications, to fit into the bandwidth budget for the managed service, the provider needs to set a high per-use price to curb the usage of each individual application. Any addition of bandwidth will be quickly filled by suppressed demand without much change in price, making allocating bandwidth to managed service more profitable. On the other hand, if there are fewer applications, each application will be used more intensively. So to create new usage of managed service, the provider has to reduce the per-use price more substantially, yielding a smaller return from allocating new bandwidth to managed service.

V. Conclusion and Future Works

This paper represents our first steps towards understanding the role of applications in the determination of customers' utility from using best-effort and managed services and also the service provider's profit. Also taking shape here is the understanding of the role of the generation of new applications to sustain a growing, dynamic network. The service provider's considerations for sustainable profit must extend to understanding these complex issues, especially the spawning of new applications.

Our model implies that the service provider's profit goals are served by the stable presence of both best-effort and managed services. Fortuitously, we find that best-effort service will survive myopic profit maximization by the service provider even without network neutrality requirements. However, this does not guarantee the provider will always provision sufficient bandwidth for best-effort service. To address the latter issue, we need to extend our analysis to identify equilibrium point(s) at which the number of network applications stabilizes, examine the process of convergence towards equilibrium and assess resulting social welfare and consumer surplus. To extend the applicability of our conclusions, we also need to generalize our model, in particular, to cover cases of non-homogeneity in applications' bandwidth and users' utility in using the applications.

From the regulator's perspective, the leading question is whether the service provider's profit maximixing incentive will be detrimental to social welfare and consumer surplus, specially due to inadequate provisioning of bandwidth to best-effort service. The analysis here provides a modeling framework and useful insights to address these questions.

References

[1] R. Cocchi, S. Shenker, D. Estrin, and L. Zhang (1991), A study of priority pricing in multiple service class networks, *IEEE Transaction on Networking,* vol 1. No 6, pp. 614-627.

[2] N. Economides and B. E. Hermalin (2012), The economics of network neutrality, *Rand Journal of Economics*, to appear.

[3] N. Economides and J. Tag (2012), Network neutrality on the Internet: a two-sided market analysis, Information Economics and Policy, 24 (2012), pp. 91-104.

[4] R. T. B. Ma and V. Misra (2011), The Public Option: A non-regulatory alternative to Network Neutrality, *Proceedings of 2011 ACM Conference on Emerging network experiment and technology (CoNEXT 2011),* Tokyo, Japan, December, 2011.

[5] R. Lee and T. Wu (2009), Subsidizing creativity through network design: zero-pricing and net neutrality, vol 23, no 3, Summer 2009, pp. 61-76.

[6] J. Musacchiio, G. Schwartz, and J. Walrand (2009), A two-sided market analysis of providers investment incentives with an application to the net neutrality issues, *Review of Network Economics,* vol 8 (1), pp. 3

[7] P. Njoroge, A. Ozdaglar, N. Stier-Moses, and G.Weintraub (2012), Investment in Two-sided Markets and the Net Neutrality Debate, *Columbia Business School working paper.*

[8] J. M. Peha (2006), The benefits and risks of mandating network neutrality and the quest for balanced policy proofs, *34th Telecommunications Policy Research Conference,* Arlington, VA, September 2006.

[9] J. M. Peha, W. M. Lear, and S. Wilkie (2007), The State of the Debate on Network Neutrality, International Journal of Communication, 1 (2007), pp. 709-716.

[10] C. S. Yoo (2010), Network neutrality or Internet innovation, *Telecommunications and Technology*, Spring 2010, pp. 22-29.

VI. Appendix: Proofs

Proof of Lemma 2

We want to show that

$$\frac{d\theta_b^2}{d\left(B^{BE}\right)^2} < 0.$$

From (16),

$$\left(D + \theta_b \frac{\partial D}{\partial \theta_b}\right) \frac{d\theta_b}{dB^{BE}} = -\theta_b \frac{\partial D}{\partial B^{BE}}. \tag{29}$$

Following (5) and (6),

$$\begin{aligned} \frac{\partial D}{\partial B^{BE}} &= -\frac{D}{A}\left(\frac{B^{BE}}{A} + \bar{\lambda}\left(\theta_b - \underline{\theta}\right)\right)^{-1} \\ \text{and } \frac{\partial D}{\partial \theta_b} &= \left(\frac{\omega}{\theta_b} - \bar{\lambda} D\right)\left(\frac{B^{BE}}{A} + \bar{\lambda}\left(\theta_b - \underline{\theta}\right)\right)^{-1} \end{aligned} \tag{30}$$

Applying the above to (29),

$$\frac{d\theta_b}{dB^{BE}} = \frac{\theta_b}{A\left(\omega/D + B^{BE}/A - \bar{\lambda}\underline{\theta}\right)}. \tag{31}$$

Observe that $d\theta_b/dB^{BE} > 0$ because (2) shows that for $\lambda^{BE*}(\theta_b) > 0$, $\omega/D > \bar{\lambda}\theta_b > \bar{\lambda}\underline{\theta}$. It follows that

$$\begin{aligned} & \frac{d^2\theta_b}{\left(dB^{BE}\right)^2} \\ = & \frac{d\theta_b/dB^{BE}\left(\omega/D + B^{BE}/A - \bar{\lambda}\underline{\theta}\right) - \theta_b/A}{A\left(\omega/D + B^{BE}/A - \bar{\lambda}\underline{\theta}\right)^2} \\ & + \frac{\omega\theta_b\left(\frac{\partial D}{\partial B^{BE}} + \frac{\partial D}{\partial \theta_b}\frac{d\theta_b}{dB^{BE}}\right)}{A\left(\omega/D + B^{BE}/A - \bar{\lambda}\underline{\theta}\right)^2} \\ = & \frac{\omega\theta_b\left(\frac{\partial D}{\partial B^{BE}} + \frac{\partial D}{\partial \theta_b}\frac{d\theta_b}{dB^{BE}}\right)}{A\left(\omega/D + B^{BE}/A - \bar{\lambda}\underline{\theta}\right)^2}. \end{aligned}$$

Therefore we only need to prove that

$$\frac{\partial D}{\partial B^{BE}} + \frac{\partial D}{\partial \theta_b}\frac{d\theta_b}{dB^{BE}} < 0$$

which is true because from (30) and (31),

$$\begin{aligned} & A\left(\frac{B^{BE}}{A} + \bar{\lambda}\left(\theta_b - \underline{\theta}\right)\right)\left(\frac{\partial D}{\partial B^{BE}} + \frac{\partial D}{\partial \theta_b}\frac{d\theta_b}{dB^{BE}}\right) \\ = & -D + \frac{\theta_b\left(\omega/\theta_b - \bar{\lambda}D\right)}{\omega/D + \left(B^{BE}/A - \bar{\lambda}\underline{\theta}\right)} \\ = & -\frac{B^{BE}/A + \bar{\lambda}\left(\theta_b - \underline{\theta}\right)}{\omega/D + \left(B^{BE}/A - \bar{\lambda}\underline{\theta}\right)} D \\ > & 0. \end{aligned}$$

Proof of Lemma 4

To prepare for the proof of (27) and (28), we first derive some relevant quantities. Applying (20) to (21),

$$-\omega \ln \frac{\bar{\lambda}p}{\omega} - \omega + \bar{\lambda}p = s/A, \tag{32}$$

and thus

$$\frac{dp}{dA} = \frac{sp}{A^2\left(\omega - \bar{\lambda}p\right)}. \tag{33}$$

To satisfy (32),

$$\lim_{A\to 0} p = 0 \text{ and } \lim_{A\to+\infty}\left(\omega - \bar{\lambda}p\right) = 0.$$

Therefore

$$\lim_{A\to 0} \frac{1}{A\left(\omega - \bar{\lambda}p\right)} = +\infty,$$

and by L'hopital's rule and (33),

$$\begin{aligned} \lim_{A\to+\infty} \frac{1}{A\left(\omega - \bar{\lambda}p\right)} &= \lim_{A\to+\infty} \frac{A^{-2}/\bar{\lambda}}{dp/dA} \\ &= \lim_{A\to+\infty} \frac{\left(\omega - \bar{\lambda}p\right)}{sp\bar{\lambda}} = 0. \end{aligned} \tag{34}$$

From the definition of θ_b in (16),

$$\frac{d\theta_b}{dA} = \left(D + \theta_b \frac{\partial D}{\partial \theta_b}\right)^{-1}\left(\frac{dp}{dA} - \theta_b \frac{\partial D}{\partial A}\right). \tag{35}$$

Following (30),

$$\frac{\partial D}{\partial \theta_b} = \left(\frac{\omega}{\theta_b} - \bar{\lambda}D\right)\left(\frac{B^{BE}}{A} + \bar{\lambda}\left(\theta_b - \underline{\theta}\right)\right)^{-1},$$

and using (5)

$$\frac{\partial D}{\partial A} = \frac{B^{BE}D - 1}{AB^{BE} + A^2\bar{\lambda}\left(\theta_b - \underline{\theta}\right)}. \tag{36}$$

From the definition of $g(A)$ and (18),

$$\begin{aligned} \frac{dg}{dA} &= \frac{1}{b^{MS}}\left[f'(p)\frac{dp}{dA} - \frac{sp}{A^2\left(\omega - \bar{\lambda}p\right)}\right] \\ &= \frac{\omega}{b^{MS}}\frac{1}{A\left(\omega - \bar{\lambda}p\right)^2}\frac{dp}{dA} \end{aligned} \tag{37}$$

and from the definition of $h(A)$ and using (31) for $d\theta_b/dB^{BE}$ and D for θ_b/p,

$$\begin{aligned}\frac{dh}{dA} &= \kappa^{-2}\left(\left(B^{BE}-\bar{\lambda}\theta A\right)\frac{d\theta_b}{dA}\right.\\ &\quad\left.-\theta_b\left(\frac{\omega}{D}-\bar{\lambda}\underline{\theta}\right)+\frac{A\omega}{D^2}\frac{dp}{dA}\right)\\ \text{where } \kappa &= s\left(B^{BE}+A\left(\omega\frac{\theta_b}{p}-\bar{\lambda}\underline{\theta}\right)\right). \quad (38)\end{aligned}$$

To prove (27), observe that (37) shows that $dg(A)\,dA > 0$. Thus we only need to show

$$\lim_{A\to+\infty}\frac{dh}{dA}<0,$$

which is true because in (38), $\omega, \theta_b, \bar{\lambda}, \underline{\theta}$ are finite quantities, by (34) and (38)

$$\lim_{A\to+\infty}A\frac{dp}{dA}=\lim_{A\to+\infty}\frac{sp}{A\left(\omega-\bar{\lambda}p\right)}=0$$

and similarly by (36), (35), (38), and (34),

$$\lim_{A\to+\infty}A\frac{d\theta_b}{dA}=0.$$

To prove (28), observe that as $A\to 0$, $dp/dA\to+\infty$, and the conclusions from from that dg/dA in (37) is on the order of $(dp/dA)/A$ and dh/dA in (33) is on the order of dp/dA (using (35) for $d\theta_b/dA$.)