Emerging Internet Content and Service Providers' Relationships: Models and Analyses of Engineering, Business and Policy Impact

Debasis Mitra
Columbia Univesity
debasismitra@columbia.edu

Qiong Wang
University of Illinois at Urbana-Champaign
qwang04@illinois.edu

Ao Hong
Columbia University and Google
ah3209@columbia.edu

***Abstract*—:** We study engineering and business relationships between Content Providers and broadband access ISPs in various organizational and policy environments. We focus on pricing and capacity decisions for bandwidth and caches for delivery of the CP's content over the "last mile" of the ISP's infrastructure. We model the CP-ISP interaction by the Stackelberg "leader-follower" game and the Integrated Operations model. We consider cases where premium bandwidth is offered to subscribers of the CP's service over the last mile, and cases where this is prohibited by Net Neutrality regulations. We develop a uniform solution procedure for all four resulting models. We explore the connections between optimal bandwidth and cache deployments, and, together with fees, their impact on the number of users, and related business and policy topics. We show that the decrease in profitability of caching due to Net Neutrality regulations is greater than the decrease from Integrated Operations to the Stackelberg game. In the Stackelberg game we prove that if a certain condition is satisfied, then with Net Neutrality the ISP will increase the cache price so that it is unprofitable for the CP to use caches. Moreover, this condition is satisfied in a typical case studied in detail.

I. Introduction

A. Evolution of Relationships

Netflix is exemplary of innovative Content Providers (CP) considered in this paper. The history of the evolution of the company until about 2012 is well covered in [1]. The company had traditionally relied on Amazon's AWS cloud service for content storage, and CDNs, such as Level 3, Akamai and Limelight. Since 2013 the company has evolved considerably. It has become a global leader in providing original video streaming content, while making substantial progress in engineering, and business relationships with ISPs. Netflix has developed and deployed the pioneering Open Connect caching system in collaboration with ISPs ([2],[3],[4]). The relationships between CPs and access ISPs had until recently been fraught with pricing disputes accompanied by user-perceived deterioration of service. However, these relationships have evolved from adversarial to cooperative, and customers have benefitted [5]. Moreover, the new model is being generally adopted in the industry ([1],[6]).

This paper takes note of these important developments, and offers models of various relationships between Content Providers and broadband access Internet Service Providers (ISP), some that are emerging and others that may emerge in the near future. The models integrate (i) essential elements of the engineering of content delivery systems, particularly content caching and bandwidth allocation, with (ii) business decisions made by the CPs and access ISPs, notably in pricing (for customers and CP-ISP transactions) and system sizing, (iii) users' subscription decisions based on Quality of Experience (QoE) and price, and, finally, (iv) the impact of regulations, notably Net Neutrality, and anti-trust laws, on incremental bandwidth allocations and caching, and, more generally, on investments in the network, and also on integrated operations.

B. High-Level Description of Modeled System

We assume that potential subscribers to the CP's service are already subscribers to the ISP's broadband service, and the CP's subscription fee is additional to the ISP's fee. The decision to subscribe is made based on the content, as well as the QoE in viewing and the price.

It is helpful to review the primary factors that impact the QoE of streaming video (see, for instance, [7]). These are the join time, i.e., initial time to fill the playback buffer, buffering ratio, i.e., fraction of total session time spent in filling playback buffer, the rate of buffering events and the average bitrate. These factors are all affected by the "effective bandwidth" received by the user, or, equivalently, its proportional inverse, the mean delay perceived by the user. In modeling this quantity it is useful to employ the following abstraction wherein the path from the content to the user comprises the Internet backbone, i.e. core, and the "last mile", i.e., edge, respectively the domains of transport and access ISPs. Owing to DWDM, ROADMs and big routers, etc., the total bandwidth of the Internet core dominates that of any individual "last mile". However, the allocation to an individual user is quite a different matter, on account of the multiplexing of the core bandwidth by the much larger number of users, many of which are very big.

Two mechanisms are considered in the paper for enhancing the CP's subscriber's QoE, namely, enhanced effective bandwidth over the last mile and caches located at or close to the ISP's "last mile" facilities. The improvement in QoE from caching depends on the size of the cache, which determines the "hit" probability that the customer's content selection is in the cache. In the event of a hit, the content traverses only the last mile; in the event of a "miss", the content traverses the length of the backbone, which is accompanied by delay due to the longer distance and, importantly, the combined delay due to buffering, scheduling and processing at various intermediate Internet backbone nodes. That is, the lower mean delay, or, equivalently, higher effective bandwidth, on the last mile compared to the backbone translates to the gain in QoE from caching.

In the paper CP-ISP interactions are based on the Stackelberg "leader-follower" model, which is well-known in economics and game theory [8]. In our application the ISP is the leader, and it sets unit prices for the extra bandwidth and caches. The CP is the follower and price-taker, and it responds by doing its profit maximization with respect to the extra bandwidth and cache that it will contract from the ISP, and the subscription fee for its service. Finally, the ISP does its own profit maximization with respect to the unit prices, and in so doing it incorporates knowledge of the CP's response policy. These are the unit prices used by the CP.

The mantle of leader falls on the ISP quite reasonably since it owns the last mile infrastructure, which gives it a stronger bargaining position with the CP. The ISP's price-setting takes into account that its prices for the CP if set too high will lead to lower demand.

C. Modeling the Impact of Net Neutrality and Anti-Trust Regulations

In recent years in the USA, Europe and also countries such as Australia, India and Brazil, Net Neutrality has been a topic of great interest. A key recent development has been the FCC's Feb 26, 2015 announcement of a new three-point "Bright Line Rules" for the Internet, and the application of Title II regulations [9]. Given recent history, it is reasonable to conclude that the regulatory and legal environment will remain dynamic, evolutionary and uncertain. (For an excellent discussion of these uncertainties, see [10]). It is hoped that the work presented will inform upcoming debates in the field.

The alternative scenarios that we consider are in two separate dimensions, namely, (i) Non-Net Neutrality (NNN) or Net Neutrality (NN), and (ii) Non-Integrated Operations (NIO) or Integrated Operations (IO). In (i) the distinction is whether or not subscribers to the CP's service are allowed extra bandwidth in excess of the allocation to subscribers to the ISP's basic broadband service only, over the ISP's access network.

In the second dimension, (ii) above, the distinction is whether the CP and ISP cooperate, or not, in its business decisions on pricing and resource allocations. With Integrated Operations, the CP and ISP jointly maximize their combined (not individual) profit. In particular, the price-setting by the ISP and price-taking by the CP for bandwidth and cache are circumvented.

The 2013 report of the FCC's Specialized Services Working Group [10] offers two possible paths by which the ISP could enable extra bandwidth to be available to the CP's specialized service. First, "An access ISP might set aside capacity separate from the BIAS service to carry the traffic for the third-party services that are using it." (BIAS stands for Broadband Internet Access Service.) The second path "assumes the access ISP agrees to implement prioritization of the OTT service's traffic amongst all the BIAS traffic, but only if a given customer elects to have that prioritization of their traffic."

It is an open question as to whether the caches interconnected to the ISP's access network would be a violation of Net Neutrality regulations. We have taken the arguable position here that it would not be a violation. This is because the Net Neutrality debate has focused on the last mile, and, while regulators have promised to look closely at interconnection issues, they have by and large left it untouched. (A counter-argument may quite easily be inferred from the work here, which shows that the two issues are intertwined.)

In the Integrated Operations (IO) case there is the matter of distribution of the combined revenue to the CP and ISP. We do not consider it here, in part because it is a large subject on its own, and well-known methods exist, e.g., the Shapley Value method [11]. In the analysis here, Non-Net Neutrality and Non-Integrated Operations, NNN-NIO, is the base case, and the other three models (NN-NIO, NNN-IO and NN-IO) are obtained as special cases.

D. Main Contributions

Models: a) A basic model of the content delivery engineering system is developed to capture essential features on which this study is focused. The essential features are the incremental bandwidth over the ISP-owned "last mile" for the CP's service, and the caches that hold the CP's content at a point of interconnection with the ISP's facility. The key design variables are the incremental bandwidth and the cache size. b) Users' decision-making on whether to subscribe to the service is modeled as a combination of the perceived QoE, which is a function of the effective data rate, and willingness-to-pay the CP's fee. c) A "Special Case" of the general model of user behavior combines elements of prior work from different sources; it is used to illustrate the general results and provide the base for the numerical investigations. d) The base case of the business relationship is modeled as a Stackelberg game with the ISP as the leader and price setter, and the CP as the follower and price taker, with each maximizing their individual profit. In an alternative model, the business relationship is such that the two enterprises integrate their operations with the objective of maximizing their combined profit. e) Net Neutrality regulations are modeled as allowing or not allowing incremental bandwidth over the last mile as an optional feature of the service. In summary four model variants are developed and analyzed in the paper.

Analyses and Results: a) The behavior of the users, the CP and the ISP are obtained from solutions of constrained optimization problems in which the objective functions are profits. b) For the most general model, the NNN-NIO case, we obtain a set of four nonlinear algebraic equations which completely characterize the stationary points of the optimization problems if these occur in the interior of the feasibility region. c) In the NN-NIO case in which no incremental bandwidth over the last mile is allowed, but caching is allowed, we show that if the function describing the user's QoE satisfies a certain condition, then the ISP's profit maximization leads it to raise the unit price for the cache to a point where the CP finds it unprofitable to use caches. Moreover, we show that in the Special Case, the aforementioned condition is satisfied by the QoE function. This important result shows that even though incremental bandwidth and caching may appear to be independent, prohibiting the former has the effect of curbing the latter. d) In contrast, in the NNN-IO case, with everything unchanged except that the CP and ISP integrate operations, we show that a profit maximizing solution does exist in which caching is used.

Numerical Results: a) We solve the four nonlinear equations to obtain the optimum solutions for the Stackelberg game in the NNN-NIO case for price elasticity in the 1.15- 1.35 range. b) The results show significant trends in changes in price, incremental bandwidth, QoE and the number of subscribers as the elasticity changes, whereas the cache size and hit probability remain approximately constant. We give an explanation for the high QoE, high price strategy at low elasticities, and also the near constancy in the cache size. c) The results for the NNN-NIO case also display well defined maxima at stationary points as the incremental bandwidth, hit probability and unit price for bandwidth are varied. We provide an explanation. d) We obtain results for Integrated Operations in both the NNN-IO and NN-IO cases as price elasticity is varied. In the NNN-IO case, gains over Non-Integrated Operations lie in the 20-30% range. The loss in combined profits in the NN case compared to NNN is in the 70-80% range. There are also striking differences in price and the number of subscribers. e) With elasticity fixed at mid-range, in the NN-NIO case the numerical date provides confirmation of the theory that the ISP' profit incentive causes it to raise the price of cache to a point where it becomes unprofitable for the CP to use the cache. The data also illuminates this important point: there are points in the NN-NIO case where the combined profit of the CP and ISP is as high as the highest such profit in the NNN-NIO case, but the ISPs profit incentive prevents it from being realized.

E. Prior Work

There has been considerable past work on the design and performance of caching in computing and information retrieval. See, for instance, early work by Serpanos et. al. [12], which we use, and more recent work on caching for Content Delivery Networks, with focus on algorithms and the effectiveness of caching [13] and deployment [14]. Price optimization for CDNs is treated in [15]. Our economic modeling relies on price elasticity of demand [8]. A special case of particular interest is when the price elasticity is constant, i.e., independent of price. There has been prior work in network economics that has modeled Internet demand by constant price elasticity; for examples see [16] and [17] and references therein. The reader will find in [17] examples from the communication and electricity power industries in which multi-year data support a constant price elasticity of demand in excess of unity.

Turning to the impact of QoE of users, research by Reichl et. al. [18] and the ITU [19] have shown that the user's QoE may be modeled by a logarithmic function.

The "Special Case" of this paper, see II-B, uses the expression for the cache hit rate probability in Serpanos et al. [12], the logarithmic function in Reichl et. al. [18] for the users' QoE, and constant price elasticity of demand in excess of unity for modeling users' response to price. Although labelled "special", this case has wide applicability, we believe.

Prior work that is closest in spirit to the work here is that of Maille and Tuffin [20]. The paper considers interactions between users, ISPs and CPs, as here, and also CDNs. Importantly, it considers integration of the ISP and CDN. Also, Net Neutrality issues are considered. Major differences exist in the modeling of the demand, the bandwidth allocation, its impact on QoE, the formulation as a Stackelberg game, and the analysis of the game, which form the core of this paper.

II. Models

A. System description and user decisions

Consider a typical subscriber to the ISP's broadband access service, see Figure 1. Content is typically located at a distance from the user, and it traverses the length of the Internet backbone to the ISP's facilities with effective bandwidth r_2 (recall that this quantity is inversely proportional to the mean delay). The effective bandwidth over the access network, i.e., the last mile, is r_1. As argued in Sec. I-B, $r_2 < r_1$. The final user-perceived effective bandwidth, $r_{\text{eff}} = \min(r_1, r_2) = r_2$.

Now suppose that the subscribers to the CP's service get the enhanced effective bandwidth over the last mile of $(r_1 + \beta)$ $(\beta > 0)$. Note that the user-perceived effective bandwidth is unchanged at r_2.

Next consider the case in which the CP has negotiated with the ISP for (i) as above, an effective bandwidth enhancement of β over the last mile for subscribers to its service, and (ii) the installation of a content cache of size S co-located with the ISP's facilities. Let Σ denote the size of the total content, or repertoire, of this CP. Also let $h(S, \Sigma)$ $(0 \leq h \leq 1)$, denote the "hit probability", i.e., the likelihood that the cache contains the contents of the typical subscriber's selection.

The effective bandwidth enhancement and the hit probability have the potential to significantly influence the user-perceived effective bandwidth, which is now,

$$r_{\text{eff}} = (r_1 + \beta)h + r_2(1 - h) \tag{1}$$

In the rest of the paper, following Serpanos et. al. [12], we let $S \geq 1$, and

$$h(S, \Sigma) = \log S / \log \Sigma \tag{2}$$

Define $Q(\beta, S)$ to be the QoE received by subscribers to the CP's service, which depends on both β and S through r_{eff}. We assume that

$$Q_\beta > 0, \;\; Q_S > 0, \;\text{ and } Q_{SS} < 0, \tag{3}$$

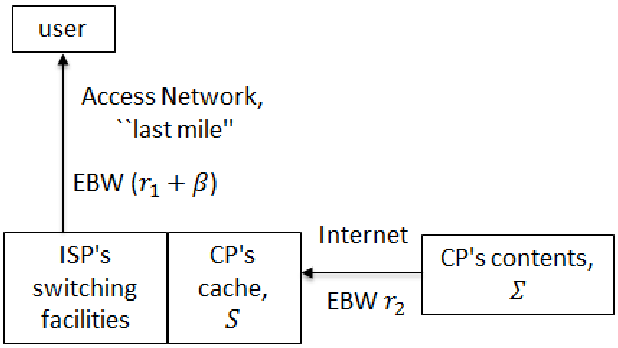


Fig. 1. Schematic of System Investigated. EBW: Effective Bandwidth

where, generally, Q_x denotes the partial derivative of Q with respect to x, and Q_{xy} the partial derivative with respect to x and y.

The number of subscribers to the content service, n, depends on β and S, and also on the subscription fee for the service p. We assume a product-form structure for the function $n(p, \beta, S)$,

$$n(p, \beta, S) = D(p)Q(\beta, S). \tag{4}$$

The function $D(p)$ gives the dependence on price p, and the function $Q(\beta, S)$ gives the dependence on QoE. We assume that D monotonically decreases with p, i.e.,

$$D_p \triangleq \partial D/\partial p < 0,$$

and define the price elasticity of demand,

$$\epsilon(p) \triangleq -\frac{p}{D}\frac{\partial D}{\partial p}. \tag{5}$$

We assume that $\epsilon(p) > 1$ for all $p \geq 0$, and that it is non-decreasing in p.

We define price elasticity of elasticity of demand, a non-standard variable,

$$E(p) \triangleq \frac{p}{\epsilon(p)}\frac{\partial \epsilon(p)}{\partial p}.$$

Since $\epsilon(p)$ is non-decreasing in p, $E(p) \geq 0$.

The three arguments in the function $n()$ comprise the complete set of decision variables for the CP, namely, (i) p, subscription fee, (ii) β, the incremental bandwidth for each subscriber, and (iii) S, the size of the cache that it will install.

B. Special Case

A special case on which we base our illustrations and numerical investigations is both important and natural. It is characterized by a logarithmic QoE function and constant price elasticity of demand.

First, following Reichl et al. [18], subscribers to the CP's service receive QoE

$$Q(\beta, S) = \log(r_{\text{eff}}/r_0). \tag{6}$$

where r_0 is a positive normalizing constant. In this case,

$$Q_\beta = \frac{1}{r_{\text{eff}}}\frac{\log S}{\log \Sigma} \geq 0; \;\; Q_S = \frac{r_1 + \beta - r_2}{S r_{\text{eff}} \log \Sigma} > 0;$$

$$Q_{SS} = -(r_1 + \beta - r_2)\frac{r_{\text{eff}} \log \Sigma + r_1 + \beta - r_2}{(S r_{\text{eff}} \log \Sigma)^2} < 0.$$

Second, the demand function takes the form $D(p) = Ap^{-\epsilon}$, where $\epsilon (> 1)$ is a constant, and $E(p) = 0$. Thus this is a special case of the above formulation. A well-known fact [8] about this demand function is that the profit-maximizing price can be obtained by simply attaching a mark-up $\frac{1}{\epsilon-1}\%$ to the marginal cost, i.e., the optimal price is the product of the marginal cost and the constant $\epsilon/(\epsilon - 1)$.

C. The Stackelberg Model

We formulate the Stackelberg model [8] with the ISP as the leader and the CP as the follower. The CP's problem is to maximize its profit Π_{CP} with respect to decision variables p, β, S, i.e.,

$$\max\{\Pi_{\text{CP}}(p, \beta, S | t_\beta, t_S)\}. \tag{7}$$

Here t_β, t_S are parameters, and

$$\begin{aligned} & \Pi_{\text{CP}}(p, \beta, S | t_\beta, t_S) \\ = \;& (p - c)n(p, \beta, S) - t_\beta \beta n(p, \beta, S) - (t_S + c_S)S \\ = \;& (p - c - t_\beta \beta)D(p)Q(\beta, S) - (t_S + c_S)S. \end{aligned} \tag{8}$$

Note that n is the number of subscribers to the CP's service, c is the cost per subscriber to the CP for the service offering, t_β and t_S are unit charges imposed on the CP by the ISP for bandwidth and cache interconnection, respectively. Recall that $\beta n(p, \beta, S)$ is the total incremental bandwidth over the last mile dedicated to the CP's service.

Throughout the paper, with the exception of the special cases, we shall understand the feasibility region of the decision variables to be $\{p \geq 0, \beta \geq 0, S \geq 1\}$. But for the expression for the hit probability, h in (2), we would have simply constrained $S \geq 0$.

Let the solutions to the CP's problem be denoted by $p^*(t_\beta, t_S)$, $\beta^*(t_\beta, t_S)$, $S^*(t_\beta, t_S)$.

Now consider the ISP's problem as the leader in the Stackelberg model. The ISP's problem is to maximize its profit Π_{ISP} with respect to decision variables t_β, t_S, $(t_\beta \geq 0, t_S \geq 0)$, i.e.,

$$\max \Pi_{\text{ISP}}\{t_\beta, t_S, p^*(t_\beta, t_S), \beta^*(t_\beta, t_S), S^*(t_\beta, t_S)\}, \tag{9}$$

where

$$\Pi_{\text{ISP}} = (t_\beta - \eta_\beta)\beta^* D(p^*)Q(\beta^*, S^*) + (t_S - \eta_S)S^*. \tag{10}$$

Here η_β and η_S are unit costs for bandwidth and cache interconnection, respectively, that are incurred by the ISP.

Let the solution to the ISP's problem be denoted by t_β^*, t_S^*. These are the unit prices set by the ISP for the CP. The final operational values of the CP's decision variables are $p^*(t_\beta^*, t_S^*)$, $\beta^*(t_\beta^*, t_S^*)$, $S^*(t_\beta^*, t_S^*)$. This concludes the formulation of the Stackelberg model.

D. Alternative Models

The Stackelberg is the base model, and it corresponds to the case of Non-Net Neutrality (NNN) and Non-Integrated Operations (NIO). Alternative models follow as special cases. For instance, the Net Neutrality case is obtained by setting $\beta = 0$; also, in this case, $t_\beta = 0$.

In the IO model the CP and the ISP jointly make business decisions to maximize their combined profit; hence there are no transfer of funds from the CP to the ISP for payments for bandwidth and caching. Hence the Non-Net Neutrality-Integrated Operations (NNN-IO) model requires maximization of a single profit function given by the sum of the two profit functions in the NIO case.

This problem is equivalent to maximizing the CP's profit function in the Stackelberg model with the following substitutions: η_β for t_β, and η_S for t_S. That is, for NNN-IO, on defining $\Pi_{\text{IO}} = \Pi_{\text{CP}}$ as defined in (8), the equivalent problem is

$$\max\{\Pi_{\text{IO}}(p, \beta, S|\eta_\beta, \eta_S)\}, \tag{11}$$

Lastly, the case of NN-IO is a special case of NNN-IO with $\beta = 0$.

III. Analysis

A. Stackelberg Model

We begin by considering the CP's problem in (7) and (8). A solution to the problem in the interior of the feasible region must be a stationary point, which are points where the partial derivatives of the function $\Pi_{\text{CP}}(p, \beta, S)$ with respect to the decision variables are zero.

$$\partial\Pi_{\text{CP}}/\partial p = \partial\Pi_{\text{CP}}/\partial\beta = \partial\Pi_{\text{CP}}/\partial S = 0. \tag{12}$$

Following (4) and (5),

$$\partial\Pi_{\text{CP}}/\partial p = -n/p[\epsilon(p)(c + t_\beta\beta) - p(\epsilon(p) - 1)].$$

Therefore, from (12),

$$p(\epsilon(p) - 1)/\epsilon(p) = c + t_\beta\beta. \tag{13}$$

This equation has a unique solution p, since the left-hand-side is increasing in p under our assumption that $\epsilon(p)$ is non-decreasing in p. With the optimal price p obtained from (13) represented as a function of β, the CP's problem reduces to the determination of two variables, β and S.

For the special case where ϵ is a constant, the optimal price p, has the explicit expression

$$p = (c + t_\beta\beta)\epsilon/(\epsilon - 1). \tag{14}$$

Now, for the general case,

$$\begin{aligned} \partial\Pi_{\text{CP}}/\partial\beta &= -nt_\beta + (p - c - \beta t_\beta)(\partial n/\partial\beta) \qquad (15)\\ &= D(\beta)[-t_\beta Q + (c + t_\beta)\beta Q_\beta/(\epsilon(\beta) - 1)]. \end{aligned}$$

In the equation, we have written $D(\beta)$ and $\epsilon(\beta)$ for $D(p(\beta))$ and $\epsilon(p(\beta))$, respectively, where $p(\beta)$ is the optimal price determined by (13). The condition $\partial\Pi_{\text{CP}}/\partial\beta = 0$ is therefore equivalent to

$$F_1(t_\beta, t_S, \beta, S) = 0, \tag{16}$$

where

$$F_1(t_\beta, t_S, p, \beta, S) \triangleq (c + t_\beta\beta)Q_\beta/[(\epsilon(\beta) - 1)Q] - t_\beta. \tag{17}$$

Considering next $\partial\Pi_{\text{CP}}/\partial S$, we note that

$$\partial\Pi_{\text{CP}}/\partial S = (c + t_\beta\beta)DQ_S/(\epsilon(\beta) - 1) - (t_S + c_S).$$

Hence $\partial\Pi_{\text{CP}}/\partial S = 0$ is equivalent to

$$F_2(t_\beta, t_S, \beta, S) = 0, \tag{18}$$

where

$$F_2(t_\beta, t_S, \beta, S) \triangleq \frac{(c + t_\beta\beta)}{(\epsilon(\beta) - 1)}D(\beta)Q_S - (t_S + c_S). \tag{19}$$

For the Special Case defined in Section II-B, the two functions are specialized to

$$\begin{aligned} F_1(.) &= \frac{c + t_\beta\beta}{\epsilon - 1}\frac{1}{r_{\text{eff}}\log(r_{\text{eff}}/r_0)}\frac{\log S}{\log\Sigma} - t_\beta \\ F_2(.) &= \frac{A(\epsilon - 1)^{\epsilon-1}}{\epsilon^\epsilon(c + t_\beta\beta)^{\epsilon-1}}\frac{r_1 + \beta - r_2}{r_{\text{eff}}S\log\Sigma} - (t_S + c_S). \end{aligned}$$

Note that F_2 decreases as S increases while all other variables are held constant.

To summarize, any solution $\beta^*(t_\beta, t_S)$ and $S^*(t_\beta, t_S)$ to the CP's problem in the interior of the feasible region must be a stationary point, which must satisfy conditions (16) and (18), while $p^*(t_\beta, t_S)$ is obtained from $\beta^*(t_\beta, t_S)$ and (13).

It is useful to collect here for future reference expressions derived from (13) on the dependence of the solution p on β and t_β.

$$\begin{aligned} \partial p/\partial\beta &= \epsilon(p)t_\beta/[\epsilon(p) - 1 + E(p)], \qquad (20)\\ \partial p/\partial t_\beta &= \epsilon(p)\beta/[\epsilon(p) - 1 + E(p)], \\ \text{and}\quad \partial(p/\epsilon(p))/\partial t_\beta &= [1 - E(p)]\beta/[\epsilon(p) - 1 + E(p)]. \end{aligned}$$

We turn next to the ISP's problem, which is stated in (9) and (10). We transform the problem to the following constrained optimization problem with decision variables t_β and t_S:

$$\max\{\Pi_{\text{ISP}}(t_\beta, t_S, \beta, S)\} \tag{21}$$

$$\text{such that}\quad F_1(t_\beta, t_S, \beta, S) = 0, \tag{22}$$

$$\text{and}\quad F_2(t_\beta, t_S, \beta, S) = 0, \tag{23}$$

$$\text{where}\quad \Pi_{\text{ISP}}(t_\beta, t_S, \beta, S) = (t_\beta - \eta_\beta)\beta n + (t_S - \eta_S)S. \tag{24}$$

Equations (22) and (23) ensure that (β, S) are interior stationary points for the CP's problem.

We use the method of Lagrange multipliers to solve (21)-(23). Define the Lagrangian:

$$\begin{aligned} \mathcal{L} &= \Pi_{\text{ISP}}(t_\beta, t_S, \beta, S) \qquad (25)\\ &\quad + \lambda_1 F_1(t_\beta, t_S, \beta, S) + \lambda_2 F_2(t_\beta, t_S, \beta, S), \end{aligned}$$

where λ_1 and λ_2 are Lagrange multipliers corresponding to the constraints (22) and (23) to the maximization. Setting to zero the partial derivatives of the Lagrangian $\mathcal{L}$ with respect to t_β and t_S, we obtain the expressions for the Lagrange multipliers in terms of t_β, t_S, β, and S:

Proposition 1: *The Lagrange multipliers for the ISP's constrained optimization problem in (21)-(23) for the Stackelberg model are as follows.*

$$\lambda_1 = \beta D \frac{Q\left(\frac{(t_\beta-\eta_\beta)(\epsilon-1)\epsilon}{(c+t_\beta\beta)(\epsilon-1+E)}-1\right)+SQ_S}{\frac{(1-E)}{(\epsilon-1+E)}\frac{\beta Q_\beta}{Q}-1}, \quad (26)$$

$$\lambda_2 = S. \quad (27)$$

For the Special Case in Section II-B, the expression for λ_1 simplifies to:

$$\lambda_1 = \left[\frac{r_1+\beta-r_2}{r_{\text{eff}}\log\Sigma} - \log\left(\frac{r_{\text{eff}}}{r_0}\right)\left(1-\frac{\epsilon\beta(t_\beta-\eta_\beta)}{c+t_\beta\beta}\right)\right] \times \frac{\beta A}{p^\epsilon} \Big/ \left[\frac{\beta}{(\epsilon-1)}\frac{1}{r_{\text{eff}}\log(r_{\text{eff}}/r_0)}\frac{\log S}{\log\Sigma}-1\right]. (28)$$

Next, for the general case, we obtain the derivatives of $\mathcal{L}$ with respect to β and S and set each to zero. On substituting the expressions for the Lagrange multipliers in (26)-(27), this procedure yields two more equations in t_β,t_S, β, and S, which are obtained in the Appendix, (A-1) and (A-6).

Proposition 2: *The set of four equations comprised of (A-1) and (A-6), and the constraints in (22) and (23) completely characterize the stationary points t_β, t_S, β, S of the ISP's optimization problem formulated in (21)-(23), which identify interior solutions to the Stackelberg model if they exist.*

Section IV gives numerical results obtained by solving the set of four nonlinear equations using MATLAB. Also given in Section IV are validating solutions to the optimizations in the Stackelberg model obtained by exhaustive search.

As we shall see below while examining special cases of the Stackelberg model, the optima may importantly occur at the boundary of the feasible region and not in the interior.

B. Integrated Operations model

The IO model has been introduced in Section I-B and developed in Section II-D. The IO problem in (11) is the maximization of Π_{IO} with respect to the decision variables p, β,S, where

$$\Pi_{\text{IO}}(p,\beta,S) = (p-c-\eta_\beta\beta)n(p,\beta,S)-(c_S+\eta_S)S. \quad (29)$$

Now, the equation $\partial\Pi_{\text{IO}}/\partial p = 0$ yields the unique solution that satisfies an equation analogous to (13),

$$p(\epsilon(p)-1)/\epsilon(p) = c+\beta\eta_\beta. \quad (30)$$

The remaining analysis for the IO model follows from simplifications to (16) and (18). For purposes of this model, where η_β, η_S are constant parameters, we abbreviate the functions F_1 and F_2 in (16) and (18) thus,

$$F_i(\beta,S) = F_i(\eta_\beta,\eta_S,\beta,S) \; i=1,2. \quad (31)$$

The conditions for stationary points, $\partial\Pi_{\text{IO}}/\partial\beta = 0$ and $\partial\Pi_{\text{IO}}/\partial S = 0$, are equivalent to $F_1(\beta,S) = F_2(\beta,S) = 0$, which gives rise to the following result.

Proposition 3: *The solutions to the following two equations give β and S,*

$$(c+\eta_\beta\beta)Q_\beta - \eta_\beta(\epsilon(\beta)-1)Q = 0, \quad (32)$$

$$(c+\eta_\beta\beta)D(\beta)Q_S - (\epsilon(\beta)-1)(\eta_S+c_S) = 0, \quad (33)$$

and p is obtained from (30), which completes the characterization of the stationary points of $\Pi_{IO}(p,\beta,S)$ for the Integrated Operations model.

C. Net Neutrality Model

As discussed in Section II-D, the Net Neutrality model is obtained from the Stackelberg model by setting $\beta = 0$, to correspond to zero incremental allocation of bandwidth to subscribers of the CP's service over the access network. We analyze this model in two parts, first for Non-Integrated Operations followed by Integrated Operations.

1) Non-Integrated Operations: For the CP's problem, from the first-order condition (13), the optimum value of the subscriber's fee p reduces to the solution to the following single variable equation:

$$p(\epsilon(p)-1)/\epsilon(p) = c. \quad (34)$$

The other important implication for the CP's problem is, from (17) and (19),

$$F_1(t_S,S) \triangleq 0$$

$$\text{and } F_2(t_S,S) \triangleq cD(p)Q_S/[\epsilon(p)-1]-(t_S+c_S),$$

where we have introduced abbreviation $F_i(t_S,S)$ for $F_i(0,t_S,0,S)$ $(i=1,2)$. We thus have from (23),

$$cDQ_S/[\epsilon(p)-1]-(t_S+c_S) = 0. \quad (35)$$

We turn to the ISP's problem, for the Net Neutrality model, specializing (24),

$$\Pi_{\text{ISP}}(t_S,S) = (t_S-\eta_S)S. \quad (36)$$

We maximize Π_{ISP} with respect to t_S subject to the constraint (35). Now,

$$\partial\Pi_{\text{ISP}}/\partial t_S = S+(t_S-\eta_S)(\partial S/\partial t_S). \quad (37)$$

To obtain $\partial S/\partial t_S$, we take the derivative of (35) with respect to t_S,

$$\partial S/\partial t_S = (\epsilon(p)-1)/[cD(p)Q_{SS}(S)]. \quad (38)$$

On substituting in (37),

$$\partial\Pi_{\text{ISP}}/\partial t_S = S+[(t_S-\eta_S)(\epsilon(p)-1)]/[cD(p)Q_{SS}(S)].$$

Hence using (35),

$$\partial\Pi_{\text{ISP}}/\partial t_S = S+[(t_S-\eta_S)Q_S(S)]/[(t_S+c_S)Q_{SS}(S)]. \quad (39)$$

In summary we have

Proposition 4: *In the Net Neutrality model, the stationary point of the CP's optimization with respect to p and S is characterized by (34) and (35), and the behavior of the ISP's profit with respect to t_S is characterized by (39).*

The following proposition gives an important implication.

Proposition 5: *For all $S > 1$,*

$$\textit{if } SQ_{SS}+Q_S < 0, \quad \textit{then } \partial\Pi_{\text{ISP}}/\partial t_S > 0. \quad (40)$$

Proof: From (39),

$$\frac{\partial\Pi_{\text{ISP}}}{\partial t_S} = \frac{t_S(SQ_{SS}+Q_S)+c_SSQ_{SS}-\eta_SQ_S}{(t_S+c_S)Q_{SS}}.$$

Now, from (3), $Q_S > 0$ and $Q_{SS} < 0$; also, c_S and η_S are positive constants. Hence the result follows.

In the Special Case of II-B,

$$SQ_{SS}+Q_S = -(r_1-r_2)^2/[S(r_{\text{eff}} \log \Sigma)^2] < 0 \text{ for all } S > 1. \tag{41}$$

That is, the condition for the result in Proposition 5 to hold is indeed satisfied in the Special Case.

The importance of Proposition 5 is that it implies that in the Stackelberg model specialized to the NN-NIO case, the ISP's profit increases monotonically with increasing t_S, which is the unit price for the cache charged to the CP by the ISP. Since we also know that the cache size installed by the CP decreases as t_S is increased, the implication is that the ISP can maximize its profit by increasing the unit price of cache to such levels that it leads the CP to curb caching.

This result establishes that if Net Neutrality rules prohibit the use of enhanced bandwidth for the content service, i.e., $\beta = 0$, and if the condition (40) in Proposition 5 is satisfied, then caching will get curbed on account of the profit maximizing incentives of the ISP and the CP.

2) Integrated Operations: For the NN-IO case, from (11),

$$\frac{\partial \Pi_{\text{IO}}}{\partial S} = F_2(\eta_S, S) = \frac{c}{\epsilon(p)-1} D(p) Q_S - (c_S + \eta_S), \tag{42}$$

In the second equation, the first term on the right hand side is monotonically decreasing with increasing S, it is large for small S and approaches 0 as $S \to \infty$. Hence there exists a unique solution S^* such that for $S < S^*$ ($> S^*$) Π_{IO} is increasing (decreasing) with S, and achieves its maximum at S^* ($S^* > 1$).

This result highlights an important distinction within the Net Neutrality framework: with Integrated Operations, caching is compatible with joint profit maximization; in contrast, in the Stackelberg leader-follower model of Non-Integrated Operations, caching is curbed because the profit incentive of the ISP causes it to raise the unit price to the point where the CP finds it unprofitable to cache. The next section will illustrate this behavior.

IV. Numerical Investigations

We present results from our numerical investigations, and comment on important observed features. We have employed two different methods for obtaining numerical results. First, the results in Table I, which give the optimal solution to the Stackelberg game for the base NNN-NIO case, are obtained by using the MATLAB routine fsolve to solve the four nonlinear equations that are specified in Proposition 2. The remaining numerical results are obtained by exhaustive search using the MATLAB routine fminsearch. The first method, where applicable, is more time-efficient by orders of magnitude; the latter offers validation and is less restrictive. The results presented here are based on the user model in the Special Case, Section II-B. The parameters are as follows:

$$A = 200,000;\ r_1 = 2.0;\ r_2 = 1.0;\ r_0 = 1.0; \Sigma = 100,000.$$

The cost parameters: $c = 1$; $c_S = 0.1$; $\eta_\beta = 0.1$, $\eta_S = 0.1$. These parameters are used throughout.

Table I gives the analytical solutions to the base NNN-NIO model for various values of the price elasticity of demand, ϵ, in the range $1.15 - 1.35$.

$\boldsymbol{\epsilon}$	**1.15**	**1.2**	**1.25**	**1.3**	**1.35**
$\boldsymbol{p}$	5401.5	639.1	173.6	72.7	39.0
$\boldsymbol{\beta}$	1011.0	186.5	66.8	33.2	19.9
$\boldsymbol{h}$	0.77	0.78	0.78	0.78	0.77
QoE	6.66	4.99	3.98	3.32	2.84
$\boldsymbol{n}$	68	429	1,264	2,520	4,040
$\boldsymbol{t_\beta}$	0.70	0.57	0.51	0.48	0.46
$\boldsymbol{t_S}$	0.65	0.58	0.53	0.51	0.50
$\boldsymbol{\Pi_{\text{CP}}}$	313,759	223,208	170,708	136,398	112,462
$\boldsymbol{\Pi_{\text{ISP}}}$	44,894	40,823	37,485	34,479	31,681

TABLE I
Optimal solutions for NNN-NIO case for various price elasticity of demand, ϵ

The table shows strikingly different optimal values of the incremental bandwidth (β) for different price elasticities (ϵ). Intuitively, with more deployment of the incremental bandwidth, β, the CP's marginal cost per customer, $c + t_\beta \beta$, rises and so does its optimal subscription fee, p. With a lower price elasticity, the same percentage increase of the subscription fee causes a smaller percentage loss of subscribers, so it is more profitable for the CP to pursue this high-bandwidth (and hence high QoE) and high price strategy. On the other hand, high price elasticity induces exactly the opposite, which explains large variations of p, β, QoE, and n in the table.

In contrast, the other factor that determines QoE, the hit probability, h, barely changes with the price elasticity. Observe from (2) that changing the hit probability requires an exponential amount of change of the cache size S. Therefore for hit probabilities to vary noticeably with price elasticity, there needs to be a significant change of the marginal revenue with respect to h to justify the associated exponential change of CP's cost, $(t_S + c_S)S$. Such a revenue impact is absent here because QoE is logarithmic in the hit probability. Also, unlike the incremental bandwidth β, cache is a common resource, i.e., its deployment cost is not directly attributable to individual users. So changing the size of the cache and the hit probability has little bearing on the optimal subscription fee p, which critically depends on the price elasticity and determines the marginal revenue.

It is also important to observe that the stable value of the hit probability is relatively high, which can be explained by the ISP's desire to sell more incremental bandwidth. With $r_1 > r_2$, the incremental bandwidth has little value in cases where the hit probability is small, and the QoE is largely constrained by the capacity in the core. To make it worthwhile to invest in additional bandwidth in the "last mile", the ISP has to set sufficiently low price to induce the CP to deploy a substantial amount of cache to allow users to avoid the bottleneck beyond the access network.

Figures 2 and 3 further demonstrate some trends that give rise to results in Table I. Both correspond to the case with $\epsilon = 1.25$. All decision variables except those demonstrated in the figures are held at their optimal values given in Table I.

Figure 2 illustrates CP's profit as a function of the two decisions that determine the QoE, β and h. As the figure shows, the profit change with respect to β is somewhat symmetric around the maxima. The change with respect to h features a slow growth before reaching the maxima and a steep decline after the peak. The latter pattern is consistent with the fact that CP's cost on caching is exponential in the hit rate, the rise of which accelerates as the probability becomes increasingly higher.

Figure 3-a shows the optimal bandwidth per user, β, as a function of ISP's bandwidth price t_β, and Figure 3-b shows the change of ISP's profit with the change of that price. As expected, a higher price by the ISP leads to lower bandwidth demand from the CP, and the ISP's profit declines once the price passes the optimal point. However, both the demand loss and profit decline happen at an increasingly slower rate as the price keeps increasing. This is because QoE is logarithmic in the effective bandwidth, and hence drops rapidly with respect to β when the latter value is small. Therefore, even though the CP needs to reduce β as the ISP raises t_β, in the range of high price (and low β), the reduction needs to be slow-paced to avoid a devastating drop of QoE.

Tables II and III give results for Integrated Operations, Table II for the NNN case (optimized β) and Table III for the NN case ($\beta = 0$). The gains due to Integrated Operations are shown in Table II to vary in the $20 - 30\%$ range depending on price elasticity. The loss in joint profit when subject to NN rules is substantial, in the $70-80\%$ range. The data also shows striking divergences in price (p) and number of subscribers (n). Comparing results from Tables I and II, Integrated Operations give lower prices and substantially increased subscribers, which is exactly as expected, given that the pricing decision under integrated operations is based on the actual bandwidth and cache costs instead of the inflated ones (i.e., $\eta_\beta \leq t_\beta$ and $\eta_S \leq t_S$). Also, within integrated operations, NN continues this trend, giving lower prices (almost an order of magnitude at low price elasticity) and increased subscribers. This is because by setting $\beta = 0$, the marginal cost is kept at the minimum (i.e., $c \leq c + t_\beta \beta$), leading to a business regime with low cost, low QoE, and low price.

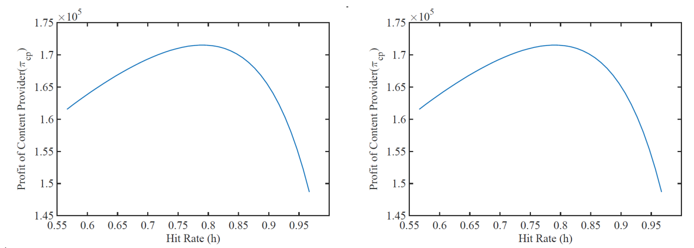
ϵ	**1.15**	**1.2**	**1.25**	**1.3**	**1.35**
$\boldsymbol{p}$	715.3	125.5	47.7	26.0	17.1
$\boldsymbol{\beta}$	922.9	199.2	85.5	50.1	34.4
$\boldsymbol{h}$	0.90	0.90	0.89	0.89	0.88
QoE	6.72	5.20	4.36	3.84	3.47
$\boldsymbol{n}$	701	3,150	6,948	11,091	15,026
$\boldsymbol{\Pi_{IO}}$	430,010	323,340	259,530	216,490	185,220
gain	19.9%	22.5%	24.1%	26.7%	28.5%

TABLE II
NNN-IO CASE, Π_{IO} IS PROFIT FROM INTEGRATED OPERATIONS, AND THE INTEGRATION GAIN COMPARES Π_{IO} TO $(\Pi_{CP} + \Pi_{ISP})$ FROM TABLE I

The final Table IV gives data for the NN-NIO case, i.e., $\beta = t_\beta = 0$, for price elasticity $\epsilon = 1.25$. The data gives optimum values of the CP's profit maximization for various values of t_S, the unit price for caches, i.e., the displayed ISP's profit is for this parameter value, and not from optimization. As the theory predicts, see Section III-C1, the ISP finds it in its interest to continually increase t_S to derive the maximum profit from the minimum amount of cache deployment by the CP. Thus in the NN-NIO case, extra bandwidth, which is withheld from users due to regulation, also has the unintended effect of curbing caching! Note that this effect is rooted in profit incentives and allocations as the following data point will illustrate. Table IV shows that for $t_S = 0.1$, $\Pi_{CP} = 62,521$ and $\Pi_{ISP} = 0$, so the sum is almost identical to the corresponding optimal profit $(62,522)$ for the NN-IO case shown in Table III! This example also highlights the importance of stable and fair allocations of profit in service offerings in which all parties, including users, benefit from cooperation. This dimension is not reflected in the Stackelberg game.

ϵ	**1.15**	**1.2**	**1.25**	**1.3**	**1.35**
$\boldsymbol{p}$	7.7	6.0	5.0	4.3	3.9
$\boldsymbol{n}$	12,271	14,776	16,868	18,644	20,173
$\boldsymbol{h}$	0.84	0.84	0.83	0.82	0.82
$\boldsymbol{\Pi_{IO}}$	75,931	68,513	62,522	57,550	53,341
loss	82.3%	78.8%	75.9%	73.4%	71.2%

TABLE III
NN-IO CASE, HENCE $\beta = 0$, Π_{IO} IS PROFIT FROM INTEGRATED OPERATIONS, AND LOSS DUE TO NET NEUTRALITY COMPARES Π_{IO} TO Π_{IO} IN NNN-IO CASE, GIVEN IN TABLE II



(a) Change of Π_{CP} w.r.t. β (b) Change of Π_{CP} w.r.t. h

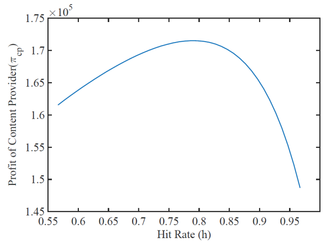
Fig. 2. CP's profit, Π_{CP}, for various β, bandwidth per user, and h, cache hit probability

t_s	h	n	p	Π_{CP}	Π_{ISP}
0.1	0.83	16,867	5	62,521	0
100	0.35	8,351	5	26,603	6,787
500	0.22	5,655	5	15,101	7,518
1000	0.17	4,411	5	9,765	7,989
2000	0.12	3,110	5	4,169	8,271
3250	0.08	2,163	5	2,163	8,570
3260	0.08	2,157	5	55.3	8,572
3270	0.08	2,152	5	29.1	8,579
3280	0.08	2,145	5	2.9	8,577

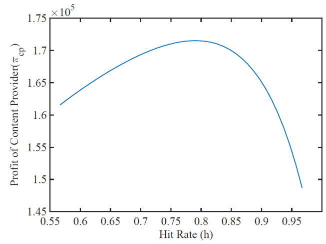
TABLE IV
NN-NIO CASE, HENCE $\beta = 0$, PRICE ELASTICITY $\epsilon = 1.25$, VARIOUS UNIT CACHE PRICE t_S.

V. CONCLUSIONS

Our study spans multiple dimensions: engineering rules that relate allocation of resources (bandwidth, cache) to the user-perceived effective bandwidth; end user models that connect willingness to pay with QoE and quantify demand response to price; industry structure that governs profit-maximizing behavior of the CP and the ISP; and policy environment that imposes boundaries on companies' decision making. We show that changes of parameter values and policy regimes can lead to substantially different outcomes.



(a) Change of β w.r.t. t_B



(b) Change of Π_{ISP} w.r.t. t_B

Fig. 3. Bandwidth per user, β, and ISP's profit,Π_{ISP}, for various unit price for bandwidth, t_β

Our work can be expanded along different directions. For instance, some CPs have become dominant market players, which enable them to gain much more bargaining power with the ISP than a price-taking "follower". On the other hand, these CPs are under intense scrutiny by regulators and the public, making it hard for them to cooperate with the ISP to maximize their joint profits. To address such cases, we need to go beyond both Integrated Operations and the Stackelberg Game, and develop new models of interplay between the CP and ISP. Moreover, while it is important to understand profit implications of NN regulation, the purpose of the regulation is for social benefits. Thus it is fitting that recently we have extended our work [21] to include social welfare.

Acknowledgement: The work reported in the paper is supported by the National Science Foundation under Grant No. CMMI-1435378 (Debasis Mitra and Ao Hong) and CMMI-1435371 (Qiong Wang).

VI. Appendix

The analysis of the Stackelberg model, see Section III-A, is completed here with the characterization of the stationary values of t_β, t_S, β, and S in the ISP's constrained optimization problem in (21)-(24). Recall that the Lagrangian for the problem, $\mathcal{L}$, is defined in (25), and that $\partial\mathcal{L}/\partial t_\beta = \partial\mathcal{L}/\partial t_S = 0$ yield solutions for the Lagrange multipliers, λ_1 and λ_2, which are given in Proposition 1. Here we derive $\partial\mathcal{L}/\partial\beta = \partial\mathcal{L}/\partial S = 0$ in terms of t_β, t_S, β, and S. These two equations, together with the constraints in (22) and (23), constitute the complete characterization.

From (25) and $\partial\mathcal{L}/\partial\beta = 0$,

$$\partial\Pi_{\text{ISP}}/\partial\beta + \lambda_1(\partial F_1/\partial\beta) + \lambda_2(\partial F_2/\partial\beta) = 0. \tag{A-1}$$

Now

$$\frac{\partial\Pi_{\text{ISP}}}{\partial\beta} = (t_\beta - \eta_\beta)\left[1 - \frac{\epsilon(\epsilon-1)t_\beta\beta}{(\epsilon-1+E)(c+t_\beta\beta)} + \beta\frac{Q_\beta}{Q}\right] DQ. \tag{A-2}$$

In deriving (A-2), we used the expression for $\partial p/\partial\beta$ in (20) to show that

$$\frac{\partial D}{\partial\beta} = \frac{\partial D}{\partial p}\frac{\partial p}{\partial\beta} = -\frac{\epsilon(\epsilon-1)t_\beta D}{(\epsilon-1+E)(c+t_\beta\beta)}. \tag{A-3}$$

Next

$$\frac{\partial F_1}{\partial\beta} = \frac{(1-E)}{(\epsilon-1+E)}\frac{t_\beta Q_\beta}{Q} + \frac{(c+t_\beta\beta)}{\epsilon-1}\left\{\frac{Q_{\beta\beta}}{Q} - \left(\frac{Q_\beta}{Q}\right)^2\right\}. \tag{A-4}$$

Finally,

$$\partial F_2/\partial\beta = D[(c+t_\beta\beta)Q_{S\beta}/(\epsilon-1) - t_\beta Q_S] \tag{A-5}$$

With λ_1 and λ_2 in (26) and (27), the left hand side of (A-1) is now obtained. Note that the demand function D is a common factor in $\partial\Pi_{\text{ISP}}/\partial\beta$, λ_1 and $\partial F_2/\partial\beta$, and hence may be eliminated.

Next consider $\partial\mathcal{L}/\partial S = 0$. From (25),

$$\partial\Pi_{\text{ISP}}/\partial S + \lambda_1(\partial F_1/\partial S) + \lambda_2(\partial F_2/\partial S) = 0. \tag{A-6}$$

Now $$\partial\Pi_{\text{ISP}}/\partial S = (t_\beta - \eta_\beta)\beta DQ_S + (t_S - \eta_S). \tag{A-7}$$

$$\partial F_1/\partial S = (c+t_\beta\beta)[Q_{S\beta}Q - Q_\beta Q_S]/((\epsilon-1)Q^2). \tag{A-8}$$

and $$\partial F_2/\partial S = (c+t_\beta\beta)DQ_{SS}/(\epsilon-1). \tag{A-9}$$

With λ_1 and λ_2 in Prop. 1, the left hand side of (A-6) is now proved.

References

[1] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, Z.-L. Zhang, M. Varvello, and M. Steiner, "Measurement study of Netflix, Hulu, and a tale of three CDNs," *IEEE/ACM Trans. Networking*, vol. 23, no. 6, pp. 1984–1997, 2014.

[2] Netflix, "Netflix OpenConnect content delivery for ISPs." [Online]. Available: https://openconnect.netflix.com.

[3] J. Brodkin, "Netflix's many-pronged plan to eliminate video playback problems," *arstechnica*, May 13, 2014. [Online]. Available: http://arstechnica.com/information-technology/2014/05/netflixs-many-pronged-plan-to-eliminate-video-playback-problems/

[4] J. Roettgers, "How Netflix built its OpenConnect cache to speed up your video streams," *gigaom*, June 20, 2013. [Online]. Available: https://gigaom.com/2013/06/20/how-netflix-built-its-openconnect-cache-to-speed-up-your-video-streams/

[5] K. Russell, "What the Netflix-comcast deal really means in plain english," *Business Insider*, Feb 23, 2014. [Online]. Available: http://www.businessinsider.com/netflix-comcast-deal-explained-2014-2.

[6] E. Steel, "Netflix, Amazon and Hulu no longer find themselves upstarts in online streaming," *The New York Times*, March 24, 2015. [Online]. Available: http://nyti.ms/1HyRtYN

[7] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang, "Understanding the impact of video quality on user engagement," in *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 4. ACM, 2011, pp. 362–373.

[8] H. R. Varian, *Intermediate microeconomics with Calculus: A Modern Approach*. WW Norton & Company New York, NY, 2014.

[9] FCC, "FCC adopts strong, sustainable rules to protect the open internet," February 26, 2015. [Online]. Available: www.fcc.gov/guides/open-internet

[10] FCC Open Internet Advisory Committee: Specialized Service Working Group, "Specialized services: Summary of findings and conclusions," August 20, 2013.

[11] M. O. Jackson, *Social and economic networks*. Princeton university Press, Princeton, 2008.

[12] D. N. Serpanos, G. Karakostas, and W. H. Wolf, "Effective caching of web objects using Zipf's law," in *Proceeding of 2000 IEEE International Conference on Multimedia and Expo*, vol. 2. IEEE, 2000, pp. 727–730.

[13] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," *Proc. IEEE INFOCOM 2010*, pp. 1478–1486.

[14] S. Hasan, G. S, C. Dovrolis, and R. K. Sitaraman, "Trade-offs in optimizing cache deployments in cdns," *Proc. IEEE INFOCOM 2014*, pp. 460–468.

[15] K. Hosanagar, J. Chuang, R. Krishnan, and M. Smith, "Service adoption and pricing of content delivery network (CDN) services," *Management Science*, no. 9, pp. 1579–1593, 2008.

[16] R. Edell and P. Varaiya, "Providing internet access: What we learn from index," *IEEE Network*, no. 5, pp. 18–25, 1999.

[17] S. Lanning, D. Mitra, Q. Wang, and M. Wright, "Optimal planning for optical transport networks," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 358, no. 1773, pp. 2183–2196, 2000.

[18] P. Reichl, B. Tuffin, and R. Schutz, "Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience," *Telecom Systems*, vol. 52, pp. 587–600, 2013.

[19] ITU-T(Rec.G.100/P.10), "Vocabulary for performance and quality of service, amendment 2: new definition for inclusion in Recommendation P.10/G.100," 2008.

[20] P. Maillé and B. Tuffin, "Impact of content delivery networks on service and content innovation," *ACM SIGMETRICS Performance Evaluation Review*, vol. 43, no. 3, pp. 49–52, 2015.

[21] D. Mitra, Q. Wang, and A. Hong, "Social welfare and policy impact in emerging isp-content provider relationships," in *Proc. TPRC44, The 44th Research Conference on Communications, Information and Internet Policy*, Arlington,VA, September 2016.