Evolution of the Internet Economic Ecosystem

Richard T. B. Ma† John C. S. Lui‡ Vishal Misra⋆

***Abstract*—The evolution of the Internet has manifested itself in many ways: the traffic characteristics, the interconnection topologies and the business relationships among the autonomous components. It is important to understand why (and how) this evolution came about, and how the interplay of these dynamics may affect future evolution and services. We propose a network aware, macroscopic model that captures the characteristics and interactions of the application and network providers, and show how it leads to a market equilibrium of the ecosystem. By analyzing the driving forces and the dynamics of the market equilibrium, we obtain some fundamental understandings of the cause and effect of the Internet evolution, which explain why some historical and recent evolutions have happened. Furthermore, by projecting the likely future evolutions, our model can help application and network providers to make informed business decisions so as to succeed in this competitive ecosystem.**

***Index Terms*—Internet Economics, Macroscopic Evolution**

I. Introduction

The Internet has been and is still changing unexpectedly in many aspects. Started with elastic traffic and applications, e.g., emails and webpage downloading, we have seen significant rise in inelastic traffic, e.g., video and interactive web traffic, across the Internet. According to [21], from 2007 to 2009, web content traffic had increased from 41.68% to 52%, reaching more than half of the total Internet traffic. From a network perspective, the Internet originated from government-owned backbone networks, i.e., the ARPANET, and then evolved to a network of commercial Autonomous Systems (ASes) and Internet Service Providers (ISPs). Meanwhile, ISPs formed a hierarchical structure and were classified by tiers, with higher tier ISPs cover larger geographic regions and provide transit service for smaller/lower tier ISPs. However, recent study [19] has reported that large content providers, e.g., Google and Microsoft, are deploying their own wide-area networks so as to bring content closer to users and bypassing Tier-1 ISPs on many paths. This is known as the *flattening phenomenon* of the Internet topology.

Changes in the content or network topology do not happen independently. Rather, they are driven by the changes in the business relationships among the players in the Internet ecosystem. Not surprisingly, we have observed dramatic changes in the business relationships between the content providers and the ISPs and among the ISPs themselves. Traditionally, ISP settlements were often done bilaterally under either a (zero-dollar) peering or in the form of a customer-provider relationship. Tier-1 ISPs, e.g., Level 3 [5], often charge lower tier ISPs for transit services and connect with each other under settlement-free peering. However, the Tier-1 ISPs do not have any guarantee in their profitability as the Internet evolves. For instance, we have seen exponential decrease (around 20% a year) in IP transit prices [26]. Also, peering disputes happened, e.g., the de-peering between Cogent [3] and Level 3 in 2005, where the lower tier ISPs that are closer to content or users refused to pay for the transit charge. This leads to the recent debate of network neutrality [29], which reflects the ISPs' willingness to provide value-added and differentiated services and potentially charge content providers based on different levels of service quality.

The situation is further complicated by the emergence of new players in the ecosystem: Content Delivery Networks (CDNs), e.g., Akamai [1] and Limelight [6], and high-quality video streaming providers, e.g., Netflix [7]. From content providers' perspective, CDNs can deliver their content faster and more efficiently; from local ISPs' perspective, CDNs can reduce the traffic volume from upstream, saving transit costs from their providers. Very often, ISPs do not charge the CDNs for putting servers in their networks. When the video streaming giant Netflix moved online a few years ago, its traffic surged immediately. Now it accounts for up to 32.7% of peak U.S. downstream traffic [8] and its traffic volume is higher than that of BitTorrent [9] applications. Netflix used Limelight, one of the biggest CDNs, for content delivery, and later, the Tier-1 Level 3 also obtained a contract to deliver Netflix's traffic. Since most of the Netflix customers are based in the U.S., they often use Comcast, the biggest access ISP, as the last-mile access provider. Interestingly, Comcast managed to enter a so-called paid-peering relationship [17] with Level 3 and Limelight, under which the Tier-1 ISP and the CDN have to pay the access ISP for higher bandwidth on the last mile connection. This has totally *reversed* the nominal customer-provider relationship where the Tier-1 ISP was the service provider and should have received payment for connectivity.

It is important to understand how these changes come about, and what the driving factors are behind these changes. In this work, we model the Internet evolution from a macroscopic view that captures the cause and effect of the evolution of the individual players in the ecosystem. Our model expends the traditional view of a single best-effort service model to capture multiple value-added services in the Internet. The main approaches and contributions are as follows.

- We model the preferences and business decisions of the

This study is supported in part by the research grant for the Human Sixth Sense Programme at ADSC from Singapore's Agency for Science, Technology and Research (A*STAR), Ministry of Education of Singapore AcRF grant T1 251RES1306 and R-252-000-448-133, and Hong Kong GRF funding 2150728 (Ref number: 415112).

†R. Ma is with Advanced Digital Sciences Center, Illinois at Singapore and National University of Singapore; tbma@comp.nus.edu.sg.

‡J. C. S. Lui is with Department of Computer Science and Engineering, The Chinese University of Hong Kong; cslui@cse.cuhk.edu.hk.

⋆V. Misra is with Department of Computer Science, Columbia University; misra@cs.columbia.edu.

application providers for purchasing Internet services, based on the application characteristics and the price and quality of the transport services (Section II).
- We characterize the market price and the market share of the Internet transport services by using general equilibrium theory in economics (Section III).
- We analyze the driving forces of the evolution of the Internet economic ecosystem (Section IV), which provide qualitative answers (Section IV-F) to questions like: *1) Why have the IP transit prices been dropping? 2) Why have the CDNs emerged in the ecosystem? 3) Why has the pricing power shifted to the access ISPs? 4) Why are the large content providers building their own wide-area networks toward users?*
- We incorporate Internet price and capacity data into our model, and quantitatively fit historical prices and project the future evolution of the ecosystem and its price trends (Section V).
- We demonstrate how our model can help the network providers to make business decisions, e.g., capacity expansion and peering decisions, based on the future price projections under various scenarios (Section V-C).

Our paper sheds new light on the macroscopic evolution of the Internet economic ecosystem and concretely identifies the driving factors of such an evolution. In particular, our model provides a tool to analyze and project the evolutionary trends of the ecosystem. The fundamental understanding of the preferences of application providers and the market equilibrium of the Internet services will also help the business decisions of the application and network transport providers to succeed in this competitive ecosystem.

II. The Macroscopic AP-TP Model

We start with a macroscopic model of the Internet ecosystem that consists a set of Application Providers (APs) and Transport Providers (TPs). The TPs differ by their service qualities and the prices they charge. We model and analyze the APs' choice of TP based on their own characteristics: how profitable the AP is and how sensitive the AP traffic is to the obtained level of service quality. In essence, this macroscopic model can help us to understand the decision process of these players in the Internet ecosystem and how these decisions may influence their business relationships.

A. The Application and Transport Providers

We consider an Internet service market of a geographic region and denote $(\mathcal{M}, \mathcal{N})$ as a macroscopic model of the ecosystem, consisting of a set $\mathcal{M}$ of TPs and a set $\mathcal{N}$ of APs. The APs provide the content/service for the Internet end-users; the TPs provide the network infrastructure for delivering the APs' data to their end-users.

Our notion of an AP broadly includes content providers, e.g., Netflix, online services, e-commerce, and even cloud services, e.g., Amazon EC2 [2]. Our notion of a TP is based on the APs' point of view. In other words, the transport services provided by the TPs are for the APs to reach their customers/users. The scope of a TP is *broader* than an ISP, and it includes CDNs. ISPs, depending on different taxonomies [17], [16], [24], include 1) eyeball/access ISPs that serve the last-mile for end-users, 2) backbone/Tier 1 ISPs that provide transit services for lower tier ISPs, and 3) content ISPs that serve APs and host content servers. A TP can be an ISP of any type. Although access and transit ISPs traditionally do not have business relationships with APs explicitly, with the emergence of video streaming APs, e.g., Netflix, we have seen more and more APs' direct or indirect contracts with the access and transit ISPs. For example, Level 3 contracted with Netflix for content delivery and Comcast managed to charge Level 3 and Limelight via paid-peering contracts (for delivering Netflix's traffic to Comcast's customer base faster) [26]. Although "whether ISPs should be allowed to differentiate services/charges for APs" is hotly debated under the network neutrality [29] argument, legitimate service differentiations will also induce more extensive business relationships among the APs and ISPs. In general, a TP can be any facilitator that delivers content to end-users. An important example of a TP that does not even own network infrastructures in the current Internet ecosystem is Akamai [1], which represents the CDNs.

We characterize each TP $I \in \mathcal{M}$ by its type, denoted as a triple (p_I, q_I, ν_I). p_I denotes the per unit traffic charge for the APs to use TP I. q_I denotes the service quality of TP I, e.g., queueing delay or packet loss probability. Without loss of generality, we assume that $q_I \geq 0$ and smaller values of q_I indicate better quality of services. ν_I denotes the bandwidth capacity of TP I. We characterize each AP $i \in \mathcal{N}$ by its utility function $u_i(\cdot)$. In particular, we define $u_i(p_I, q_I)$ as AP i's utility when it uses TP I, which depends on the service quality q_I and the per unit traffic charge p_I.

Assumption *1:* $u_i(\cdot, \cdot)$ is non-increasing in both arguments.

Assumption *2:* For any set $\mathcal{M}$ of TPs, each AP $i \in \mathcal{N}$ chooses to use a TP, denoted as $I_i \in \mathcal{M}$, that satisfies

$$u_i(p_{I_i}, q_{I_i}) \geq u_i(p_I, q_I), \quad \forall\ I \in \mathcal{M}.$$

The above assumes that each AP is rational and chooses a TP that provides the highest utility. Technically, there might exists multiple TPs that provide the same amount of utility for the AP. We assume that every AP has certain preference to break the tie and choose one of the TPs. We further denote $\mathcal{N}_I \subseteq \mathcal{N}$ as the set of APs that choose to use TP I, or the market share of TP I, defined as $\mathcal{N}_I = \{i \in \mathcal{N} : I_i = I\}$. Based on Assumption 1 and 2, if two TPs I, J have the same quality, i.e., $q_I = q_J$, then they have to price equally, i.e., $p_I = p_J$; otherwise, the one with higher price will not obtain any market share. As TPs differ only by price p_I and quality q_I from the APs' perspective, we aggregate the TPs that have the same value pair (p_I, q_I) into a single TP with a capacity that equals the summation of individual TPs' capacity. Similarly, if a TP performs service differentiations, we conceptually treat it as multiple TPs, each with a service class (p_I, q_I) and the corresponding capacity ν_I. More precisely, our abstraction of a TP I models a competitive market segment that provides a quality level q_I and has a total capacity ν_I.

B. Throughput and Types of the APs

Although the utility function u_i can be used to model all the characteristics of AP i, the setting does not yet capture the traffic dynamics and the profitability of the APs. We model AP i's profitability by denoting v_i as its per unit traffic revenue. This revenue is related to the AP's core business, e.g., online adverting or e-commerce, and we do not assume how it is generated. We denote $\lambda_i(\cdot)$ as AP i's throughput function, where $\lambda_i(q_I)$ defines the aggregate throughput of AP i toward its consumers under a quality level q_I. Thus, we model any AP i's utility as its total profit (profit margin multiplied by the total throughput rate), defined by $u_i(p_I, q_I) = (v_i - p_I)\lambda_i(q_I)$.

Assumption *3:* For any AP $i \in \mathcal{N}$, $\lambda_i(\cdot)$ is a non-increasing function with $\alpha_i = \lim_{q_i \to 0} \lambda_i(q_i)$ and $\lim_{q_i \to \infty} \lambda_i(q_i) = 0$.

Assumption 3 says that the throughput will not decrease if an AP uses a better service. λ_i reaches a maximum value of α_i when it receives the best quality $q_i = 0$ and decreases to zero if the quality deteriorates infinitely, i.e., q_i tends to $+\infty$. In particular, we consider the following canonical form of the throughput function $\lambda_i(q_I) = \alpha_i e^{-\beta_i q_I}$, where AP i's throughput is characterized by a parameter β_i that captures its sensitivity to the received quality q_I. Notice that the throughput also depends on the APs' charge to their customers, which affects their customer demand. Because this work focuses on the interaction between the APs and the TPs, we capture the user demand in the parameter α_i. To understand the sensitivity parameter β_i, we calculate the *throughput elasticity of quality* ϵ_i, defined by $\epsilon_i = \dfrac{d\lambda_i(q_I)}{dq_I}\dfrac{q_I}{\lambda_i(q_I)} = -\beta_i q_I$. The throughput elasticity captures the ratio of the percentage change in throughput caused by the percentage change in the quality. The above shows that any AP's throughput elasticity of quality is proportional to its sensitivity β_i, and therefore, our model uses a single sensitivity parameter to capture APs with different sensitivities to the quality. Figure 1 illustrates the

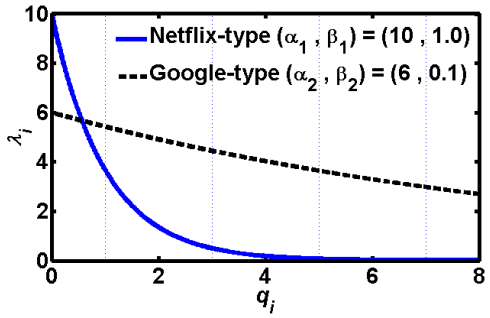


Fig. 1. Throughput of different type of APs.

throughput of two APs with parameters $(\alpha_1, \beta_1) = (10, 1.0)$ and $(\alpha_2, \beta_2) = (6, 0.1)$ under varying service qualities, interpreted as network delays in this case, along the x-axis. For an illustrative purpose, AP 1 models a Netflix-type of application that is more sensitive to delay and has a high maximum rate $\alpha_1 = 10$ Mbps; however, AP 2 models a Google-type of query application that is less sensitive to delay. We observe that when delay increases, the throughput of delay-sensitive application decreases sharply, while the delay-insensitive application decreases only mildly.

Because α_i is just a linear scaling factor of the throughput, it does not affect the AP's preference over different TPs. Consequently, APs with the same (β_i, v_i) value pairs will choose the same TP; and therefore, we can conceptually aggregate them as a single AP. Similar to a TP I representing a market segment, each AP i can be interpreted as a group of APs with the same characteristics and α_i represents the aggregate maximum traffic intensity, which depends on the number of APs in the group and the individual traffic intensities. Although α_i does not play a role in the AP's decision of choosing TPs, we will see later that α_i reflects the demand of the APs and affects the market prices of the TPs. In summary, based on our throughput model, we define

$$u_i(p_I, q_I) = (v_i - p_I)\lambda_i(q_I) = \alpha_i(v_i - p_I)e^{-\beta_i q_I}. \quad (1)$$

Similar to each TP I's type (p_I, q_I, ν_I), we can characterize any AP i's type as another triple (α_i, β_i, v_i).

C. APs' Choice of Transport Providers

When facing a set $\mathcal{M}$ of TPs, each AP i's best choice I_i depends on the price-quality pairs $\{(p_I, q_I) : I \in \mathcal{M}\}$ and its own characteristics (β_i, v_i). The APs' choices satisfy the following results.

Theorem *1 (Monotonicity of AP Choices):* For a fixed set $\mathcal{M}$ and any two APs i and j with $\beta_j \geq \beta_i$, $v_j \geq v_i$ and $(\beta_j, v_j) \neq (\beta_i, v_i)$, their chosen service qualities satisfy $q_{I_i} \geq q_{I_j}$.

Theorem 1 says that if an AP j is more profitable and more sensitive to service quality than another AP i, then the chosen quality of AP j will be at least as good as that of AP i. This property holds regardless how the services are priced.

Theorem *2:* For any $\kappa_1, \kappa_2, \kappa_3 > 0$, and system $(\mathcal{M}, \mathcal{N})$, we define a scaled system $(\mathcal{M}', \mathcal{N}')$ as $\mathcal{M}' = \{(\kappa_1 p_I + \kappa_2, q_I/\kappa_3, \nu_I) : I \in \mathcal{M}\}$ and $\mathcal{N}' = \{(\alpha_i, \kappa_3\beta_i, \kappa_1 v_i + \kappa_2) : i \in \mathcal{N}\}$, then system $(\mathcal{M}', \mathcal{N}')$ satisfies $\mathcal{N}_I(\mathcal{M}', \mathcal{N}') = \mathcal{N}_I(\mathcal{M}, \mathcal{N})$ for all $I \in \mathcal{M}$.

Theorem 2 says that if 1) the AP profitability v_i and the TP price p_I are linearly scaled in the same way, and/or 2) the quality q_I of the TPs and the sensitivity β_i of the APs scale inversely at the same rate, then the APs' choices of TP will not change. The intuition is that the abovementioned scaling in v_i and p_I does not change the APs' optimal choices of TPs, and the scaling in β_i and q_I does not change the throughput. This result will help us normalize different systems and make a fair comparison of various solutions.

Theorem *3:* For any $\kappa > 0$ and a fixed set $\mathcal{N}$ of APs, let $\mathcal{M}' = \{(p_I, \kappa q_I, \nu_I) : I \in \mathcal{M}\}$, then for all $i \in \mathcal{N}$, 1) $q_{I_i'} \leq \kappa q_{I_i}$ if $\kappa > 1$ and 2) $q_{I_i'} \geq \kappa q_{I_i}$ if $\kappa < 1$.

Theorem 3 says that if all the qualities in the market deteriorate ($\kappa > 1$) linearly at the same rate, APs will not use worse quality TPs than before. The opposite is also true: when qualities improve linearly, APs will not use better quality TPs than before. Intuitively, this result captures the fact that quality becomes a more (less) important concern when all the TPs provide worse (better) of it.

With this framework, we can understand the choices made by APs when there are multiple TPs. To illustrate, we consider

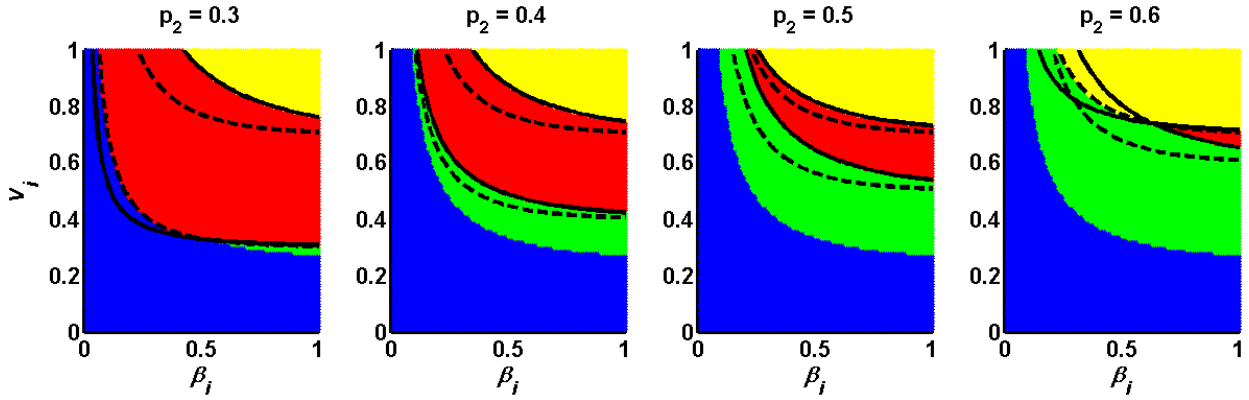


Fig. 2. Shift of market share for four TPs under $(q_1, q_2, q_3, q_4) = (1, 3, 5, 7)$ and $(p_1, p_3, p_4) = (0.7, 0.25, 0.1)$.

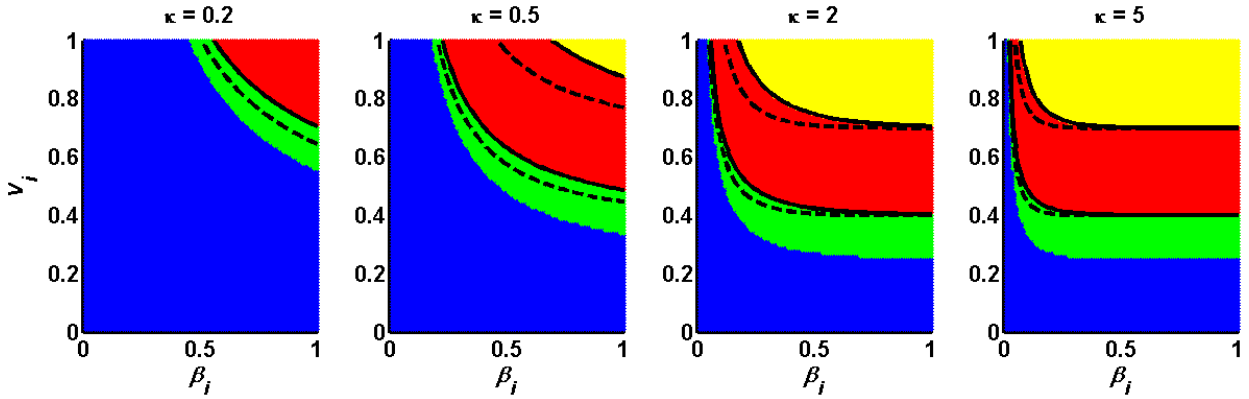


Fig. 3. Shift of market share for four TPs under $(q_1, q_2, q_3, q_4) = \kappa(1, 3, 5, 7)$ and $(p_1, p_2, p_3, p_4) = (0.7, 0.4, 0.25, 0.1)$.

the collective choices of the APs under a market of four TPs. In Figure 2, We fix the qualities to be $(q_1, q_2, q_3, q_4) = (1, 3, 5, 7)$ and the prices to be $(p_1, p_3, p_4) = (0.7, 0.25, 0.1)$ and vary p_2 from 0.3 to 0.6 in the four subfigures from left to right. In each subfigure, we vary β_i on the x-axis and v_i on the y-axis. Each point (β_i, v_i) on the plane represents a type of AP. The APs located on the top are more profitable and the APs located on the right are more sensitive to the quality of service. Notice from Figure 1 that a Netflix-type AP i, i.e., $\beta_i = 1$, would obtain around 40% and 5% of its maximum throughput under quality q_1 and q_2; however, under q_3 and q_4, its obtainable throughput almost reaches zero. Thus, APs with higher value of β_i will more likely choose higher quality TPs. The sets $\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3$ and $\mathcal{N}_4$ are shown in yellow, red, green and blue respectively. For example, $\mathcal{N}_1$ ($\mathcal{N}_4$) represents the set of APs that eventually choose the TP that provides the highest (lowest) quality with the highest (lowest) price. For any $I, J \in \mathcal{M}$, we define $\mathcal{N}_{IJ} = \{(\beta_i, v_i) : u_i(p_I, q_I) = u_i(p_J, q_J)\}$ to be the set of APs that obtain equal utility from I and J. In each subfigure, we plot $\mathcal{N}_{12}$ and $\mathcal{N}_{23}$ in solid lines and $\mathcal{N}_{13}$ and $\mathcal{N}_{24}$ in dashed lines. Thus, Figure 2 illustrates the *shift of market shares* for these four TPs when we vary the price p_2 of TP 2.

We make the following observations. First, with the increase (decrease) of p_2, $\mathcal{N}_2$ decreases (increases) monotonically. Second, if we keep increasing (decreasing) p_2 to p_1 (p_3), $\mathcal{N}_2$ ($\mathcal{N}_3$) will become empty. Third, the upper-right APs always choose TPs with better qualities (by Theorem 1). Finally, when $p_2 = 0.4$ or 0.5, each set $\mathcal{N}_i$ forms a distinct band; however, when $p_2 = 0.3$ and 0.6, we find $\mathcal{N}_3$ and $\mathcal{N}_2$ to be isolated regions respectively. We explain this by the convexity of TPs' pricing and the quasiconcavity of the APs' utility as follows.

Definition *1 (Convexity):* The pricing of $\mathcal{M}$ is *convex* if for any TPs $I, J, K \in \mathcal{M}$ with $q_I < q_K < q_J$, we have $p_K \leq \eta p_I + (1 - \eta) p_J$, where $\eta = (q_J - q_K)/(q_J - q_I)$.

The above definition is a discrete version of a continuous convex pricing function. Convex pricing often reflects the underlying convex cost structure where the marginal cost monotonically increases with the level of quality.

Definition *2 (Quasi-Concavity):* The utility function u_i is *quasi-concave* if the upper contour sets $\{(p_i, q_i) \in \mathbb{R}_+^2 : u_i(p_i, q_i) \geq u\}$ are convex for all $u \in \mathbb{R}$.

The quasi-concavity of the utility function implies that if two choices (p_1, q_1) and (p_2, q_2) provide at least u amount of utility for AP i, then any linear combination of the choices will induce at least that amount of utility for AP i. In practice, an AP often prefers better quality services until a certain level at which the price becomes a concern. Combined with a convex pricing, a quasi-concave utility function implies this kind of single-peak preference of the AP as follows.

Lemma *1 (Single-Peak Preference):* When the pricing of $\mathcal{M}$ is convex and u_i is quasi-concave, for any TPs $I, J \in \mathcal{M}$ with $u_i(p_I, q_I) > u_i(p_J, q_J)$, then $u_i(p_J, q_J) \geq u_i(p_K, q_K)$ if $q_I < q_J < q_K$ or $q_I > q_J > q_K$.

Lemma 1 gives a condition under which if an AP prefers a higher (lower) quality TP I over a lower (higher) quality TP J, then it prefers I over any TP whose quality is inferior (superior) to that of J. This condition will help us to understand the collective choice of APs of different types in Section II-C.

Lemma *2:* The utility function $u_i(p_I, q_I) = (v_i - p_I)\lambda_i(q_I)$ is quasiconcave in the domain $(v_i, \infty) \times \mathbb{R}_+$ if $\lambda_i(\cdot)$ is in the form of $\lambda_i(q_I) = \alpha_i e^{-\beta_i q_I}$.

Notice that when $p_2 = 0.4$, the pricing of $\mathcal{M}$ becomes convex and by Lemma 1 and 2, each AP has a single-peak preference among the TPs, where the bands show the preference peaks of the APs. When $p = 0.3$ or $p = 0.6$, the non-convexity in pricing induces non-single-peak preferences of some APs. For example, when $p_2 = 0.3$ ($p_2 = 0.6$), we can identify APs that prefer TP 2 and TP 4 (TP 1 and TP 3) over TP 3 (TP 2), where $\mathcal{N}_3$ ($\mathcal{N}_2$) shrinks to be an isolated region.

Let us illustrate the shift of market share when TPs vary their capacity. In Figure 3, we fix the prices $(p_1,p_2,p_3,p_4) = (0.7,0.4,0.25,0.1)$ and qualities $(q_1,q_2,q_3,q_4) = \kappa(1,3,5,7)$, and scale the capacities by $\kappa = 0.2, 0.5, 2$ and 5 from left to right. We observe that when the qualities degrade, APs' choices move to better quality TPs gradually (Theorem 3).

In summary, we presented a framework to help us to analyze (and understand) the APs' decision on choosing TPs based on each TP I's quality and price (q_I, p_I), and the AP i's profitability and sensitivity to quality (v_i, β_i). In reality, the prices of the TPs fluctuate due to competition. Next, we will study what affect the market prices and characterize the equilibrium market prices, which also depends on the traffic intensity α_i of the APs and the capacity ν_I of the TPs.

III. Market Equilibrium

In this section, we start with the definition of a market equilibrium, under which the prices of the TPs are stable and the claimed service qualities can be achieved when APs choose their best TPs. We then proceed to characterize the market equilibrium and calculate the equilibrium prices.

A. The Existence of Market Equilibrium

Although any TP I claims to provide service quality q_I, it cannot keep its promise if more APs choose this TP than its capacity can support. We model the achieved quality $Q_I(\lambda_I, \nu_I)$ as a function of the actual throughput λ_I going through I and its capacity ν_I.

Assumption *4:* The achieved quality $Q_I(\lambda_I, \nu_I)$ for any TP $I \in \mathcal{M}$ is non-decreasing in λ_I and non-increasing in ν_I.

Definition *3:* A set $\mathcal{X} \subseteq \mathcal{N}$ of APs is *feasible* for TP I with quality q_I, if $Q_I\big(\lambda_I(\mathcal{X}), \nu_I\big) \leq q_I$, where $\lambda_I(\mathcal{X}) = \sum_{i\in\mathcal{X}} \lambda_i(q_I)$ defines the induced throughput of the set $\mathcal{X}$ of APs under quality q_I.

In a market $\mathcal{M}$ of TPs, each TP would adjust its strategies to accommodate its customer APs' traffic demand and keep its service quality promise. For example, if the current capacity of TP I cannot support quality q_I, it might 1) expend its capacity ν_I, 2) increase price p_I, or 3) reduce the quality level q_I. Next, we define a market equilibrium where the APs' demand are *feasible* and the TPs' strategies are *stable*.

Definition *4:* Let p_I^{min} be the cost (or minimum price) of TP I. Let $\mathcal{M}'$ be identical to $\mathcal{M}$ except for $p_I' \neq p_I$ for some $I \in \mathcal{M}$ and $\mathcal{N}_I'$ be the set of APs choosing TP I under $\mathcal{M}'$. A system $(\mathcal{M}, \mathcal{N})$ forms a *market equilibrium* if 1) all APs' aggregate demands are feasible, i.e., $Q_I\big(\lambda_I(\mathcal{N}_I), \nu_I\big) \leq q_I$, for all $I \in \mathcal{M}$, and 2) each price p_I maximizes the utilization of capacity for acceptable throughput at TP I, i.e., for any $p_I' \geq p_I^{min}$ with the corresponding $\mathcal{N}_I'$ satisfying $Q_I\big(\lambda_I(\mathcal{N}_I'), \nu_I\big) \leq q_I$, we have $\lambda_I(\mathcal{N}_I') \leq \lambda_I(\mathcal{N}_I)$.

One way to understand the above definition of a market equilibrium is that given a set $\mathcal{N}$ of APs and a set $\{q_I : I \in \mathcal{M}\}$ of service qualities for them to choose from, the price p_I and capacity ν_I of each market segment should be consistent in that 1) when the APs make their choices of TP, their expected service quality can be achieved and, 2) the capacities of the TPs are not under-utilized, unless the charge p_I reaches the TP's cost p_I^{min}. If APs' quality expectations are not fulfilled, their choices of TP will change. Furthermore, if capacity ν_I is under-utilized with $p_I > p_I^{min}$, then the market segment I is not correctly priced. That being said, we assume that none of the market segment is controlled by a monopoly, which might want to under-utilize capacity and keep a higher price for profit-maximization. Notice that under a competitive market segment, the prices of the TPs cannot deviate too much and therefore, the goal of maximizing one's profit is the same as maximizing its capacity utilization. We will summarize and discuss the limitations of our model in Section V-D. The interesting aspect here is that although p_I, like all other prices, mainly depends on the supply ν_I and the demand $\mathcal{N}_I$ of the APs, all the TPs (or market segments) are correlated, which serve substitutions for the APs.

In practice, the TPs might not have enough capacities to accommodate all APs. As a result, market prices will rise and some APs cannot afford the prices and will not use any of the TPs. However, under Assumption 2, each AP needs to choose a TP even it cannot afford to use any of the TPs, so a market equilibrium might not exist under this assumption. To fix this minor technical issue, we make the following assumption to allow any AP not to use any of the TPs if they all induce negative utilities.

Assumption *5:* There always exists a dummy TP $D \in \mathcal{M}$ with quality $q_D = \infty$ and price $p_D = 0$.

By Assumption 3, quality q_D always induces zero throughput for any AP, and therefore, the dummy TP guarantees a zero utility and can accommodate as many APs as possible in equilibrium. Effectively, the set $\mathcal{N}_D$ models the APs that cannot afford to use any TP in the market in reality.

Theorem *4:* For any fixed set $\mathcal{N}$ of APs and any set $\mathcal{M}$ of TPs with fixed values of p_I^{min}, q_I and ν_I for all $I \in \mathcal{M}$, there exists a set $\{p_I : I \in \mathcal{M}\}$ of prices that makes $(\mathcal{M}, \mathcal{N})$ a market equilibrium.

Although TPs might be able to adopt new technologies to improve or differentiate their services, the quality that they can provide is often physically constrained by the nature of the TP, for example, if a TP is a Tier 1 ISP, it cannot guarantee end-to-end delays for the customers unless the access ISP's link is not congested. Similarly, although TPs might execute a long-term capacity planning, the supply of capacity does

not change in a small time scale. Compared to service quality and capacity, market prices change more frequently and easily. Theorem 4 says that even in a small time-scale where prices adapt to market conditions, prices might still converge to an equilibrium, which reflects the short-term market structure of the Internet ecosystem.

B. Characteristics of a Market Equilibrium

In theory, one might find multiple sets of prices that make $(\mathcal{M},\mathcal{N})$ a market equilibrium. For example, from any existing equilibrium, one might find a TP I such that with only a small change in p_I, no APs will change their choices. This new price also constitutes a market equilibrium. In practice, these price differences can happen by two reasons. First, even without a monopoly in a market segment, oligopolistic providers might implicitly collude on the price so that they keep a relatively high price simultaneously. When one of them starts to reduce price, the price of that segment will converge to a lower price. Second, the preferences of the APs are quite different so that the price change in one segment might not affect the demand choices of the APs.

Definition *5:* A market equilibrium $(\mathcal{M},\mathcal{N})$ is *competitive* if there does not exist any $p_I^{min} \le p_I' < p_I$ with the corresponding $\mathcal{N}_I^{'}$ satisfying $Q_I\big(\lambda_I(\mathcal{N}_I^{'}),\nu_I\big) \le q_I$.

If the AP types are very diverse or each market segment consists of many competing providers, one can focus on the above definition of a competitive market equilibrium. Technically, a competitive market equilibrium might not exist, since the minimum price might not exist when all the feasible equilibrium prices form an open set. However, prices in practice have a minimum unit, e.g., one cent, and we can always find such a competitive market equilibrium.

We will later show how to calculate competitive market equilibria. We would like to point out that our model is not limited to competitive market equilibria, i.e., if a segment I is not competitive enough, we can use a higher price for p_I. As a result, competitive equilibrium prices might be biased downward if the real market structure is not perfectly competitive; nevertheless, our qualitative results do not depend on whether the market equilibrium is competitive or not.

Theorem *5:* Let $\mathcal{N}' = \{(\kappa\alpha_i,\beta_i,v_i) : i \in \mathcal{N}\}$ and $\mathcal{M}' = \{(p_I,q_I,\kappa\nu_I) : I \in \mathcal{M}\}$ for some $\kappa > 0$. If $(\mathcal{M},\mathcal{N})$ is a market equilibrium and the quality function $Q_I(\cdot,\cdot)$s are homogeneous of degree 0, i.e., $Q_I(\lambda_I,\nu_I) = Q_I(\kappa\lambda_I,\kappa\nu_I)$, $\forall\kappa > 0, I \in \mathcal{M}$, then $(\mathcal{M}',\mathcal{N}')$ is a market equilibrium too.

Theorem 5 says that if the quality only depends on the ratio of the incoming traffic rate and the capacity, then when the number of APs (and their traffic intensity) and the capacities scale at the same speed, the original market equilibrium prices will remain in equilibrium. As a typical example, the quality function $Q_I(\lambda_I,\nu_I) = \lambda_I/\nu_I$ is homogeneous of degree 0, which models the capacity sharing [14], [18] nature of network services. If we consider the queueing delay as the quality metric, because of statistical multiplexing, the average queueing delay reduces when both arrival rate and service rate scales up at the same rate. In this case, Theorem 5 also implies that each TP I can accept more and more traffic for a fixed delay q_I, and as a consequence, the market prices will move downward in a new equilibrium.

C. Calculating Market Equilibrium Prices

We denote μ_I as the maximum throughput that TP I can accept when it can still fulfill the quality q_I, defined as

$$\mu_I = \arg\max_{\lambda_I} \; Q_I(\lambda_I,\nu_I) \le q_I. \tag{2}$$

For instance, if the quality metric is the average queueing delay under M/G/1 systems and TP I implements a FIFO scheduling policy, by the Pollaczek-Khinchine mean formula,

$$Q_I\big(\lambda_I,\nu_I\big) = \frac{\lambda_I}{\nu_I-\lambda_I}E[R],$$

where $E[R]$ is a constant that denotes the expected residual service time of jobs. If we want λ_I to be feasible, we need

$$\frac{\lambda_I}{\nu_I-\lambda_I}E[R] \le q_I \Rightarrow \lambda_I \le \frac{q_I}{E[R]+q_I}\nu_I = \mu_I.$$

We define $\eta_I = \mu_I/\nu_I$ as the maximum acceptable throughput per unit capacity, or the conversion factor from raw capacity to achievable throughput. Notice that given a fixed capacity ν_I, the smaller delay TP I wants to provide, the smaller maximum amount of traffic it can accept. For the M/G/1 case, η_I tends to 0 when the required quality q_I tends to 0, which also shows a convex cost structure for the TP.

Based on the monotonicity of Q_I (Assumption 4), we can translate the equilibrium conditions in terms of the maximum acceptable throughput μ_I rather than Q_I, q_I and ν_I. A market equilibrium can be characterized alternatively as follows.

Definition *6:* A system $(\mathcal{M},\mathcal{N})$ forms a *market equilibrium* if for all TP I, 1) $\lambda_I(\mathcal{N}_I) \le \mu_I$, and 2) there does not exist $p_I' \ge p_I^{min}$ with the corresponding $\mathcal{N}_I^{'}$ satisfying $\lambda_I(\mathcal{N}_I) < \lambda_I(\mathcal{N}_I^{'}) \le \mu_I$.

Based on the above alternative definition of a market equilibrium, we can calculate the competitive equilibrium prices without evaluating Q_I repeatedly as follows.

Calculate Price Equilibrium$\big(\mathcal{N}$**,** $\{p_I^{min},q_I,\nu_I : I \in \mathcal{M}\}\big)$

1. *Set* $p_I = \infty$ *for all TP* $I \in \mathcal{M}$;
2. *Calculate* μ_I *for all TP* $I \in \mathcal{M}$ based on q_I and Q_I;
3. **while** *there exists* $p_I' \in [p_I^{min},p_I)$ *such that* $\lambda_I(\mathcal{N}_I) \le \lambda_I(\mathcal{N}_I^{'}) \le \mu_I$
4. **set** $p_I = p_I'$;
5. **return** $\{p_I : I \in \mathcal{M}\}$;

In the above algorithm, we do not restrict which TP I to choose in step 3 if multiple TPs satisfy the condition. However, any sequence of updates will make the price vector converge, because each p_I will only be decreasing monotonically until convergence. Similarly, we can also set $p_I = p_I^{min}$ for all TPs, and the price vector will increase monotonically until convergence.

Based on Theorem 2 and 5, we have the following result.

Corollary *1:* Let $\mathcal{N}' = \{(\kappa\alpha_i,\kappa_3\beta_i,\kappa_1 v_i+\kappa_2) : i \in \mathcal{N}\}$ and $\mathcal{M}' = \{(\kappa_1 p_I+\kappa_2, q_I/\kappa_3,\nu_I') : I \in \mathcal{M}\}$ for positive

$\kappa, \kappa_1, \kappa_2, \kappa_3$ with $\mu'_I = \kappa\mu_I$ for all $I \in \mathcal{M}$. If $(\mathcal{M}, \mathcal{N})$ is a market equilibrium, then $(\mathcal{M}', \mathcal{N}')$ is a market equilibrium.

Although the prices of the TPs influence the APs' choices, which further affect the capacity utilization of the TPs, equilibrium prices are the fixed points in which both the APs' choices and the TPs' prices do not change. However, external factors could move the resulting equilibrium. In the next section, we will study these fundamental driving forces for the evolution of the Internet economic ecosystem. By understanding these factors, we will know why the market prices change and why certain evolutions happen.

IV. Price Dynamics in Equilibrium

In this section, we look deeper into the qualitative dynamics of the equilibrium market prices. In particular, we explore how the different characteristics of the APs and the TPs can affect the market prices in equilibrium.

A. Evaluation Setting

Each AP i is characterized by three parameters (α_i, β_i, v_i); each TP I is characterized by three parameters (p_I, q_I, ν_I). To make a fair comparison between equilibrium prices under different settings, we carefully normalize the system parameters as follows. We define $v_{max} = \max\{v_i : i \in \mathcal{N}\}$, $\beta_{max} = \max\{\beta_i : i \in \mathcal{N}\}$, and $p_{min} = \min\{p_I^{min} : I \in \mathcal{M}\}$. Based on Theorem 2, we normalize any system $(\mathcal{M}, \mathcal{N})$ by factors $\kappa_1 = 1/(v_{max} - p_{min})$, $\kappa_2 = p_{min}/(v_{max} - p_{min})$, and $\kappa_3 = 1/\beta_{max}$. As a result, we normalize each β_i or v_i within the interval $[0, 1]$ and the equilibrium prices will also be scaled accordingly with $[0, 1]$. If p_I^{scaled} is the derived market equilibrium price in the normalized system, we can recover the real market price p_I as $p_I = (v_{max} - p_{min})p_I^{scaled} + p_{min}$. When the normalized price p_I^{scaled} tends to 0, it reflects that the real market price p_I goes down to the cost p_{min}; when p_I^{scaled} tends to 1, it reflects that the real market price p_I goes to the maximum AP profitability v_{max}. We describe the TPs' capacity in terms of the maximum acceptable rates μ_Is. We define $\alpha = \sum_{i\in\mathcal{N}} \alpha_i$, $\mu = \sum_{I\in\mathcal{M}} \mu_I$ and the ratio $\rho = \mu/\alpha$. Based on Corollary 1, any price equilibrium sustains when α_is and μ_Is scale at the same rate. Thus, we normalize the APs' aggregate maximum traffic intensity α to be 1. We define $\sigma_I = \mu_I/\mu$ as the capacity share of TP I, and under the normalized system, each TP I has $\mu_I = \sigma_I\rho$.

After the above normalization, we can describe any system by the following four parameters:

1) a set of qualities $\{q_I : I \in \mathcal{M}\}$,
2) the normalized aggregate capacity ρ,
3) the distribution of α_i over the domain $[0, 1]^2$ of (β_i, v_i),
4) the capacity distribution $\{\sigma_I : I \in \mathcal{M}\}$.

We focus on three different quality types: 1) q_A, the highest quality for real-time content delivery, 2) q_B, medium quality, mostly for web applications, and 3) q_C, the best-effort quality, mostly for elastic traffic. As analyzed in [28], IP transit markets will be quite efficient if two tiers of services are provided; thus, q_B and q_C can be considered as the higher and lower tier services of such an IP transit market. To differentiate the three qualities, we set $q_A : q_B : q_C = 1 : 5 : 25$. We vary ρ from 0 to 1, where the system's total capacity varies from extremely scarce to abundant. We discretize the AP domain with 50 levels of v_i and β_i, with a minimum value of 0.02 and a maximum value of 1.0. This forms the 2500 types of APs used in our numerical evaluations in this section. We assume that APs' profitability and quality-sensitivity follow probability distributions F_v and F_β respectively, and α_i follows the joint distribution of F_v and F_β. We use the various distributions in Figure 4 for F_v and F_β. For instance, when a

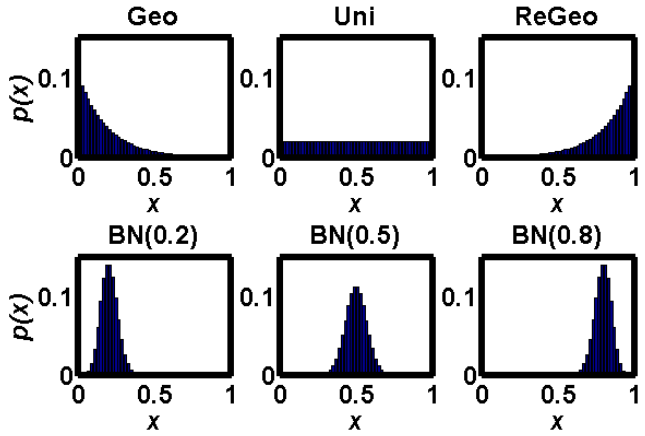


Fig. 4. Common distributions: geometric, uniform, reversed geometric, binomial with $p = 0.2, 0.5$ and 0.8.

geometric distribution Geo is used to describe F_β, it models the scenario where most of the AP traffic are elastic and the amount of quality-sensitive traffic decreases exponentially with its sensitivity level β_i. The binomial distributions $BN(p)$ are often used to approximate a normal distribution of the profitability v_i, or quality sensitivity β_i, where p determines the mean value.

B. Impact of TP Capacity on Prices

In this subsection, we study how the capacities of the TPs affect the equilibrium prices. We initially set $(q_A, q_B, q_C) = (0.2, 1, 5)$. We will evaluate how the quality may impact the equilibrium prices in the next subsection.

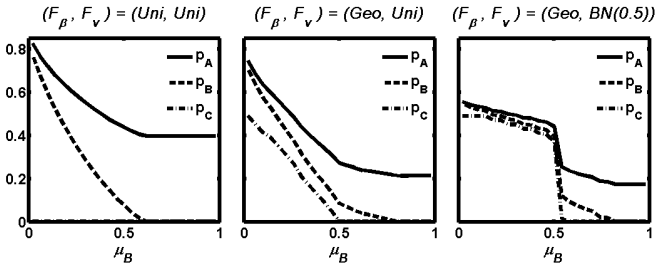


Fig. 5. Shift in market prices as μ_B varies: with $(q_A, q_B, q_C) = (0.2, 1, 5)$, $\mu_A = 0.05$ and $\mu_C = 0.25$.

In Figure 5, we fix $\mu_A = 0.05, \mu_C = 0.25$ and vary μ_B from 0 to 1 along the x-axis. The three sub-figures show the equilibrium prices when α_i follows the joint distributions of $(F_\beta, F_v) = (Uni, Uni), (Geo, Uni)$ and $(Geo, BN(0.5))$ respectively. We observe that when μ_B is scarce, equilibrium

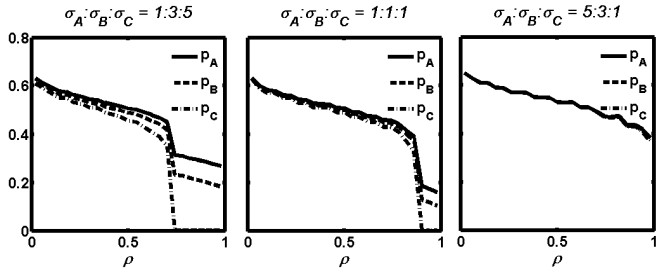


Fig. 6. Shift in market prices as ρ varies: with $(q_A, q_B, q_C) = (0.2, 1, 5)$ and $(F_\beta, F_v) = (Geo, Uni)$.

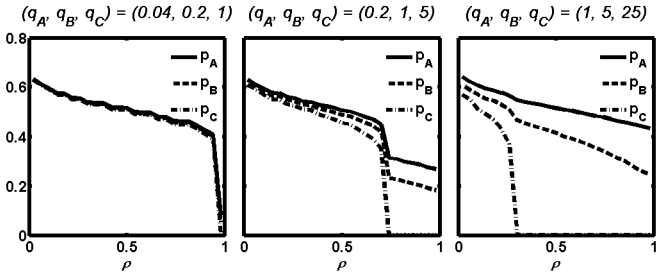


Fig. 7. Shift in market prices as ρ varies: with $(q_A, q_B, q_C) = \kappa(0.2, 1, 5)$ where $\kappa = 0.2, 1$ and 5.

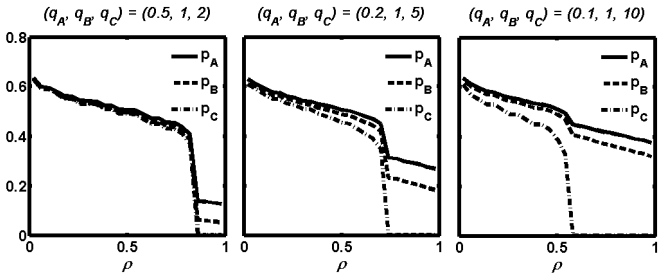


Fig. 8. Shift in market prices as ρ varies: with $q_B = 1$, $q_A : q_B = q_B : q_C = 1 : \kappa$ where $\kappa = 2, 5, 100$.

price p_B is close to (but strictly less than) the price p_A of its upper class TP. When μ_B increases, p_B diverges from p_A and moves to the price p_C of its lower class TP. When μ_B becomes abundant, its market price goes down to the minimum price after p_C. In general, when the capacity of a particular TP, i.e., μ_B, increases, it drives all equilibrium prices down; however, the prices of higher quality TPs, e.g., p_A, might not go down to the minimum price.

In the rest of this section, we often use $F_\beta = Geo$, which models the case where more APs were elastic, and $F_v = BN(0.5)$, which approximates that the AP profitability follows a normal distribution centered at $v_i = 0.5$. Note that our qualitative results do not depend on these settings.

In Figure 6, we vary the system capacity ρ from 0 to 1 along the x-axis. α_i follows the joint distribution $(F_\beta, F_v) = (Geo, BN(0.5))$. The sub-figures show the equilibrium prices when the capacity ratio $\sigma_A : \sigma_B : \sigma_C$ equals $1:3:5$, $1:1:1$ and $5:3:1$ respectively. In all three cases, when the total capacity ρ is small, all equilibrium prices are very close and high. When we increase ρ, all market prices drop. By comparing the price curves across the three subfigures, we observe that when the capacity share of the higher class TP is smaller (the left subfigure), 1) the three market prices differ more from each other, 2) p_C drops faster, and 3) all the prices drop to the minimum price faster than the other two cases. Because price differences exist in practice, we will use $\sigma_A : \sigma_B : \sigma_C = 1:3:5$ in the rest of this section.

Lessons (the TP capacity effects on prices) learned:

- Capacity expansion drives market prices down.
- The capacity expansion of a particular TP I would affect not only its own price p_I, but also other TPs' prices, due to the substitution effect of TP I to other TPs.
- When TP I's capacity share σ_I is small (big), its price p_I is close to that of its next higher (lower) class TP.

C. Impact of TP Quality on Prices

Let us explore how the quality q_I of the TPs may affect the equilibrium prices. We use the setting that the capacity distribution follows $\sigma_A : \sigma_B : \sigma_C = 1:3:5$ and α_i follows the joint distribution $(F_\beta, F_v) = (Geo, BN(0.5))$.

In Figure 7, we keep the quality ratio $q_A : q_B = q_B : q_C = 1:5$ and use $(q_A, q_B, q_C) = \kappa(0.2, 1, 5)$, where κ equals $0.2, 1$ and 5 in the three subfigures. We vary the system capacity ρ from 0 to 1 along the x-axis. We observe that when all the TPs improve quality by the same ratio, i.e., $\kappa = 0.2$, the market prices of the TPs are very close; when all the TPs degrade their quality by the same ratio, i.e., $\kappa = 5$, the market prices of the TPs diverge greatly. This observation can be explained by Theorem 3. When κ decreases and all qualities are improved, more APs will choose lower class TPs, which move the prices of the lower class TPs upward and the prices of upper class TPs downward. As a result, all TPs prices will move closer. On the other hand, when κ increases and all qualities are degraded, more APs will choose to upper class TPs, which move the prices of the upper class TPs upward and the prices of lower class TPs downward. This will further diverge the price differences among the TPs of different qualities.

In Figure 8, we keep $q_B = 1$ and vary the quality ratio $q_A : q_B = q_B : q_C = 1 : \kappa$, where κ equals $2, 5$ and 10. We observe that the price differences are positively correlated with the quality ratio. In particular, when quality ratio is high, e.g., $\kappa = 10$, the price of the lowest class TP, i.e., p_C, drops earlier and sharper when the total capacity ρ expands. At the same time, higher class TPs can still maintain a non-zero market price even after p_C drops down the minimum price. The general trend is that when the quality ratio keeps increasing, the price curves will move higher and toward the left. In the rest of this section, we will often use the quality ratio $1:5$ and $(q_A, q_B, q_C) = (0.2, 1, 5)$ for our evaluations. Again, our qualitative results do not depend on this setting.

Lessons (the TP quality effects on prices) learned:

- The market prices of the TPs would be close to (far from) one another if the quality ratio is small (big) or/and the overall qualities of the market are high (low).
- In reality, the qualities provided by the TPs are becoming better and better, which implies that market prices for different services might converge.
- High-end market segments can still maintain a price difference if they can differentiate their quality from the

lower class TPs substantially.

Next, we will see that the TP price differences also depend on the demand side: the characteristics of the APs.

D. Impact of AP Wealth on Prices

Let us explore how the profitability distribution F_v may affect the equilibrium prices. We still keep $\sigma_A : \sigma_B : \sigma_C = 1 : 3 : 5$ and $(q_A, q_B, q_C) = (0.2, 1, 5)$. α_i follows the joint distribution of F_β and F_v, where F_β is distributed as Geo.

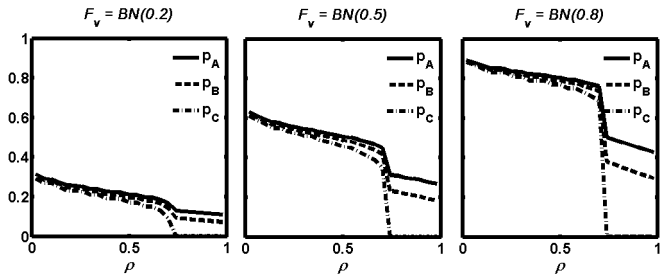


Fig. 9. Shift in market prices as ρ varies: $(q_A, q_B, q_C) = (0.2,1,5)$, $\sigma_A : \sigma_B : \sigma_C = 1:3:5$, $F_\beta = Geo$.

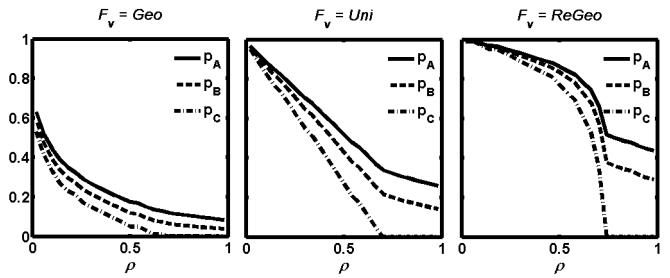


Fig. 10. Shift in market prices as ρ varies: $(q_A, q_B, q_C) = (0.2,1,5)$, $\sigma_A : \sigma_B : \sigma_C = 1:3:5$, $F_\beta = Geo$.

In Figure 9, we vary ρ from 0 to 1 along the x-axis and plot the equilibrium prices where the profitability distribution F_v follows a binomial distribution $BN(p)$ parameterized by $p = 0.2, 0.5$ and 0.8 respectively. By doing this, we simulate the normal distributions of the APs' wealth varying the mean value from small to large. We observe that despite the difference in mean profit of the APs, p_C drops to the minimum price at the same time. The price curves in all cases keep the same shape; however, they scale differently on the vertical axis. This indicates that the market prices depend on how much the APs are able to pay for the services, and how they demand for the TPs based on their values of (β_i, v_i).

In Figure 10, we vary F_v to be Geo, Uni and $ReGeo$. We observe that the shapes of the price curves are very different: prices decrease convexly, linearly and concavely in the three subfigures. In general, how fast the prices drop depends on the density of the APs whose profitability are around that price range, and the shape of the curves look like the complimentary cumulative distribution function (CCDF) of F_v.

Lessons (the AP wealth effects on prices) learned:

- The market prices of the TPs are positively correlated with the mean profitability of the APs.
- At a certain price range where the density of the APs is high (low), more (less) competition among the APs drives the prices close to (far below) their profitability.

E. Impact of AP Quality-Sensitivity on Prices

In this subsection, we study how the quality sensitivity distribution F_β affects the equilibrium prices. We set $\sigma_A : \sigma_B : \sigma_C = 1 : 3 : 5$ and $\rho = 0.5$. In the following cases, F_β follows a binomial distribution $BN(p)$, where we vary the parameter p along the x-axis. By doing this, we simulate the cases where the APs become more and more sensitive to quality when the mean sensitivity increases with p.

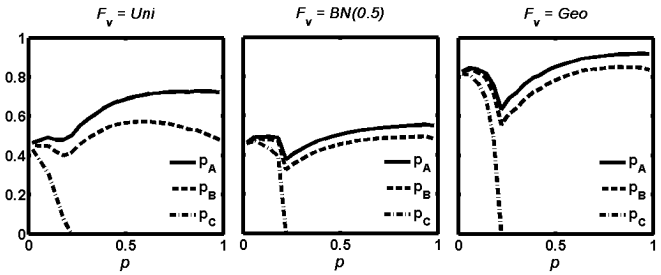


Fig. 11. Shift of market prices when we vary AP's sensitivity to quality: with $\sigma_A : \sigma_B : \sigma_C = 1 : 3 : 5$, $\rho = 0.5$, $(q_A, q_B, q_C) = (0.2, 1, 5)$ and $F_\beta = BN(p)$.

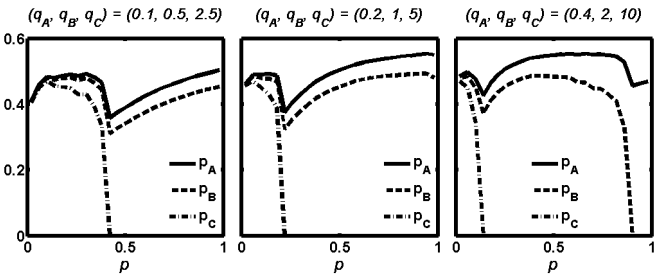


Fig. 12. Shift of market prices when we vary AP's sensitivity to quality: with $\sigma_A : \sigma_B : \sigma_C = 1 : 3 : 5$, $\rho = 0.5$, $F_v = BN(0.5)$ and $F_\beta = BN(p)$.

In Figure 11, we fix $(q_A, q_B, q_C) = (0.2, 1, 5)$ and vary F_v to be Uni, $BN(0.5)$ and Geo in the three sub-figures. We observe that although the profitability distribution affect the absolute price values, the shape of the price curves look similar. When the quality sensitivity of the APs increases, the lowest quality service price, i.e., p_C, drops sharply and quickly. Although p_A and p_B drops accordingly with p_C, after p_C reaches the minimum price, both p_A and p_B rebound. With further increase in quality sensitivity, p_B shows a trend to decrease slowly; however, p_A always stays at a high level. When the APs become more sensitive to quality, more and more APs start to move to higher class TPs. As a result, the capacity μ_C becomes under-utilized, which also drives p_C down very quickly. Although p_C's drop pulls down the overall market prices, more APs move to higher class TPs, which make TPs A and B in demand, and therefore, keep p_A and p_B steadier. After p_C reaches the minimum price, p_C stops decreasing. As the APs' quality sensitivity keeps increasing, even the minimum market price of p_C becomes

relatively expensive to the APs. This makes even more APs move to TP A and B and drives p_A and p_B upward.

In Figure 12, we fix $F_v = BN(0.5)$ and vary the qualities to be $(q_A, q_B, q_C) = \kappa(0.2, 1, 5)$, where $\kappa = 0.5, 1$ and 2. We observe the same trends as in Figure 11 that p_C drops quickly and sharply to the minimum price as the APs' quality sensitivity increases. As κ increases, all the price curves move to the left and the price drop of lower class TPs becomes quicker and sharper. This also coincides with the observations made in Figure 7 that when the qualities degrade, the price of the lower class TP drops much quicker.

In the above illustrations, we vary the distribution F_β. It is also possible that all the APs' sensitivity increase by $\beta_i' = \xi\beta_i$ for some $\xi > 1$. By Theorem 2, we can rescale the system by $\kappa_3 = 1/\xi$, as if the APs keep their quality sensitivity constant and all the qualities become poorer. By Theorem 3 and the TP quality effect result in Figure 7, we also conclude that more APs will prefer higher quality TPs and the price of the lower quality TPs will drop sharply.

Lessons (the AP quality-sensitivity effects) learned:

- When the APs become more sensitive to the service quality, the price of lower class TPs will drop quickly.
- When the price of the lowest quality TP goes down to its cost, the prices of higher quality TPs might increase due to their relatively cheap prices and high demand.

F. Internet Evolution: Some Explanations

By understanding the factors that drive the market equilibrium, we reason about the evolution of the Internet ecosystem and reach plausible answers to the questions raised in Section I. We do not claim that our answers below are exhaustive and the limitations of our model will be discussed in Section V-D.

1) Why have the IP transit prices been dropping? The capacity effect tells that the price drop can be a consequence of the capacity expansion of the transit providers. Compared to the capacities at the last-miles, the capacity in the backbone grows faster than demand and is abundant [15]. Also, the price drop in better quality services, i.e., CDN prices, will drive the transit prices further down. The quality effect tells that when the transit quality differs a lot from the CDN services, the prices will diverge greatly. The wealth effect tells that since the majority of the elastic APs might not be very profitable, transit providers cannot fully utilize its capacity and charge a high price at the same time. This is also why they are looking for providing value-added and differentiated services. Last, the AP quality-sensitivity effect tells that when AP traffic becomes more and more inelastic, e.g., the surge of Netflix traffic, lower quality service will become less valuable and therefore its price will drop quickly.

2) Why have the CDNs emerged in the ecosystem? The capacity effect tells that when the capacity of higher quality service is small, it can maintain a price difference with the lower quality services. The quality effect tells that if a CDN service's quality differs a lot from the transit services, it can be priced much higher. When the capacity of the transit market was limited and priced high, the demand for even higher quality service drove the price for potential CDN services even higher. This explains why CDNs emerged in the first place. The wealth effect tells that when the APs' profitability is not high, the market prices cannot be high. However, due to the low cost structure of the CDNs, they can still help small APs who could not afford the infrastructure to support large demand. The AP quality-sensitivity effect further tells that with the traffic being more and more sensitive to quality, the price of high quality CDN can sustain at a high level.

3) Why has the pricing power shifted to the access ISPs? This can be partially explained by the AP quality-sensitivity effect and the TP quality effect. When the AP traffic becomes more and more sensitive to service quality, they are more willing to pay for the higher quality services. Because access ISPs are physically closer to the users, their service quality is naturally much better than other providers who have to go through the access ISPs to reach the end-users anyways. Consequently, the difference in service quality makes it possible for the access ISPs to charge services at higher prices. Furthermore, Comcast's monopolistic position in the U.S. market could be another reason, under which its price will be set higher than the competitive market price under Definition 5.

4) Why are the large content providers building their own wide-area networks toward users? Mostly because the APs become more sensitive to service quality, they cannot rely on the transit providers to deliver content. As high quality services are limited and access ISPs would obtain more pricing power, large APs might consider establishing their own networks toward users as a cheaper alternative than paying access ISPs for better services in the future.

V. Internet's Economic Evolution

Besides understanding how each isolated factor might affect the market prices, we incorporate ground truth data [10], [21], [4], [11], e.g., the historical trends of the TPs' capacity expansion and the APs' characteristics, and project *possible future price dynamics* of the Internet ecosystem. Through this, our model can help the TPs make various long-term business decisions. Let us demonstrate this.

We take a macroscopic view and categorize network services as two types: $\mathcal{M} = \{A, B\}$. B models the IP transit service that provides interconnection based on "best-effort"; A models the CDN or private peering type of service that provides better service quality than B. We categorize the APs as three types: $\mathcal{N} = \{a, b, c\}$. a models the video or realtime interactive applications that are very sensitive to quality. b models the web applications that are elastic but more tolerate to quality than type a applications. c models the inelastic applications, e.g., email and P2P file download.

By Corollary 1, we know that when quality and the sensitivity parameters scale inversely, the equilibrium remains the same; therefore, without loss of generality, we set $q_B = 1$ as the baseline best-effort quality level. We set the quality sensitivity parameters to be $(\beta_a, \beta_b, \beta_c) = (10, 1, 0.1)$. Under this setting, type a APs would only obtain $e^{-10} \approx 4.5^{-5}$ of their maximum throughput under q_B, which implies that the best-effort service cannot support quality sensitive applications. Also, under q_B, a type b AP could get $e^{-1} \approx 37\%$

of its maximum throughput; however, a type c AP could get $e^{-0.1} \approx 90\%$ of its maximum throughput. When measured by delay, the quality of service for realtime applications often require the delay to be at the order of milliseconds [30], compared to the best-effort service delays at the order of seconds. Thus, we choose $q_A = 0.01$ to reflect the same order of magnitude of service difference. As a result, even type a APs would obtain $e^{-0.1} \approx 90\%$ of their maximum throughput under the better quality level q_A.

Next, we try to estimate the capacity of the TPs on the Internet. We take the Equinix Internet Exchange at New York (Equinix-NY) as a reference market and estimate the capacities based on the data provided by PeeringDB [10]. At the end of year 2011, there were 102 ISPs listed on at Equinix New York Exchange in PeeringDB, among which 44 use *Open* peering policy and the remaining 58 use either *Selective* or *Restricted* peering policy. The total capacity was around 21 Tbps, among which the ISPs using Open peering policy contributed 7 Tbps and the remaining ISPs contributed 14 Tbps. Since *Selective* and *Restricted* policies are used for private and often paid-peering agreements, we set ν_A and ν_B to be 14 and 7 Tbps, for the reference time of the year 2011.

From the Global Internet Geography [4] report, between 2007 to 2011, the international Internet capacity increased six-fold and the bandwidth to the U.S. had increased nearly 50 percent per year. To a first approximation, we assume that the capacity ν of the TPs increases 50% per year. We define $\alpha = \alpha_a + \alpha_b + \alpha_c$ and ω_a, ω_b and ω_c as the weight of the throughput upper bound of each application type. Given α and the weight of AP i, we obtain α_i as $\alpha_i = \omega_i \alpha / (\omega_a + \omega_b + \omega_c)$ for all $i = a, b, c$. Based on the observed traffic distribution of various applications in [21], we set $(\omega_a, \omega_b, \omega_c) = (2\%, 75\%, 23\%)$ for the year 2007, and assume that the weight for video (ω_a), web (ω_b) and inelastic applications (ω_c) increase at an annual growth rate of 150%, 50% and 20% respectively. Notice that IP transit prices are often quoted for per Mbps-month, while CDN prices are often quoted for per terabit. If capacity is fully utilized 24/7, \$1 per Mbps-month can be translated into \$0.386 per terabit. We assume that the maximum per unit traffic revenue for the APs is \$10 Mbps-month and the APs' revenue are uniformly distributed.

A. A First Approximation Benchmark

We use our macroscopic model to fit the historical prices starting from 2007 to 2011 and project future Internet prices of 2012 to 2014. In a first approximation, we choose the following parameters.

1) α at year 2007 (denoted as α_{07}) equals 10 Tbps.
2) α increases at an annual growth rate $r_\alpha = 22\%$.
3) $\eta_A = \mu_A / \nu_A = 0.3$ and $\eta_B = \mu_B / \nu_B = 0.9$.

In Figure 13, the upper left subfigure plots the achievable throughput for the CDN (μ_A) and IP transit (μ_B) services from 2007 to 2014 and the lower left subfigure plots the maximum demand α_a, α_b and α_c for the same time period. The upper right subfigure plots the price dynamics of both IP transit and CDN services and the lower right subfigure plots the percentage of price change for both services. We observe that

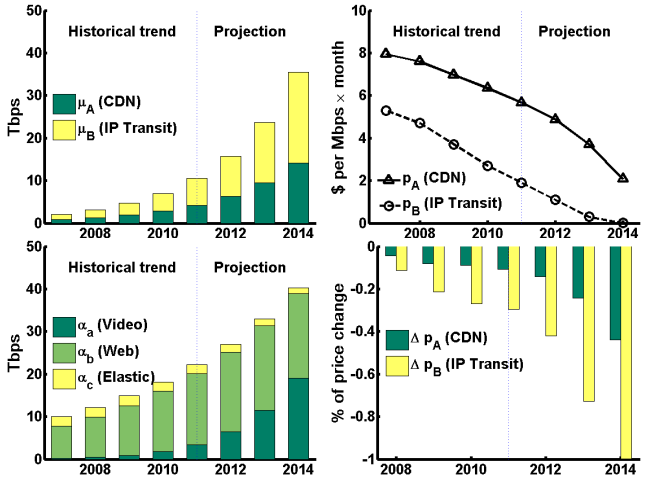


Fig. 13. Historical price and future price projection.

average price drop from 2007 to 2011 is approximately 20%, which coincides with the price drop surveyed in the Global Internet Geography [4] report. Also, the price of IP transit is below \$2 per Mbps-month, very close to the mean of IP transit prices, where the lowest price fell to \$1 per Mbps-month.

Compared to the video delivery pricing [11], our price projection shows that the CDN price drops around 8% annually from 2007 to 2011, and reaches \$5.67 per Mbps-month, or \$2.18 per terabit. This price is lower than the \$7.5 per terabit price for APs with volume of 5PB data and the price drop is slower than the observed 20% price drop in the CDN industry [11]. The difference could come for two reasons: 1) since CDN service charges based on traffic volume, we cannot assume that the APs would always use the capacity 24/7, and therefore, the CDN providers should charge some premium on top of the basic per Mbps per month charge, 2) in contrast to our competitive model for CDN service, the industry might be less competitive and could charge a much higher price; therefore, when the industry becomes more competitive, we expect to see much sharper price drops.

Based on the trend from 2007 to 2011, our model projects that both the IP transit and CDN prices will further drop, at an even faster rate, and IP transit price will drop to its minimum price. Of course, this projection is based on the assumption that the capacity of the TPs will keep expanding at the 50% annual rate. We will further discuss potential trends of future prices in a later subsection.

B. Sensitivity of the Benchmark

In this subsection, we show the sensitivity of our price projection with respect to the chosen parameters.

First, we want to see how the demand parameter α affects the price dynamics. Figure 14 shows a projection of service prices when the initial value α_{07} and the growth rate r_α change. In the left subfigure, we vary α_{07} to be 9 and 11 Tbps compared to the benchmark value of 10 Tbps. We observe that the prices are positively correlated with α_{07}. In the right subfigure, we vary the growth rate r_α to be 20% and 24% compared to the benchmark value of 22%. We observe that

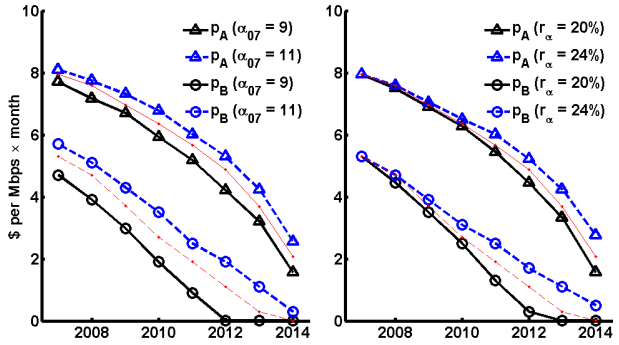


Fig. 14. Sensitivity to initial demand α_{07} and rate r_α.

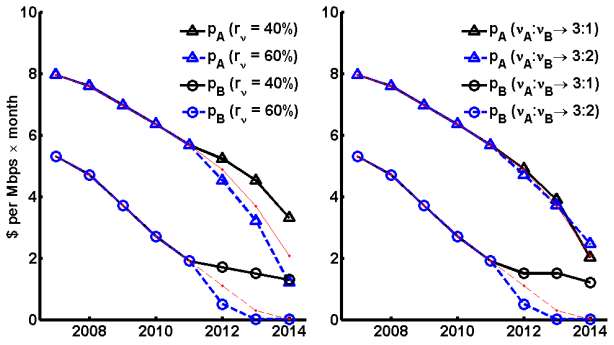


Fig. 16. Price projection under various capacity ratios $\nu_A : \nu_B$ and capacity expansion rate r_ν.

the prices again are positively correlated with r_α. Both tells that when the demand increases, so do the prices.

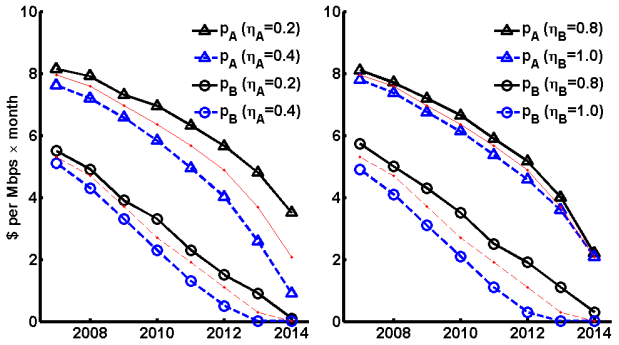


Fig. 15. Sensitivity to capacity utilization η_A and η_B.

Second, we want to see how the capacity utilization factor $\eta_I = \mu_I/\nu_I$ affects the price dynamics. Figure 15 shows a projection of Internet service prices when η_A and η_B vary. In the left subfigure, we vary η_A to be 0.2 and 0.4 compared to the benchmark value 0.3. In the right subfigure, we vary η_B to be 0.8 and 1.0 compared to the benchmark value 0.9. We observe that the all prices are negatively correlated with the capacity utilization factors. Also, the IP transit prices are sensitive to both η_A and η_B; while the CDN prices are only sensitive to its utilization factor η_A.

C. Price Projection and TP Business Decisions

Now, we demonstrate that by using the price projection from our model, we can help the TPs to make business decisions on 1) how aggressive the TPs should expend their capacity, and 2) whether the TPs should/would tend towards *Open* or *Selective* peering policies.

To answer the first question, we vary the capacity growth rate of the TPs r_ν to be 40% and 60%, compared to the historical benchmark rate 50% and plot the price projections in the left subfigure of Figure 16. We observe that when the capacity grows at 60% per year, both the CDN and IP transit price drop fast and the IP transit price will down to its cost next year; however, when the growth rate is 40%, the IP transit price will be decreasing at a very slow rate. These observations tell us that the ISPs providing IP transit services might want to slow down their investment in capacity expansion; however, CDN providers and ISPs that sell private-peering and QoS might want to continue to expand their capacity when their profit margins are still above zero. As the price of IP transit drops, we believe that the investment in the transit capacity will slow down, which will also stabilize the price of the IP transit services.

To answer the second question, we vary the capacity ratio $\nu_A : \nu_B$ from the benchmark ratio $2:1$ (14 Tpbs : 7 Tbps) to $3:1$ and $3:2$ for the year 2014 in the right subfigure. These two projections model the scenarios where ISPs will tend to be more *Selective* and more *Open* in their peering policies respectively. We observe that if more ISPs are going to use an *Open* peering policy, the IP transit price will drop to its cost quickly; otherwise, it will get closer to the CDN price and be stable. This observation implies that ISPs would have strong incentives to move towards *Selective* peering policies if possible, which coincides with the reality that the access ISP, Comcast, started to use private peering exclusively.

In summary, we predict that although the CDN price will still be dropping, the price of IP transit will be more stable. Furthermore, the capacity expansion will slow down and more ISPs will tend to use *Selective* rather than *Open* peering policies in the near future.

D. Limitations of the Model

Although we have demonstrated potential usages of our model, we want to mention the limitations of the model so as to avoid misinterpreting the results obtained from our macroscopic model.

First of all, our general equilibrium model implicitly assumes that each market segment is competitive. In practice, some market segment could be lack of competition and form a monopolistic or oligopolistic market structure. Thus, the real market prices will be higher than what our model predicts. Second, our equilibrium model does not capture the off-equilibrium and transit dynamics that could happen in practice. Third, our model is in nature macroscopic, and it does not capture detailed information like peering agreement, topology, traffic patterns and etc. Nevertheless, our model does capture the type of different services the TPs provide via implicitly encoding all the relevant information into the quality level q_I. From the APs' point of view, they do care about *quality* rather than other details of the TPs. Fourth, since our focus

is on the transit/CDN market, our model does not intend to capture the end-user market aspects. For example, modeling the bundle of access services and other service differentiations are out of scope. Last but not the least, our macroscopic model provides some qualitative reasons for the Internet evolution, which we do not claim to be exhaustive. There might be additional factors/reasons that are not captured by our model, e.g., the lack of competition in the market.

VI. Related Work

Many empirical studies have been tracking the evolution of the Internet using measurements and public data sets [21], [16], [19], [27], [12]. Labovitz et al. [21] measured the inter-domain traffic between 2007 and 2009, and observed the changes in traffic patterns as well as the consolidation and disintermediation of the Internet core. Gill et al. [19] collected and analyzed traceroute measurements and showed that large content providers are deploying their own wide-area networks. Dhamdhere et al. [16] confirmed the consolidation of the core of the Internet, that brings the content closer to users. Akella et al. [12] used measurements to identify and characterize non-access bottleneck links in terms of their location, latency and available capacity. At the edge of the Internet, Sundaresan et al. [27] studied the network access link performance measured directly from home gateway devices. We focus on a macroscopic model of the Internet ecosystem that captures the application traffic going through the network transport service providers.

Many works [13], [17], [24], [25], [28], [22], [20] focused on the modeling perspective of the Internet evolution. Chang et al. [13] presents an evolutionary model for the AS topologies. Lodhi et al. [22] used an agent-based model to study the network formation of the Internet. Motiwala et al. [25] used a cost model to study the Internet traffic. Valancius et al. combined models and data to study the pricing [28] structure of the IP transit market. Faratin et al. [17] and Ma et al. [24], [23] studied the evolution of the ISP settlements. In this work, we take a holistic view and analyze the business decisions and evolutions of the APs and TPs altogether.

VII. Conclusions

In this paper, we proposed a network aware, macroscopic model to explain the evolution of the Internet. This model captures 1) the business decisions of the APs, 2) the pricing and competition of the TPs, and 3) the resulting market equilibrium of the ecosystem. By analyzing how the AP characteristics (i.e., traffic intensity, profitability and sensitivity to service quality), and the TP characteristics (i.e., quality, price and capacity, affect the market equilibrium), we obtain fundamental understanding of why historical and recent evolutions of the Internet have happened. With further estimations of the trends in traffic demand, capacity growth and quality improvements, our model can also project the future evolution of the Internet ecosystem. This model provides a tool for the Internet players to better understand their business and risks, and help them to deal with their business decisions in the complicated and evolving ecosystem.

References

[1] Akamai. http://www.akamai.com/.
[2] Amazon Elastic Compute Cloud (EC2). http://www.amazon.com/ec2.
[3] Cogent Communications, Inc. http://www.cogentco.com.
[4] "Global Internet Geography." Telegeography Research. http://www.telegeography.com/.
[5] Level 3 Communications, Inc. http://www.level3.com.
[6] Limelight Networks. http://www.limelight.com/.
[7] Netflix, Inc. http://www.netflix.com.
[8] Netflix takes up 32.7% of Internet bandwidth, CNN New. http://edition.cnn.com/2011/10/27/tech/web/netflix-internet-bandwith-mashable.
[9] Netflix Traffic Now Bigger Than BitTorrent. Has Hollywood Won? GIGAOM News. http://gigaom.com/broadband/netflix-p2p-traffic/.
[10] PeeringDB. http://www.peeringdb.com/.
[11] Video Delivery Pricing for Q4 2011. http://www.cdnpricing.com/.
[12] A. Akella, S. Seshan, and A. Shaikh. An empirical evaluation of wide-area Internet bottlenecks. In *Proceedings of the ACM conference on Internet measurement (IMC)*, 2003.
[13] H. Chang, S. Jamin, and W. Willinger. To peer or not to peer: Modeling the evolution of the Internet's AS-level topology. In *Proceedings of IEEE Infocom*, Barcelona, Spain, 2006.
[14] C.-K. Chau, Q. Wang, and D. M. Chiu. On the viability of Paris metro pricing for communication and service networks. *Proceedings of IEEE INFOCOM*, 2010.
[15] C. Courcoubetis and R. Weber. *Pricing Communication Networks: Economics, Technology and Modelling*. John Wiley & Sons Ltd., 2003.
[16] A. Dhamdhere and C. Dovrolis. Ten years in the evolution of the Internet ecosystem. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (IMC 08)*, pages 183–196, Vouliagmeni, Greece, October 2008.
[17] P. Faratin, D. Clark, P. Gilmore, S. Bauer, A. Berger, and W. Lehr. Complexity of Internet interconnections: Technology, incentives and implications for policy. *The 35th Research Conference on Communication, Information and Internet Policy (TPRC)*, 2007.
[18] R. Gibbens, R. Mason, and R. Steinberg. Internet service classes under competition. *IEEE Journal on Selected Areas of Communications*, 18(12), December 2000.
[19] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. The flattening Internet topology: natural evolution, unsightly barnacles or contrived collapse? In *Proceedings of the 9th international conference on Passive and active network measurement*, 2008.
[20] H. Haddadi, S. Uhlig, A. Moore, R. Mortier, and M. Rio. Modeling Internet topology dynamics. *ACM SIGCOMM Computer Communication Review, Volume 38 Issue 2, April 2008*.
[21] C. Labovitz, D. McPherson, S. Iekel-Johnson, J. Oberheide, and F. Jahanian. Internet inter-domain traffic. In *Proceedings of the ACM SigComm*, New Delhi, India, 2010.
[22] A. Lodhi, A. Dhamdhere, and C. Dovrolis. GENESIS: An agent-based model of interdomain network formation, traffic flow and economics. In *Proceedings of IEEE Infocom*, Miami FL, March 2012.
[23] R. T. B. Ma, D. Chiu, J. C. Lui, V. Misra, and D. Rubenstein. Internet Economics: The use of Shapley value for ISP settlement. *IEEE/ACM Transactions on Networking*, 18(3), 2010.
[24] R. T. B. Ma, D. Chiu, J. C. Lui, V. Misra, and D. Rubenstein. On cooperative settlement between content, transit and eyeball Internet service providers. *IEEE/ACM Transactions on Networking*, 19(3), 2011.
[25] M. Motiwala, A. Dhamdhere, N. Feamster, and A. Lakhina. Towards a cost model for network traffic. *ACM SIGCOMM Computer Communication Review*, 42(1), January 2012.
[26] W. Norton. *The Internet Peering Playbook: Connecting to the Core of the Internet*. DrPeering Press, 2011.
[27] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescap. Broadband Internet performance: A view from the gateway. In *Proceedings of ACM SIGCOMM*, Toronto, Ontario, Canada, 2011.
[28] V. Valancius, C. Lumezanu, N. Feamster, R. Johari, and V. Vazirani. How many tiers? Pricing in the Internet transit market. In *Proceedings of ACM SIGCOMM*, Toronto, Ontario, Canada, August 2011.
[29] T. Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 141, 2005.
[30] X. Xiao. *Technical, Commercial and Regulatory Challenges of QoS: An Internet Service Model Perspective*. The Morgan Kaufmann Series in Networking, 2008.

APPENDIX

Proof of Theorem 1: We prove by contradiction. Assume $q_{I_i} < q_{I_j}$, by Assumption 1 and 2, we know that $p_{I_i} > p_{I_j}$. By Assumption 2, we further know that $(v_i - p_{I_i})e^{-\beta_i q_{I_i}} \geq (v_i - p_{I_j})e^{-\beta_i q_{I_j}}$ and $(v_j - p_{I_j})e^{-\beta_j q_{I_j}} \geq (v_j - p_{I_i})e^{-\beta_j q_{I_i}}$. From the above two inequalities, we can derive

$$\frac{v_i - p_{I_i}}{v_j - p_{I_i}} e^{-(\beta_i - \beta_j) q_{I_i}} \geq \frac{v_i - p_{I_j}}{v_j - p_{I_j}} e^{-(\beta_i - \beta_j) q_{I_j}}.$$

However, because $p_{I_i} > p_{I_j}$, we have $\frac{v_i - p_{I_i}}{v_j - p_{I_i}} < \frac{v_i - p_{I_j}}{v_j - p_{I_j}}$ if $v_i \neq v_j$ and $\frac{v_i - p_{I_i}}{v_j - p_{I_i}} = \frac{v_i - p_{I_j}}{v_j - p_{I_j}}$ if $v_i = v_j$. Also, because $q_{I_i} < q_{I_j}$ and $\beta_j \geq \beta_i$, we have $e^{-(\beta_i - \beta_j) q_{I_i}} < e^{-(\beta_i - \beta_j) q_{I_j}}$ if $\beta_i \neq \beta_j$ and $e^{-(\beta_i - \beta_j) q_{I_i}} = e^{-(\beta_i - \beta_j) q_{I_j}}$ if $\beta_i = \beta_j$. Because $(\beta_j, v_j) \neq (\beta_i, v_i)$, we cannot have the equality condition for both of the above, and therefore, we derive the contradiction that $\frac{v_i - p_{I_i}}{v_j - p_{I_i}} e^{-(\beta_i - \beta_j) q_{I_i}} < \frac{v_i - p_{I_j}}{v_j - p_{I_j}} e^{-(\beta_i - \beta_j) q_{I_j}}$. ∎

Proof of Theorem 2: By definition, $u_i(p_I, q_I) = \alpha_i(v_i - p_I)e^{-\beta_i q_I}$. Thus, under the scaled system, we have $u_i(p'_I, q'_I) = \alpha'_i(v'_i - p'_I)e^{-\beta'_i q'_I} = \alpha_i(\kappa_1 v_i + \kappa_2 - (\kappa_1 p_I + \kappa_2))e^{-\kappa_3 \beta_i q_I / \kappa_3} = \alpha_i \kappa_1 (v_i - p_I)e^{-\beta_i q_I} = \kappa_1 u_i(p_I, q_I)$. Since all the utilities of the APs are scaled by κ_1 in the system $(\mathcal{M}', \mathcal{N}')$, their choices of TPs do not change, and thus, the market share $\mathcal{N}_I(\mathcal{M}', \mathcal{N}')$ of the TPs do not change too. ∎

Proof of Theorem 3: We show part 1) by contradiction, and part 2) can be shown by the same arguments. Assume for some $\kappa > 1$, $I'_i = J$ with $q_J > q_{I_i}$. By assumption 2, we have $(v_i - p_{I_i})e^{-\beta_i q_{I_i}} \geq (v_i - p_J)e^{-\beta_i q_J}$ and $(v_i - p_{I_i})e^{-\beta_i \kappa q_{I_i}} \leq (v_i - p_J)e^{-\beta_i \kappa q_J}$. However, the above can be rewritten as

$$(v_i - p_{I_i})e^{-\beta_i q_{I_i}} e^{-\beta_i (\kappa - 1) q_{I_i}} \leq (v_i - p_J)e^{-\beta_i q_J} e^{-\beta_i (\kappa - 1) q_J}.$$

Because $q_J > q_{I_i}$ and $\kappa > 1$, we have $e^{-\beta_i(\kappa-1)q_{I_i}} > e^{-\beta_i(\kappa-1)q_J}$. Combined with the condition $(v_i - p_{I_i})e^{-\beta_i q_{I_i}} \geq (v_i - p_J)e^{-\beta_i q_J}$, we have a contradictory condition

$$(v_i - p_{I_i})e^{-\beta_i \kappa q_{I_i}} > (v_i - p_J)e^{-\beta_i \kappa q_J}.$$ ∎

Proof of Lemma 1: We first consider the case $q_I < q_J < q_K$. By Definition 1, we know that $p_J \leq \eta p_I + (1 - \eta)p_K$, where $\eta = (q_K - q_J)/(q_K - q_I)$. By Assumption 1, we have $u_i(p_J, q_J) \geq u_i(\eta p_I + (1 - \eta)p_K, q_J) = u_i(\eta p_I + (1 - \eta)p_K, \eta q_I + (1 - \eta)q_K)$. By Definition 2, we have $u_i(p_J, q_J) \geq u_i(\eta p_I + (1 - \eta)p_K, \eta q_I + (1 - \eta)q_K) \geq \min(u_i(p_I, q_I), u_i(p_K, q_K))$. Since $u_i(p_I, q_I) > u_i(p_J, q_J)$, we must have $u_i(p_J, q_J) \geq u_i(p_K, q_K)$. The derivation of the case $q_I > q_J > q_K$ is similar. ∎

Proof of Lemma 2: We show the sufficient condition for functions with two-dimensional domains:

$$2\frac{\partial u_i}{\partial p_I}\frac{\partial u_i}{\partial q_I}\frac{\partial^2 u_i}{\partial p_I \partial q_I} - (\frac{\partial u_i}{\partial p_I})^2 \frac{\partial^2 u_i}{\partial q_I^2} - (\frac{\partial u_i}{\partial q_I})^2 \frac{\partial^2 u_i}{\partial p_I^2} > 0. \quad (3)$$

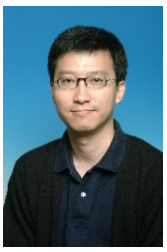
We know that $\partial u_i / \partial p_I = -\alpha_i e^{-\beta_i q_I}$, $\partial u_i / \partial q_I = -\alpha_i \beta_i (v_i - p_I)e^{-\beta_i q_I}$, $\partial^2 u_i / \partial p_I^2 = 0$, $\partial^2 u_i / \partial q_I^2 = \alpha_i \beta_i^2 (v_i - p_I)e^{-\beta_i q_I}$, and $\partial^2 u_i / \partial p_I \partial q_I = \alpha_i \beta_i e^{-\beta_i q_I}$. By substituting the above into (3), we get $\alpha_i^3 \beta_i^2 (v_i - p_I)e^{-3\beta_i q_I}$, which is positive. ∎

Proof of Theorem 4: We start with $p_I = p_I^{min}$ for all $I \in \mathcal{M}$, and for each overloaded TP, we increase its price until its capacity is fully utilized. For each step, when overloaded TP I's price is increased, APs will move to other TPs, making them possibly overloaded too. Thus, the prices of the TPs will be monotonically non-decreasing during the process, and therefore, will converge to a market equilibrium. ∎

Proof of Theorem 5: Since α_i is a linear scaling factor, it does not affect the choices of the APs and the market shares of the TPs. Thus, when α_is are scaled by κ, the aggregate throughput λ_Is are scaled by κ too. Because $Q_I(\cdot, \cdot)$s are homogeneous of degree 0, when ν_Is are scaled by κ as the same time, the achieved quality values do not change in the scaled system. By the monotonicity of Q_I of Assumption 4, we know that the equilibrium conditions of Definition 4 do not change, and therefore, $(\mathcal{M}', \mathcal{N}')$ is a market equilibrium too. ∎

Proof of Corollary 1: By Theorem 2, we know that the choices of APs and the market shares of the TPs will not change in both systems. By the same arguments for proving Theorem 5, we know that when α_is are scaled by κ and the effective capacity μ_Is are scaled by κ at the same time, the actual quality Q_is do not change. As a result, the equilibrium conditions of Definition 4 do not change. ∎

Richard T. B. Ma received the Ph.D. degree in Electrical Engineering in May 2010 from Columbia University. During his Ph.D. study, he worked as a research intern at IBM T.J. Watson Research Center in New York and Telefonica Research in Barcelona. He is currently a Research Scientist in Advanced Digital Science Center, University of Illinois and an Assistant Professor in School of Computing at National University of Singapore. His research interests include distributed systems and network economics.



John C.S. Lui received his Ph.D. in Computer Science from UCLA in 1992. He worked in the IBM San Jose/Almaden Lab and later joined CUHK. His research interests are in theoretic/applied topics in data networks, security, and optimization and performance evaluation. John received the CUHK Vice-Chancellor's Exemplary Teaching Award in 2001. He is a co-recipient of Best Paper Award in the IFIP WG 7.3 Performance 2005 and the IEEE/IFIP Network Operations and Management Conference.

Vishal Misra (S'98, M'99) received the B.Tech. degree from the Indian Institute of Technology, Bombay, India, in 1992, and the Ph.D. degree from the University of Massachusetts, Amherst, in 2000, all in Electrical Engineering. He is an Associate Professor in Computer Science at Columbia University. He received an NSF CAREER Award, a DoE CAREER Award and IBM Faculty Awards. His research emphasis is on modeling of computer systems, bridging the gap between practice and analysis.