Economics of Overlay Networks: An Industrial Organization Perspective on Network Economics

Peyman Faratin
Computer Science and AI Lab, MIT
peyman@mit.edu

Abstract

We use theory of Industrial Organization to demonstrate the demand and cost conditions on entry and scaling incentives of Internet overlay networks, restricting the problem to Internet *content* distribution. We describe the market structure, the nature of demand and the cost-allocation mechanism in both wholesale and retail markets. We show how the end-to-end coordination failures of the ISPs and content providers has resulted in wholesale market failures, inducing entry by propriety content distribution "middleboxes" that act as intermediaries, coordinating unmet demands. We also show that the intermediary has incentives to *strategically* use the Internet cost-allocation mechanism to internalize and scale using these indirect externalities from trade. The scaling incentives of such overlay markets is also briefly compared with bargaining institutions implemented by P2P content networks.

1 Introduction

This research grew from a question of whether overlay networks would "grow up", or whether they would remain "Peter Pan" technologies, confined to the lab and refusing to grow up. An overlay is roughly speaking a collection of "middleboxes" (a.k.a. "platform" or "intermediary" in economic nomenclature) that implement value-added functionalities at higher layers of the architecture (through various mechanisms including tunneling and DNS redirection). Overlays can have profound architectural, economic and policy effects on the Internet (see [CLA05] for a full treatment of these issues). Of most interest to engineers is the potential unraveling of End-to-End (E2E) principle, one of the core organizational principles of the Internet. Yet, despite this significant impact, we lack economic models to provide explanation of why overlays enter and whether they will scale, if at all. The contribution of this paper is to take a step towards this goal using models from Industrial Organizations to show what are some of the (demand and cost) conditions that induce entry and scaling incentives. We demonstrate the joint effects of these demand and cost forces on entry and scaling under two different architectures (or governance structures with different pattern of asset ownership), namely propriety and end-point P2P overlays, or wholesale and retail market architectures respectively. We have attempted to provide a brief glossary of most economic terms in this essay in Appendix 7.3 for readers unfamiliar with economic nomenclature.

We make two restrictions. Firstly, we focus on *incentives* to enter and scale and not on the absolute realization of scale of either architecture. Secondly, we restrict the focus to *content* overlays. In this economy, propriety and P2P content overlay serve the demands of different markets. In the former the economy typically consists of two distinct markets, content providers (such as Apple), who have a high willingness to pay for better than Internet's best-effort packet transport services, and content consumers. In the latter (P2P architecture) the economy consists of less distinct markets and generally not willing to pay for better than best-effort service. One of the key insights is that entry by propriety overlays is conditional on not only presence of coordination failures by incumbent ISPs but also this *asymmetric* nature of cross-market demands between content providers and consumers, and the higher this asymmetry the higher the incentive for entry. Additionally, recognizing and strategically managing these cross-market forces *ex-post* entry can also result in network effects that contribute to scaling (and potentially to "tipping") giving further *ex-ante* entry incentives. Information asymmetry has been a central topic of information economics since the early 1950s. However, what we want to show is that *demand* asymmetry can also affect the strategic behaviors, a result from Two-Sided Markets (TSM), an Industrial Organization (IO) model. Due to space constraints and the nature of the venue we refer the reader to [ROC05,ARM05] for a full treatment of the TSM theory. In this paper we focus instead on an informal and qualitative application of the theory to show how presence of demand asymmetries can lead to strategic effects when networks (including ISPs and propriety content overlays) implement two extant Internet wholesale transfer mechanisms, transit and peering contracts, to make cost allocations in the wholesale market. The technical and economic rationale (and evolution) of these transfer mechanisms, although fascinating, is also beyond the scope of this paper (see [HUS99a,b] for a brief overview). We therefore assume the cost-allocation mechanism is given, and instead focus on the strategic behavior the mechanism induces in presence of demand asymmetries Furthermore, we will also show that the Internet's wholesale cost allocation mechanism induces additional scaling incentives *ex-post* entry. Specifically, Internet wholesale transport market has considerable economies of scale, relative to the retail market, further driving scaling incentives of wholesale propriety operators relative to P2P content overlays.

In sum, we show that a propriety architecture has relatively a greater entry and scaling *incentives* because: i) ownership of assets allows the entrant to not only eliminate the transaction costs of coordinating resources, but ii) the firm can more efficiently capture the value from trade in distinct markets (propriety "intermediaries" are said to be better at internalizing the potential trade externalities in two-sided markets) and iii) manage costs better. (Bargaining institutions of) P2P on the other hand have limits to incentives for growth because of relatively greater: i) transaction costs (of search and contracting), requiring costly search and routing mechanisms (such as DHTs that introduce latencies because of their insensitivity to topological information), ii) *ex-ante* and *ex-post* information problems (trust) that lead to bargaining failures, requiring costly incentive mechanisms and iii) their inability to manage costs since costs in retail markets are allocated according to a flat-pricing mechanism which are inefficient and do not exhibit scale economies.

We also show how *entry* into propriety CDN is induced in the global content distribution economy by i) an E2E QoS coordination failure at the IP layer and ii) pairwise bargaining failures, between content providers and access providers, due to information asymmetries and significant transaction costs involved in searching for and contracting with parties (what is referred to as the "Coasian theory failure"). We focus predominantly on strategic *entry* incentives of propriety CDNs (Akamai being a canonical example) to show how such an institution solves some of the coordination failures with relatively larger scaling incentives. The intuition is that E2E coordination market failure and "Coasian theory failure" (due to large transaction costs) induce entry by a propriety platform who coordinates and internalizes the *indirect* externalities from potential transactions across distinct markets with demand asymmetries. It does so strategically using (often discriminatory fixed and/or usage) pricing structure that further contributes to scaling. The fixed price instrument (a.k.a. "high powered incentive scheme [LAF93]) is usually appropriate when transactions on the platform are unobservable requiring a "lump sum" transfer to create appropriate incentive scheme. Since packet delivery is observable we focus instead on the usage-based pricing instruments based on existing Internet (peering and transit) transfer mechanisms.

The higher level goal of this paper is methodological. We want to demonstrate through the use content overlays, a paradigmatic point from IO, that most network economics and engineering literature has to date assumed design autonomy and in fact have ignored the *market* determinates of *both* the underlying mechanisms as well as the strategic behaviors. The mechanisms and the strategic behavior are in fact the *result* of imperfect market structure (c.f. monopoly, duopoly,…) and basic environmental conditions (such as demand and supply), as well as (threat of) regulation. For instance, technology affects the productive efficiency of a firm, the basic environmental condition that affects supply. If an engineer designs a technology that reduces the average cost of production as output increases then the industry tends to have only one firm (the structure), the famous "natural" monopoly argument. If only one firm sells output in an industry then in absence of regulation it maybe able to *unilaterally* design and implement a (take-it-or-leave-it) mechanism to set a price that is well above the marginal costs of production (the strategic behavior). In a similar vein this paper demonstrates how the technological choices affect information, demand and cost allocation mechanisms and market failures, defining the basic conditions which in turn facilitate strategic entry (the structure), and growth using various pricing instruments (strategic behavior). The causal connection between "upstream" design choices to "downstream" mechanisms and strategies is not well understood in network economics literature. IO models can help engineers design "better" architectures that take into account these casual connections to incentives further "downstream".

2 Content Overlays

Content distribution became an engineering goal from the observation that although Internet infrastructure was distributed the web servers were not. Traditionally content delivery involved web servers serving many requests, a distribution mechanism that was unscalable (specially with flash crowds, e.g. Victoria Secret's show) and created latencies when servers could be a long way away from clients, where distance is measured as number of hops, delays, packet loss or geographic distance. Even if the server had enough resources to handle all requests the distance introduces delays. Furthermore, these performance problems gets worse as Broadband access gives incentive for more bandwidth intensive applications and more data to be sent along the routers and links. Consumers then face low service quality due to high delay, unstable throughput, and loss of packets in the best-effort model.

One solution to this problem is to change the service model of the Internet from best-effort to E2E QoS. But an E2E QoS solution has proven costly so far because defining a shared and interoperable E2E QoS ontology, accounting, and payment transfer (levels and structure) standards has failed. The alternative solution that emerged was evolution from traditional web model to intelligent dynamic content distribution. Historically this was achieved in two stages. First stage still assumed a centralized web server providing content and services but used load balancing server farms to overcome server side bottlenecks. The second stage involved relaxing the centralized assumption and distributing content closer to user using replication of content and web caching technologies. Content Distribution Networks (CDNs) evolved to solve this problem. A CDN is a communication network that distributes content by deploying infrastructure components operating at protocol layers 4-7. These components interconnect with each other, creating a virtual overlay network layered on top of an existing IP packet network infrastructure. CDNs that have emerged have included propriety systems (e.g. Akamai, Limelight, Savvis, etc) and P2P CDNs (CoralCDN, Bittorrent, Kazza, etc). The second stage lead to a model where static web objects is distributed at various locations, but services such as ecommerce and creation of dynamic content was still provided by a central server. Next logical step was to distribute content and *services*. Due to space constraints we will focus only on the caching of static content, first service class for most commercial CDNs.

3 Propriety CDNs

Propriety CDN are middleboxes that, although distributed in operation, are owned, controlled and managed by a single entity. Akamai is a canonical example of such a propriety CDN (which we will refer to as p-CDN to distinguish it from P2P CDNs). Akamai is estimated to have approximately 20,000 servers, collocated in 900 networks, in 70 countries and 750 cities, serving approximately 15% of the global Internet content (which at rough estimates is about 40% of traffic on the Internet).

3.1 Market Structure

Today's propriety CDN market consists of the following stakeholders: 1) **Content Providers** (CPs) such as Yahoo, MSN, CNN, Apple, etc., a market which is assumed to be a highly competitive. Furthermore, there are typically two types of CPs: i) delay insensitive but cost sensitive CPs who serve mainly static data (e.g. Microsoft patches update), and who trade-off performance for volume of completed downloads and ii) delay sensitive but more cost insensitive CPs who serve mainly dynamic, interactive content (e.g. CNN), 2) **Content Consumers** (CUs), or eyeballs, which are either "household" or corporate customers and run delay inelastic/elastic applications, have a budget constraint and buy access to the Internet through an Internet Access Provider, 3) **p-CDNs**, a differentiated and concentrated market (consisting of Akamai, Limelight, Savvis,…), 4**) Internet Access Provider** (IAP) which are most often facility-based duopolies (e.g. Verizon/Comcast) or resellers of unbundled transport services and 5) **Internet Backbone Provider** (IBP) which is considered a competitive market, specially after regulatory interventions [NUE05,GRE05]. IBPs are also referred to as Tier1s (e.g. AT&T, Sprint, Level3) who also provide access to Internet. We will refer to either IAP or IBP as IXP.

3.2 Services & Costs

Services in this market are as follows. CPs generate, market and sell content. CDN's primary service is to distribute this content. Therefore, (often large) CPs are one set of customers of CDNs. As we will show below the other potential customer are IAPs. Furthermore, most modern propriety CDNs are multiservice firms, providing not only traditional content but also "application acceleration" services (such as SureRoute, Akamai's routing overlay service) and other value-added features (such as DDoS prevention, location-based services, such as information about where users are, their language, etc). Significant efficiencies can be implemented from such economies of scope, giving further incentives for firms to scale. However, for expositional ease and comparison goals we will focus only on a single service – static (http) content distribution. IAP in turn provide retail packet transport services. Finally, IBPs are often facility-based providers of wholesale and retail packet transport services. IXPs can offer caching services, but mostly do not do so today.

The resources needed for production of these services, the "factor inputs", are often composed of some convex combination of labor and capital. The factor inputs into the p-CDN are: i) hardware (storage and redirection servers), ii) labor (for writing software for processing logs for billing, network operations and administration) and iii) distribution (collocation with IAP, peering or transit points). Hardware costs are often fixed and display economies of scale, where unit price of capital diminishes with increasing volume of purchase, giving strong incentives to scale the capital investment. Due to space limitations we focus only on the wholesale packet transport costs from perspective of CPs, IXPs and p-CDNs. The goal is to show the variable nature of these costs and how entities can strategically use the Internet transfer mechanisms (peering and transit) to allocate costs. Strategy space of packet transport entities as well as overlays is in fact much richer, playing a much more complicated ("poaching") game to create cost savings, but we will ignore this effect in this paper.

3.3 Usage (Interconnection) Costs

CPs, IXPs and CDNs need to interconnect to support E2E packet transport services. Interconnections in the Internet is coordinated through a pricing system, giving very strong incentives for entities to minimize the (fixed and usage) operating costs these prices incur. Fixed costs incur because each stakeholder must physically interconnect with other network/s to provide services, through either collocation (in case of CPs and p-CDNs) or physical links (provisioned through a third-party Interchange Exchange Points), or through private link/s. If interconnection is mutually stable then the fixed costs of a physical interconnection are then committed by investing in the physical links or collocations. Each entity then strategically uses the Internet routing algorithm (the Border Routing Protocol---BGP), *and* contracting strategies to control usage-based costs through how and what level (respectively) packets exit and enter their network.

A settlement mechanism that coordinates the E2E packet delivery is therefore needed that not only make transfers over *ex-ante* fixed interconnection costs associated with the physical investment, but also the *ex-post* usage based costs, and potentially other transaction and information costs (such as "moral hazard" costs in some peering cases, see [MIL00]). The solution in the Internet has been a spectrum of cost sharing structures for the fixed cost component and a system of bi-lateral transit and peering transfer contracts for the usage-sensitive transactions, two *standard* equilibrium *ex-post* settlement mechanisms that have satisfied the architectural constraints, implementing transfers based on the volume of traffic exchanged. Instead of a formal definition we provide an informal account of these mechanisms by example below.

In absence of a p-CDN CPs typically use transit, and less frequently peering, mechanisms to contract directly with IXPs. Figure 1 shows a simple interconnection example, involving two CPs (CP_j and CP_m), two sets of eyeballs (CU_i and CU_k) who are requesting the content, an IAP (labeled IAP1) and an IBP (labeled IBP2). Furthermore, the pairs (CU_i, CP_j) and (CU_k , CP_m) each contract for packet transport with an IAP and IBP respectively. The solid and dotted lines represent payment and traffic transfers respectively. Note the asymmetries in the direction of content and payment flow, a point which we will come back to below. The "links" in the figure are abstractions at least two levels. Firstly, the "links" are perhaps multiple physical (shared/private) links (in case of IXPs) or collocation servers hosted in datacenters where the CP collocates in the IXP. Secondly, the "links" can be abstracting away *n* number of (payments and traffic) exchanges between intermediary IXPs. We do this for parsimony purposes. Assume, unless otherwise stated, the links are direct and involve no other intermediary. P_{ij} represents payment from i to j.

The transfer mechanism in the consumer retail market (P_{i1} and P_{k2}) is usually a usage-insensitive flat-rate pricing, a transfer mechanism that end-hosts implement in P2PCDNs. There are two types of transfer mechanism implemented in the wholesale Internet market (see [HUS00]). *Transit* (P_{12}, P_{j1} and P_{m2}) is a non-linear usage-based pricing, commonly implemented as a "95-5" (less common is average). The customer (say, IAP1) of such a contract commits to and pays (to IBP2) for a level of total traffic it will send *or* receive to/from its provider (called Committed Information Rate, CIRs, measured in Mbps) *ex-ante* before the actual traffic load is realized. This price is discriminatory, decreasing non-linearly with increasing CIRs (see Appendix 7.4 for some data). The provider (IBP2 in this example) then computes the rank ordered 95th percentile of the inbound traffic and the outbound traffic and typically takes the maximum of the two values (less commonly takes the average or sum) as the traffic load for the month *ex-post*. If it is above the CIR the customer is priced (P_{12}) according to an *ex-ante* mutually agreed excess tariff rate. The other settlement mechanism (but not shown in figure 1), motivated by lower transaction costs, is called peering, where there is no payment transfers between interconnecting IXPs as long as there is symmetry to the traffic exchanged between the *customers* of IXPs, within some *ex-ante* agreed ratio levels (usually 4:1).

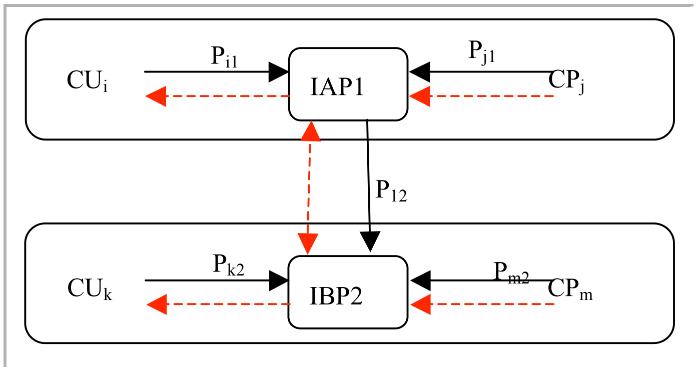


Figure 1: Classic Interconnection

In general, when a link represents multiple intermediaries then there are a series of such "vertical" transit contracts between customers and providers, with "horizontal" settlement-free contracts between peering intermediaries. Although intermediary IAPs may peer for optimization purposes (reducing transit costs due to P2P traffic), the IBPs (or Tier1 ASs) peer as their defining attribute, and do not pay transit with any other IBP. In total, the cost of end-to-end packet delivery is in fact born by end-hosts (and content providers) who share this cost through respective payments to their access providers, who in turn pay their upstream, until the peering pair (money, is said, to "move up the hierarchy" to the core).

Demand for better than best-effort transport by CPs is realized when "links" between CPs and CUs involve many intermediaries, decreasing quality (through increased latencies). As mentioned above, even though this demand can be priced on a single network (QoS can be guaranteed on a single network, c.f. AT&T v.s. Google), coordination failures between IXPs (in defining an E2E QoS) means the inter-provider transport market cannot price this E2E demand for better than best-effort service. The problem is also exacerbated because the settlement-free peering links lack QoS where settlements are independent of class of traffic (however, this is beginning to change as Tier1s begin to support QoS for VoIP).

These IXP coordination failures meant "rents were left on the table", rents that were available from a higher willingness to pay by CPs for the better than best-effort E2E service. However, the transaction costs of CPs contracting with all thousands of IAPs globally in the Internet for caching their content is prohibitively costly. Additionally, note that traffic is positively priced in transit contracts (e.g. $P_{12} > 0$). Therefore, if content that was previously collocated and served through many hops (CP_m to CU_i through multiple IXPs) for example) could instead be served closer to the destination (at say IAP_1) then there is not only a quality improvement (due to fewer number of hops), but an additional cost-saving for the IXP close to the destination (IAP1) since requests do not have to be served from "offnet", raising the transit costs of IAP. In fact lower latencies (due to either peering or caching) may increase demand because TCP sends more packets with lower latencies and users are more likely to view more web pages when the response time to load a page is quicker.

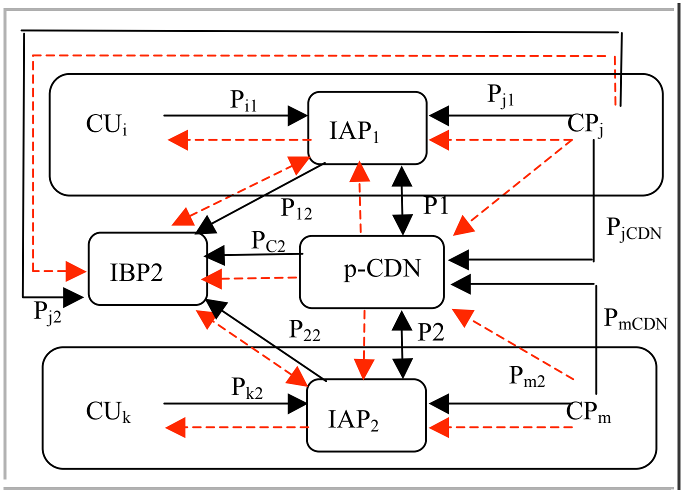


Figure 2: Interconnection with a p-CDN

The combination of rent price signal together with cost savings induced entry by p-CDNs, including Akamai, Cachelogic and Limelight who sold caching *services* as well as caching infrastructure vendors such as Inktomi. Figure 2 represents the system of interconnections if CPs' content was served by a p-CDN. Recall there are two types of CPs in this economy. Cost sensitive CPs (for example Microsoft, represented as CP_j) who are typically relatively lower margins for p-CDN (P_{jCDN}) because they seek to minimize costs for higher volumes. Microsoft could have range of possible optimizations choices, from letting Akamai serve all content (and paying P_{jCDN}), to distribute some percentage of total traffic on IBP (collocate within 4 or 5 IBPs in data centers in the world, and paying P_{j2}, with cost of a typical 95-5 transit contract) and the remaining portion be served by Akamai. The delay sensitive CPs (CP_m) on the other hand make *additional* marginal payments (P_{mCDN}) to get better performance (to end users) than current best-effort service of the Internet.

Furthermore, to a tier1 IBP a p-CDN is like any other network or CP in the system, but with a large (asymmetric) outbound traffic characteristics. A p-CDN "appears" large because it aggregates then "slices and dices" the traffic that best suits its traffic engineering needs (using the traffic as an instrument to meet peering and transit contract conditions). Therefore there is transfer of payment from a p-CDN to a IBP (P_{C2}) dictated by some transit contract, that varies in length from 1-3 years. Note, a p-CDN has strong incentive to grow because not only does traffic behave better (is less bursty and hence more amenable to better traffic engineering) the larger a p-CDN scales, but recall transit contracts induce incentives to scale because growth also lowers marginal transit costs which decrease non-linearly with increasing CIR. Finally, because a p-CDN reduces its variable usage costs (transit cost P_{12} or P_{22}) IAPs allow p-CDNs to collocate in their datacenters for free (P_1=P_2=0). In fact, there is some positive transfers to p-CDNs because often IAPs bear the fixed cost of collocation facilities such as security, power management and backup in a datacenter. Furthermore, there are no time horizon on these contracts and parties are free to cancel whenever wanted.

In sum, there is significant heterogeneity in pricing *structure* (and pricing magnitude, or *level*) in the Internet. All IBPs charge positive prices to *all* entities, CPs, CUs and p-CDNs ("double billing", because both the CU and CP is priced, an outcome that is implemented because of the bargaining power of Tier1s). The customers of modern p-CDNs are CPs who pay positive prices to p-CDNs. p-CDNs in turn make positive transfers to IBPs for IP underlay packet transport, but make no payment to smaller packet transport entities such as IAPs. p-CDNs bear both fixed costs (of capital investment and transaction costs of contracting on a global scale) and operational usage costs (of transit costs). So the interesting question is whether this is an equilibrium system of payments, and can it support cost-recovery? Although difficult to evaluate empirically the existence of the equilibrium was demonstrated when the Internet market experienced the *systematic* shock of the "bubble" collapse in 2001. An early entrant into the caching market, Inkotmi, sold caching *infrastructure* to IAPs, whereas Akamai sold caching services to CPs. Inktomi lost most of its customer base in the bubble and was acquired by Yahoo! in 2002. Akamai also lost many customers in the bubble but its recovery and continual growth is a signal that growth and cost recovery is a sustainable strategy in the complex system of transfers in the Internet only if the asymmetries in demand are properly accounted for. Two-Sided Markets (TSM) provides such a theory.

4 Demand Structure

The key insight from Two-Sided Markets (TSM) literature (which unifies multi-product with externality literature) is that a platform can enter and internalize the *indirect* externalities of trade across distinct markets (in our case, CUs and CPs), but must be sensitive to the level of demand asymmetry information when designing the pricing *structure* across markets. In other words, the basic conditions (demand in this problem instance), induces entry (the structure of the market) which in turn induces the behavior of the players (their strategies). The intuition in best parted in the example of a ("monopoly" *or* competitive) nightclub, a platform that charges men (labeled market *i*) differently to women (labeled market *j*), based on the rationale that below cost pricing to the women market will induce higher demand (q_j) by women which in turn will increase the valuation (and hence demand, q_i) of the platform by men as more women enter. The "cross-market elasticities" $e_{ji} = \partial q_i / \partial q_j$, measures the marginal change in consumption in market *i* with a marginal increase of consumption in the *j* market and represents the externality/spill-over effect market *j* consumption has on market *i*. The platform internalizes these externalities through a discriminatory pricing rule in order to recover fixed and usage costs. Recognizing and correctly price discriminating (and subsidizing) these forces can lead to nonlinear externalities/network effects that assist the growth of the platform. The exact nature and magnitude of such cross-subsidies is generally dependent on not only the sensitivity of demand in each market to prices ("native" price elasticity of demand) but also: i) the degree of cross-market elasticities, ii) extent of multi-homing and iii) the degree of membership and usage externalities. A rudimentary overview of application of TSM model is given in Appendix 7.1 (but see [FAR06] for a more in-depth exposition of an application to networking problems as well as [EVA03,ROC05,ARM05] for a detailed overview and treatment of the literature). Interesting IBPs are insensitive to this information (practices "double billing") and Inktomi misinterpreted it. As we will show next P2PCDNs in fact reintroduce symmetry, making $e_{ji} = e_{ij} = 0$.

5 p-CDNs v.s. P2P-CDNs

The interested reader is referred to Appendix 7.2 for a brief overview of P2P CDN overlays and some of their key economic properties. There are a number of very interesting points of comparison between p-CDNs and P2P-CDNs, but today propriety and P2P CDNs serve distinct quality markets. However, the P2Ps technical architecture has immediate consequences on its scaling incentives. A P2PCDN replaces the client-server model of content delivery with a model where each node is a client *and* a server simultaneously (in fact the original Internet was a P2P model, see Appendix 7.2). Additionally, P2Ps often have reciprocal incentive mechanisms (such as Tit-for-Tat or some reputation mechanism). These features effectively (re)introduce demand symmetry into the system, because peers are given incentives to balance demand with supply of content, thereby altering the two-sided nature of markets in a client-server model. In fact P2P economy is better compared to a commodity market, where market prices determine whether a node is a buyer or a seller; two-sided markets on the other hand involve "matching" *distinct* markets (see [ROT92], p.1). To best of our knowledge there are no pricing mechanisms for a P2PCDN, so it is not clear how the potential benefits from growth (quicker access to larger content) are moderated by increasing costs of coordination with scale. Existence of "marquee" peers (ones who are "heavy hitters") may support the design of discriminatory pricing institution that cross-subsidizes the usage of others. However, such a mechanism is difficult to implement and scale because peers are end-hosts that access the Internet through a DSL/Cable link, paying either a fixed tariffs or peak-tiered pricing to their IAP (who must ultimately bear the cost of implementing the transfer mechanism). Therefore either it is the IAP (in the case of flat-rate tariff) or the peer (in the case of peak-rate tariff) who will instead bear the marginal cost of these cross-subsidies, lowering their incentives. In fact symmetry introduced by P2Ps today is giving IAPs greater incentive to *peer* with other IAPs who they exchange P2P traffic with so as to reduce these transit costs due to P2P traffic, which today is estimated to account for 60% of Internet traffic. A third party platform that coordinates transactions does not have an entry incentive either because symmetry reduces its ability to design a discriminatory institution so as to match demands and internalize the *indirect* externalities.

Altogether these incentive problems increase the pressures on bargaining institutions, increasing likelihood of *ex-ante* and *ex-post* bargaining failures. Peers not only have to find costly way to search for others (DHTs for instance do not reflect topological proximity thereby increasing packet Round Trip Time and latencies and effecting quality; CoralCDN attempts to solve this problem through more intelligent IP clustering [COR04]), but must also trust each other and be given appropriate participation and cost management incentives. In a two-sided economy on the other hand a platform has incentive to enter, and scaling is not only easier but also provides better provisioning against bargaining failures (an insurance scheme in effect, thereby increasing trust). Neither does a p-CDN need to design costly incentive for cost and coordination management mechanisms needed to deliver a service because it has ownership rights and control over the capital whose costs actually decrease with scale (economies of scale), giving the platform further economic incentives to scale.

6 Conclusions and Future Work

Our goal was to demonstrate how basic economic conditions such as demand asymmetries and cost structures together with market failures facilitated strategic entry and growth of propriety middleboxes using various pricing instruments. We also tried to qualitatively show how P2P CDNs architectural choices alter the basic conditions that affect entry and relative growth in the content market.

Our higher level goal was to use CDN overlays to demonstrate a paradigmatic point that mechanisms and the strategic behavior are in fact the *result* of imperfect market structure and basic environmental conditions (such as demand and supply), as well as (threat of) regulation. There are many other Internet problems (e.g. network neutrality, ossification of IP architecture, failure of QoS and multicast) that can be analyzed using IO models. The benefit of doing so is that understanding the causal connection between "upstream" design choices to operator incentive and performance problems, can help engineers design "better" architectures that take into account these economic forces further "downstream".

Our future goal is to demonstrate two points. Firstly, we have not accounted for the effect of overlay entry and growth on the current Internet architecture (see [GRE05] for a cursory treatment). For example, what are the competitive and architectural affects of scaling of Akamai who is able to offer better SLAs than incumbent ISPs' best-effort service? How will p-CDN scaling affect topology of the Internet? What, if any, are the structural and behavioral determinates and consequences of externalities that can lead market "tipping" in favor of one overlay?

Finally, Internet itself was an overlay over the Public Switch Telephone Network (PSTN), just as overlays are a response to innovation failures at the IP layers. Many of the emerging Internet problems (BGP churn, QoS standardization, Multicast cost sharing, risk exposure by access providers in the retail market, etc), not just E2E content transport, are due to the costly coordination problems across networks, lowering their incentives to innovate (referred to as histerisis in econometric models). Could overlays be a stable pattern of innovation? What are the conditions for this to be true? We would like to use IO tools to evaluate whether coordination failures by the IP "underlay" can lead to an *institutional* response that can correct that failure.

Acknowledgements

This work was supported by funding from MIT's Communications Futures Program (cfp.mit.edu). I would also like to thanks David Clark, Arthur W. Berger, Steven Bauer, William Lehr, Patrick Gilmore and Rahul Sami for their continual support and informative discussions.

Bibliography

[ARM05] M. Armstrong (2005): *Competition in Two-Sided Markets*, forthcoming, Rand Journal of Economics, forthcoming

[CLA05] David Clark, William Lehr, P. Faratin, S. Bauer and J. Wroclawski (2005): *The Growth of Internet Overlay Networks: Implications for Architecture, Industry Structure and Policy* In the proceedings of Telecommunications Policy Research Conference (TPRC-05), Washington, DC. 2005.

[COR04] M. J. Freedman, E. Freudenthal, and D. Mazières, *Democratizing Content Publication with Coral*, In Proc. 1st USENIX/ACM Symposium on Networked Systems Design and Implementation (NSDI '04) San Francisco, CA, March 2004.

[EVA03] D. Evans (2003) The Antitrust Economics of Two-Sided Markets. Yale Journal on Regulation: 2003, 20(2), pp: 325—381

[FAR06] P. Faratin, T. Wilkening (2006): *Interconnection Discrimination: A Two-Sided Markets Perspective*. In Proceedings of Fifth Hot Topics in Networks (HotNets-V '06), Irvine, CA, US, November 29-30, 2006

[GRE05] The Economic Geography of Internet Infrastructure in the United States, in Handbook of Telecommunications Economics, Technology Evolution and the Internet, Vol.2, S.K. Majumdar, I Vogelsang and M. Cave (eds), Elsevier, 289—364, 2005.

[HUS99a,b] J. Huston (1999a,b): *Interconnection, Peering and Settlements: Part I & II*, CISCO Internet Protocol Journal, (2), 1 & 2, 1999, pp. 2-24.

[LAF93] J.J. Laffont and J. Tirole (1993): *A Theory of Incentives in Procurment and Regulation*, MIT Press, Cambridge, MA, US,1993.

[MIL00] P. Milgrom, B. Mitchell and P. Srinagesh *Competitive Effects of Internet Peering Policies*, in The Internet Upheaval, edited by Ingo Vogelsang and Benjamin Compaine, Cambridge: MIT Press (2000), 175-195.

[NUE05] H. E. Nuechterlein and P.J. Weiser (2005) *Digital Crossroads: American Telecommunications Policy in the Internet Age*, MIT Press, Cambridge, MA, US, 2005

[PAR05] G.G.Parker and M. Van Alstyne (2005): *Two-Sided Network Effects: A Theory of Information Product Design*, Management Science, (51), 10, 2005, pp. 1494-1504.

[ROC05] J. Rochet and J. Tirole (2005): *Two sided Markets*: A Progress Report, forthcoming, Rand Journal of Economics.

[ROT92] A. Roth and M. Sotomayor (1992): *Two-Sided Matching: A Study in Game Theoretic Modeing and Analysis*, Econometric Society Monographs, Cambridge University Press.

[TIR88] J. Tirole (1988): *Industrial Organization*, MIT Press, Cambridge, MA, US

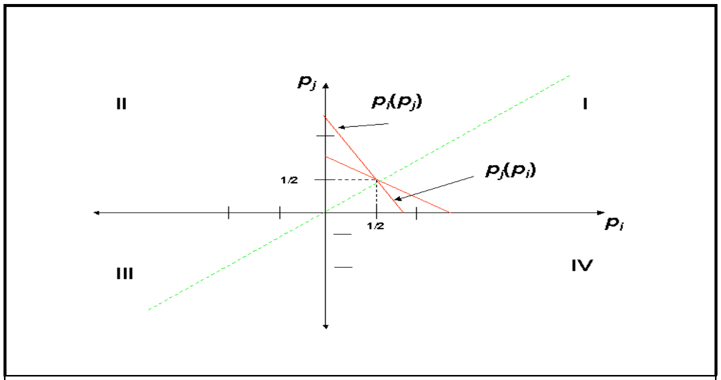
7 Appendix

7.1 A Model of TSM

Figure A1 shows the geometry of the pricing problem of a monopolist IXP who intermediates two markets *i* and *j*. Let the *i* and *j* markets be the CU and CP markets facing usage prices p_i and p_j from the IXP respectively. The problem of the IXP is to determine the structure of profit maximizing prices. Let q_k denote the total consumption of network transport services in market $k \in \{i,j\}$. One potential additive demand function is $q_i = D_i(P_i) + e_{ji}D_j(P_j)$, where $D_i(P_i)$ is the "native" demand in the *i* market at price P_i, and the additive term is the effect of the consumption in the other market (see [PAR05] for nonlinear demand). The constant $e_{ji} = \partial q_i / \partial q_j$, measures the marginal change in consumption in market *i* with a marginal increase of consumption in the *j* market and represents the externality/spill-over effect market *j* consumption has on market *i* demand. Figure X, shows the *benchmark* pricing structure case where there is no cross-market effects, $e_{ji} = e_{ij} = 0$. The solid lines represent the pricing reaction curves $p_i(p_j)$ and $p_j(p_i)$, representing the optimal prices in the *i* market given prices in the *j* market and, conversely, the optimal prices in the *j* market given prices in the *i* market respectively. Specifically, $p_i(p_j)$ is computed as the solution to the following maximization problem with p_j fixed:

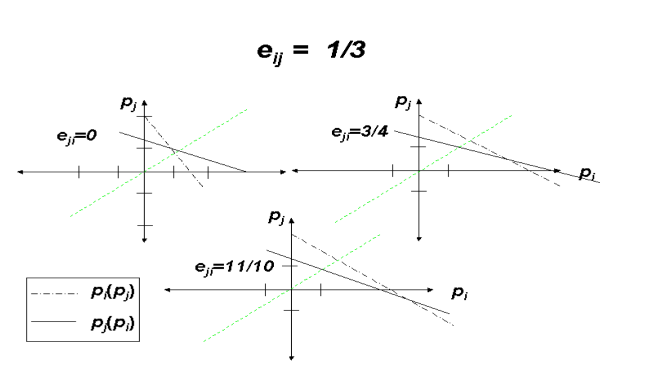
$$p_i^*(p_j) = \arg\max_{p_i} (p_i q_i) + (p_j q_j).$$

When there is no cross-market effects, $e_{ji} = e_{ij} = 0$, the equilibrium set of prices in each market lies on the 45° line, corresponding to classic monopoly prices. That is, when markets are independent then both markets are priced positively (the level of which is captured by the Lerner index and regulated by the degree of elasticity of native demand in each market, [TIR88]). IBPs charge positive prices for both content requestors and servers ("double billing", quadrant I). However, IAPs also charge considerably less (even below cost) to content providers whose content is much in demand.



8 **Figure A1: Geometry of the Pricing Problem**

Figure A2 shows how the pricing structure can diverge from the benchmark independent markets when the cross-market effect from market *i* (content users/eyeballs) is constant, and cross-market effect from market *j* to market *i* increases [PAR05]. That is, as e_{ji} increases (the demand of content users increases as demand for transport by content provider increases) then prices to the content providers decrease, to the point (e_{ji} = 11/10) that the content providers may in fact be subsidized by the platform (because users value content).



9 **A2: Pricing with Indirect Externalities**

9.1 P2P CDN Overlays

A P2P CDN (PP-CDN) achieves the same functionality but unlike the p-CDN is managed and controlled by no single ownership. It is often forgotten that the original Internet was designed as a P2P; applications such as SMTP, FTP, UCCP and DNS, are all peer applications where each end host can act as either client or server in the protocol. Part of End-to-End principle requires transparency where packets flow unaltered through the network from source to destination. Packets can then be retrieved unaltered from the source by knowing only their address. Transparency enables (but does not require) symmetry since any end point can access any other end point. However, overtime Internet became asymmetric. FTP, Telnet, HTTP are more client-server protocols that allow servers to take on a different role than clients. Note, it is the role and not the protocol the node is running that defines whether it is a P2P. As the client-server model became more of the installed base the access architecture changed, also becoming more asymmetric: servers have high bandwidth connections to Internet, clients have ADSL, providing more down than uplink bandwidths toward client There are other reasons for increasing asymmetry in the Internet. IP addresses are becoming more transient (PPP,SLIP,DHCP). Firewalls allow connection to outside but not inside. There are increasingly more private addresses. Network Address Translations (NATs) are middleboxes that change IP addresses dynamically. Application level gateways, relays, proxies and caches may alter content in ways that are unknown or uncontrollable by the endpoints. Voluntary isolation and peer networks such as WAP protocol networks do not use Internet addressing and protocols but connect end points to Internet. Split DNS. Load sharing schemes hide real endpoint's IP address behind a VPN. Modern P2Ps therefore reintroduce symmetry. In early days of Internet peers were a small number of large organizations such as academic and military. Today peers communicating are often large number of individuals, with approximate aggregate level of traffic of 65%.

However, P2Ps make a number of tradeoffs. Firstly, scalability is achieved at a large cost of organization and coordination. This tradeoff often has a phase transition where organization can become the scarce resource with scale. Secondly, scale is achieved at cost of quality. P2Ps achieve scalable, reliable, and fault resilient operation from a collection of unreliable peers with intermittent connections. Therefore constructing a Service Level Agreements, SLAs, can be prohibitively costly since servers join and leave dynamically.

The major technological cost of P2P Networks is content location. Solutions have varied from centralized directory (e.g. Napster, Archie, WAIS and Gopher), to flooding request (e.g. Gnutella), to DHT routing solutions (e.g. Chord) which themselves are not efficient since they ignore geographical/topological information. P2Ps also face social costs of trust, accountability and reputation. Content integrity is primary trust issue in PP-CDNs, because original content can be tampered with or altered during storage, transport and/or delivery, misleading the requestor and compromising the reputation of author. Both author and requestor must therefore ensure the content requestor gets exact copy of content using technologies such as digital signatures. Accountability is needed to solve the free-rider problems that lead to the tragedy of commons.

9.2 Glossary of Economic Terms

Coase's Theorem: In the absence of transaction costs, all (government) allocations of property rights are equally efficient, because interested parties will bargain privately to correct any externality. Obstacles to bargaining are often sufficient to prevent this efficient outcome, leaving normative Coase theorem to prevail over positive Coase theorem.

Economies of Scale: A reduction in long run unit costs which arise from an increase in production. Economies of scale occur when larger firms are able to lower their unit costs. This may happen for a variety of reasons. A larger firm may be able to buy in bulk, it may be able to organise production more efficiently, it may be able to raise capital cheaper and more efficiently. All of these represent economies of scale.

Economies of Scope:

Exogenous: A term which describes anything pre-determined or given in a piece of analysis. Not determined by the model

Endogenous: A term whose value is determined within the economic model.

Externalities: The spillover effects of production or consumption for which no payment is made. Externalities can be positive or negative. For example all fax users gain as new users become connected (positive); and smoke from factory chimneys (negative). Variously known as external effects, external economies and diseconomies, network effects, spillover, and neighborhood effects. It can be direct or indirect.

Ex-ante: The planned, desired or intended level of some activity

Ex-post: The realized level of some activity

Internalization: Act of capturing

Market Failure: Term used to describe a situation in which markets do not efficiently allocate goods and services

Settlement mechanism: A settlement mechanism is any rule or institution that specifies a system of cost and benefit transfers between N parties in negotiation.

Transfer mechanism: Synonym to Settlement mechanism

9.3 Transit Prices

The results below were gathered by B. Norton at he 36th Peering Bird of a Feather at North Amercian Network Operator's Group (NANOG):

Sample size: 42

Number of Tier1s: 28

Average Cost: $25/Mbps

Maximum Cost: $95/Mbps

Minimum Cost: $10/Mbps

Average Commit Level: 1440 Mbps

See www.nanog.org/mtg-0606/pdf/bill.norton.2.pdf for the nonlinear pricing schedules