THESIS IN ECONOMICS

Final Honour School of Philosophy, Politics and Economics 2016

Congestion Based Financial Instruments for the Internet Economy

Abstract

In this thesis, we answer the question: could market mechanisms be employed to tackle or relieve Internet congestion? In particular, we look at three market mechanisms: consumer-side congestion pricing, content-side paid prioritization, and bilateral risk sharing agreements. While the former two have been heavily studied in economic literature, the risk sharing approach towards congestion is a novel addition of this thesis. We treat the uncertainty in broadband congestion levels as an economic risk that consumers and Internet businesses are forced to bear. The thesis investigates the possibility of introducing congestion-based financial instruments, similar to derivatives in a stock market, that efficiently allocate risk borne out of congestion.

Chapter 1: Introduction

The Internet faces a growing congestion problem. The changing demands of today's population means that billions of people are increasingly using the same broadband pipelines for transferring much heavier files.[1] The core Internet usage has rapidly shifted from text and image to high definition videos and voice telephony. With the emergence of various online multimedia services, the bandwidth demand from an average user has massively increased, both in volume and in size. Broadband infrastructure, however, hasn't kept pace with the exploding demand, leading to frequent congestion slowdowns (Klinker, 2013).[2] The situation is expected to worsen in the coming years, with potentially devastating consequences for all Internet-based businesses, consumers and service providers (ISPs). [3][4]

Various approaches have been suggested to deal with general congestion in economic literature (in Chapter 2, we show how network congestion is an externality problem and inherits some of the same issues). One way to deal with congestion externalities is to establish norms that penalize inappropriate behavior, like free-riding, that leads to congestion. Such norms can work well in small groups where there is repeated interaction; however, they do not scale well to a system with millions of users. Another way to deal with congestion is to establish rationing or quota systems that reject additional users when the load is too high (Bohn et al., 1994). Despite the simplicity of of these approaches, economists tend to favour pricing mechanisms as a way of alleviate congestion.

Historically, Internet service providers (ISPs) have adopted *flat-rate pricing* for simplicity, which required consumers to pay a flat monthly pee for Internet access independent of usage volume. Mobile providers, on the other hand, have generally adopted *usage-based pricing*, where the price paid is proportional to the volume used. While usage-based pricing can promote judicious use of resources and improve quality of service, it doesn't eco-

[1] As of the most recent reported period (2015), the number of internet users worldwide was 3.17 billion (Stats, 2016)

[2] For example, in 2012, the subscribers to the Sky network in 34 areas of the UK faced a congestion slowdown - lasting as long as seven weeks. The network had been placed under increasing pressure to meet the growing demand through high volumes of new customers and greater levels of demand from existing customers. As one network expert observed, "Use of services like Netflix and Sky Anytime+ has accelerated the bandwidth demand curve, meaning that previously planned upgrades did not happen soon enough"(Stewart, 2012)

[3] Congestion-based financial disputes between businesses and ISPs have been going on for some time now. In January 2014, Netflix - the on-demand video streaming business, publically blamed the Comcast and Verizon service providers for congestion, and published a report showing that the performance for its streaming video service was declining on these networks. In response, the ISPs demanded that Netflix and its transit operators pay new fees for the ever-growing amounts of traffic. Netflix eventually agreed to pay Verizon a shakedown fee to avoid interconnection congestion for its streaming, but the blame game for congestion still continues (Lyons, 2016; Masnick, 2014; Solsman, 2014)

[4] Netflix alone accounts for more than a third of peak traffic in the United States. It has also been responsible for considerably slowing down broadband speeds in Australia, a phenomenon now known as 'The Netflix effect'. With expansion into 130 new countries in January 2016, many of which have network infrastructure far worse than Australia or the US (in reliability, quality and reach), questions of allocating cost of congestion becomes even more important(Bingemann, 2015; Metz, 2016)

nomically inefficient unless the prices are linked to the costs of congestion. Linking pricing to congestion requires *congestion-sensitive pricing*. The driving idea behind congestion-sensitive pricing is to convert delay and queuing costs into real costs for the users and thus internalize the negative congestion externality each user imposes on other users. Two such schemes suggested in the literature are 'Smart Market' (MacKie-Mason et al., 1993; MacKie-Mason and Varian, 1995) and 'Dynamic Capacity Contracting' (Singh et al., 2000). We review the economics behind congestion-sensitive pricing in Chapter 3, observing that such pricing models, while economically efficient in theory, come with serious implementation challenges, technical inefficiencies, and overhead costs.

Instead of consumer schemes that link broadband tariffs to congestion, Internet Service Providers (ISPs) have instead pressed for new business models where content providers (CPs) pay consumer networks for access to users. ISPs argue that since the bandwidth-intensive services of CPs bring about congestion in the first place, CPs ought to pitch in towards the massive network investments needed for mitigating it.[5] The argument has some economic merit; one reason why broadband infrastructure is underprovisioned is that network investments create positive externalities for CPs which are not factored in the private cost-benefit analysis of ISPs. Revenue from the content providers would (arguably) allow ISPs to recoup their investments in existing networks and incentivize further investment and innovation in content delivery.

Several issues have been raised against this proposal. ISPs, for one, have been accused of using congestion as an excuse to profiteer from both sides of the market (by not delivering on their promise to the customers). But there is another problem with allowing ISPs to cut deals with CPs - that it violates *net neutrality*, the principle that all packets or content must be treated equally by service providers. Policymakers are concerned that allowing differential pricing for content by ISPs would potentially open a Pandora's box of economic evils, leaving ISPs free to adopt economically inefficient business models and network management practices due to a lack of sufficient competition in the provision of broadband access services (Becker et al., 2010; Choi et al., 2015; Jang et al., 2010). We revisit the economic rationale behind net neutrality in Chapter 4.

It is clear that all stakeholders in the Internet economy - consumers, CPs, and ISPs, stand to be adversely affected by congestion. Unaccounted variation in congestion levels across time can lead to loss of utility for consumers, loss of profit for CPs, and loss of market share and revenues for ISPs. What is needed for a viable solution to the congestion problem is

[5] Interviews by CEOs of major ISPs in Europe illustrate the nature of the arguments involved:

"Service providers are flooding networks with no incentive to cut costs...It's necessary to put in place a system of payments by service providers as a function of their use." - France Telecom CEO Stephane Richard.

"[Companies such as Google and Yahoo! Inc.] use Telefonica's networks for free, which is good news for them and a tragedy for us...That can't continue." - Telefonica SA CEO Cesar Alierta

The mismatch between investments and revenue "is set to compromise the economic sustainability of the current business model for telecom companies," - Telecom Italia CEO Franco Bernabe a link between CPs and ISPs that realigns the interests of those who invest with those who capture value, without violating net neutrality. In this thesis, we propose one such solution: a contract between CPs and ISPs which transfers congestion risk from the CPs back to the ISPs. Such a contract would help any party affected by congestion decrease their exposure to congestion. Just like weather derivatives protect businesses from the vagaries of weather (Stix, 1998), or electricity derivatives protect utilities and generators from fluctuations in electricity spot prices (Deng and Oren, 2006), a class of instruments for risks from Internet congestion, would provide relief to all the stakeholders in the Internet. A market in risks, even if it doesn't solve or relieve congestion, promises to soften its adverse economic impacts.

This thesis is structured as follows:

- Chapter 2 explains how network congestion, being an externality problem, leads to market failure. In particular, we use a congestion game from Nisan et al. (2007) to demonstrate how the problem of Internet congestion is a special case of the *tragedy of the commons*.

- In Chapter 3 and 4 we review two proposed solutions for congestion: retail pricing models that re-couple congestion cost to revenue (consumer side congestion-sensitive pricing, Chapter 3), and new business models that allow CPs to subsidize ISP investment (Chapter 4). We conclude that under net neutrality regulations, these solutions are inadequate. In particular, consumer-side congestion-sensitive pricing is costly to implement, while charging CPs through paid prioritization carries potentially harmful anti-competitive effects that threaten content innovation and overall social welfare.

- In Chapter 5 we outline the economics of risk-sharing contracts that provide an alternative channel of raising money from CPs while maintaining net neutrality, providing ISPs with both funds (through upfront payments by content providers who buy the contract) and incentives to invest in Internet infrastructure. We investigate the correct pricing of a congestion based security in a market for congestion risk. Lastly, we investigate reasons (information asymmetry problems and non-diversifiability of congestion risk) that might lead to failure of our proposed market mechanisms.

Chapter 2: Congestion as Externality

Consider the problem of providing network bandwidth which is shared by many users. There are places and periods when bandwidth is scarce and periods when it is abundant. When the supply of bandwidth far exceeds the demand, there is little role for economics. But during the periods when demand exceeds the supply, bandwidth is a scarce resource, and the fundamental issues of resource allocation become important.

Under suitable assumptions, the first theorem of welfare economics guarantees that resource allocation is efficient (in a competitive equilibrium). One such assumption requires the welfare of each consumer to be dependent solely on their consumption decision. However, this is not the case when users share a common network bandwidth under congestion: each user's welfare is directly affected by the action of other users. If Alice sends a data packet that crowds out Bob's packet; Bob suffers delay, but Alice is unaffected for the cost imposed on Bob.[6] Thus, network congestion is an externality which results in market inefficiencies.

If congestion is too high, it is possible that the consumer receives no utility at all from the broadband access. This is a special case of the classic *tragedy of the commons*, with the broadband pipeline as the common resource, i.e. a resource which is both *rivalrous* and *non-exclusive*.[7] As long as users have access to unlimited usage, without instituting new mechanisms for congestion control, the network is going to suffer from "overgrazing".

To see this, consider a congestion game (Nisan et al., 2007). Suppose that $n > 1$ players are sharing a broadband pipeline (or bandwidth) of maximum capacity 1. Each player chooses to send x_i units of flow (or demands x_i bandwidth) along the channel for some value $x_i \in [0,1]$. Thus, each player has an infinite set of strategies.

Assume each player would like to have the largest fraction of the bandwidth, but the quality of the channel deteriorates with the total bandwidth used. For simplicity, assume that if the total bandwidth demanded $\sum_j x_j$ exceeds the pipeline capacity, then no one gets any benefit. If $\sum_j x_j < 1$, then the payoff for player i is $x_i(1-\sum_j x_j)$. This payoff function captures the trade-off: the benefit for a player deteriorates as the total bandwidth demand increases, but it increases with his own share (upto a point):

$$P(x_i) = \begin{cases} 0 & \text{if } \sum x_j \geq 1 \\ x_i(1-\sum_j x_j) & \text{if } \sum x_j \leq 1 \end{cases}$$

[6] The way the Internet deals with congestion is to either drop packets (so that some information must be resent by the CPs), or to delay traffic.

[7] When villagers have shared, unlimited access to a common grazing field, each will graze his cows without recognizing the costs imposed on the others. Without some mechanism to control contain congestion, the commons will be overgrazed.

To find the stable strategies at equilibria for each player, consider player i. Assume $t = \sum_{j \neq i} x_j$ is the bandwidth demanded by all other players. Now player i faces a simple optimization problem for selecting his optimal strategy:

$$\max_{x_i} x_i(1 - t - x_i)$$
$$x_i \in [0, 1]$$

Differentiating with respecting to x_i, we find that the optimal flow for player i, x_i^*, is given by:

$$x_i^* = \frac{1-t}{2} = \frac{1 - \sum_{j \neq i} x_j}{2}$$

Double differentiating the objective function $x_i(1-t-x_i)$ with respect to x_i, we get $-2x_i \leq 0$, hence x_i^* is indeed a maxima.

At Nash equilibria, *each* player is playing his optimal selfish strategy, given the strategies of all other players. Thus, for at Nash equilibria, $x_i^* = \frac{1-\sum_{j \neq i} x_j}{2}$ for all i. This is a set of n equations in n variables, and has a unique solution,[8] given by:

$$x_i^* = \frac{1}{n+1} \forall i$$

For player i, the benefit $P(x_i)$ is $\frac{1}{n+1}(1 - \sum_j \frac{1}{n+1}) = \frac{1}{(n+1)^2}$. And the aggregate net benefit (to all the players) $\sum_i P(x_i)$ is $n * \frac{1}{(n+1)^2} = \frac{n}{(n+1)^2}$.

While the Nash equilibria solution maximizes the individual payoff $P(x_i)$, the efficient solution maximizes the aggregate payoff $\sum_i P(x_i) = \sum_i x_i(1 - \sum_j x_j) = \sum_j x_j(1 - \sum_j x_j)$. Both $\sum_j x_j$ and $(1 - \sum_j x_j)$ are positive when $\sum_j x_j < 1$, so we can use the Arithmetic Mean - Geometric Mean (AM-GM) inequality[9] to show that at any solution which maximizes the aggregate payoff, $\sum_j x_j = (1 - \sum_j x_j)$, from which we get $\sum_j x_j = 1/2$ and $\sum_i P(x_i) = 1/4$.

The following table summarises the individual and aggregate payoffs:

[8] This can also be arrived at through a simple symmetry argument. Since all players are identical agents, their optimization problems and their strategies at equilibrium must be identical. Substituting x^* for all demand variables in $x_i^* = \frac{1-\sum_{j \neq i} x_j}{2}$ gives $x^* = \frac{1}{n+1}$

[9] For two positive terms a and b, $\frac{a+b}{2} \geq \sqrt{ab}$. In other words, the maximum value the geometric mean $\sqrt{ab}$ can attain is equal to the arithmetic mean $\frac{a+b}{2}$. This happens when the terms are equal:

$$\sqrt{ab} = \frac{a+b}{2}$$
$$\implies 4ab = (a+b)^2$$
$$\implies (a-b)^2 = 0$$
$$\implies a = b$$

Thus, the product of two positive terms is maximized when they are equal.

$$
\begin{bmatrix}
 & \text{Individual Payoff} & \text{Aggregate Payoff} \\
\text{Nash equilibria} & \frac{1}{(n+1)^2} & \frac{n}{(n+1)^2} \\
\text{Socially Efficient Solution} & \frac{1}{2n} & \frac{1}{4}
\end{bmatrix}
$$

Why is the congestion game a tragedy? The net aggregate benefit at equilibrium is too low. $n/(n+1)^2 \approx 1/n$, so the total welfare at equilibria is $n/4$ times worse than the optimal solution, and decreases rapidly as the total number of players increases. As the number of players reaches infinity, the aggregate welfare at Nash equilibria vanishes.

Note that our analysis does not assume which side of the market the players in the congestion game are. The players can be thought of as CPs sending data packets to consumers, or as consumers demanding bandwidth to access content. Congestion affects both sides of the markets. Both the sender and receiver of a data packet are adversely affected when it is delayed or dropped.

If the network congestion is too high, content providers can lose revenue or market share as their consumers face slow or no access. For example, if the primary revenue source is advertising (Google Adwords, for instance), then heavy congestion might lead to a massive reduction in the click-through rate on the advertisement. Different content or businesses would, however, differ in their sensitivity to congestion. Data providers such as e-mail or Wikipedia can be relatively insensitive to moderate delays. By contrast, the revenue from streaming video or Voice over Internet Protocol (VoIP) services can depend critically on congestion. The adverse impact of congestion might also be more indirect. If Netflix is not able to assure its customers of high quality streaming on a consistent basis, then it might lose subscribers in the long run.

Chapter 3: Consumer Side Congestion Pricing

The congestion game in the previous chapter assumes that the players face no fees for sending (or receiving)[10] an additional unit of data packet. This represents the current pricing models under which most ISPs operate. On one side of the market, the users pay a flat monthly fee to the ISPs for accessing the Internet. On the other side, the CPs face no fees. Flat rate pricing essentially allows high-volume users to impose costs on low-volume users. In effect, the low volume users subsidize the high volume users. Usage-based pricing, which charges users a tariff proportional to usage, solves this problem to some extent. But even usage-based pricing is inefficient if individual players contributing to congestion do not bear the true costs of congestion.

When players are charged rates that are not sensitive to congestion, the private cost of sending an additional data packet is not aligned with the true costs. Utility-maximizing players will increase their consumption of bandwidth until the marginal utility from any further increases in usage is zero, at which point the social costs associated with the last unit consumed will exceed the benefits, and welfare is reduced. Economic welfare is maximized if the market reaches equilibrium at the point where the social benefits equal the social costs. In the case of bandwidth, this would occur where the benefits each player derives from the last packet equals the costs of congestion created by the packet.

Congestion sensitive pricing aligns incentives by bringing private costs into line with the true social costs of consuming an additional unit. The price of bandwidth usage (sending a data packet) in an uncongested network should be close to zero; a higher price is socially inefficient since it does not reflect the true incremental costs of using another unit of bandwidth. The price for sending a data packet when the network is congested should be positive: if my packet eats into another user's bandwidth or delays their packet, then I should pay the cost I impose on the other user. If my data packet is more valuable than hers, then it should be sent; if hers is more valuable than mine, then hers should be sent. Charging network users for the congestion they cause can lead to more efficient network utilization by forcing them to take social costs into account.

The demand and supply curves in Figures 3.1 and 3.2 sketch out the logic of this argument.[11] Suppose the price of sending a data packet were very high: only a few users would want to send packets. As the packet price decreases, more users would be willing to send packets.

Case 1 (Figure 3.1): If the bandwidth capacity is fixed at K, then the optimal price for admitting the packets is where the demand curve crosses the capacity supply. If demand is

[10] Note again that the congestion game can be used to analyze both sides of the market. Players can be modelled as a group of CPs or as a group of consumers.

[11] We are indebted to MacKie-Mason et al. (1993) and MacKie-Mason and Varian (1995) for the analysis in this chapter

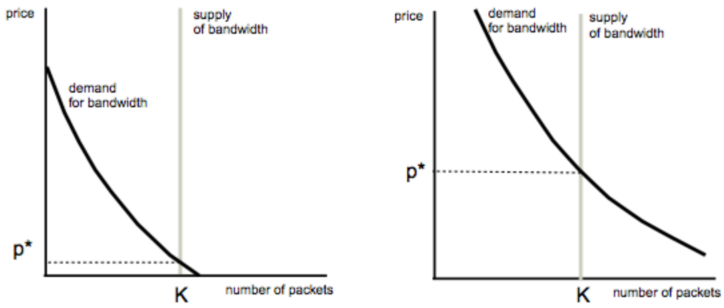

Figure 3.1: Demand for network access with fixed capacity K. When demand is low, the price for sending a packet is low. When demand is high, the packet price is high. Source: MacKie-Mason et al. (1993)

small relative to capacity, the efficient price is zero - all packets are admitted. If demand is high, only the users that are willing to pay at least the price of admission to the network are allowed to send their packets.

Case 2 (Figure 3.2): If an increase in packets from some users imposes delay on other users, but not outright exclusion, then the analysis is slightly different. If we know the costs of congestion (how delay varies with the number of packets), and that we have some idea of the costs imposed on users by a given amount of delay, then we can calculate a relationship between number of packets sent and delay costs. The cost of congestion or the marginal cost of delay, the cost added by the next single packet, determines the optimal number of users.

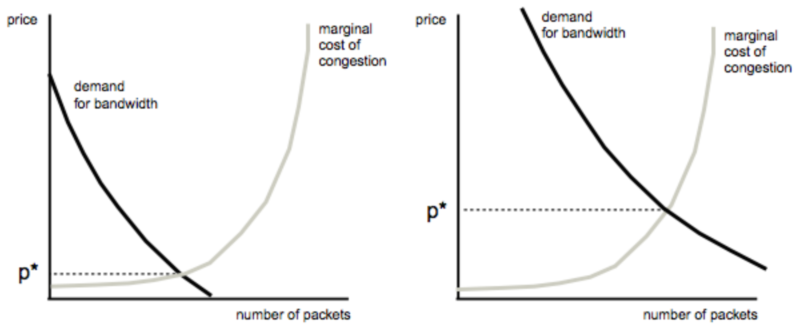

Figure 3.2: Demand for network access with marginal cost of delay. When demand is low, the price for sending a packet is low. When demand is high (congestion is high), the packet price is high. Source: MacKie-Mason et al. (1993)

The efficient price is where the user's willingness to pay for an additional packet equals the marginal increase in delay costs generated by that packet. If a potential user faces this price, she can be able to compare her own benefit from sending a packet to the marginal delay costs she imposes on other users.

To see this mathematically, let x_i denote user i's use of bandwidth and $t = \sum_{j=1}^{n} x_j$ the total use of the bandwidth. The user cares about her own use x_i, and the delay that they encounter. Instead of the utility function $u_i(x_i, t) = x_i(1-t)$ used in the congestion game, let us use a more general utility function $u_i(x_i, \eta)$ where $\eta = t/K$ is a measure of bandwidth utilisation or congestion (K is the overall capacity. Summarize the preference of the user by $u_i(x_i, \eta) + m_i$, where m_i is the money user has to spend on other things. Assume that $u_i(x_i, \eta)$ is a differentiable, concave funciton of x_i and a decreasing concave function of η. Delay can be interpreted as a general congestion cost: it can include the cost of exclusion, congestion, etc. This generic formulation also captures the relationship between usage and capacity: if total usage t is double and capacity K is also doubled, then there is no change in congestion or utilization $\eta = t/K$ and the delay remains constant.

Let the efficient pattern of usage be given by:

$$W(K) = \max_{x_j} \sum_{j=1}^{n} u_j(x_j, \eta) - c(K) \tag{3.1}$$

From (3.1) and the observation that $\frac{\partial u_j(x_j,\eta)}{\partial x_i} = \frac{\partial u_j(x_j,\eta)}{\partial \eta}\frac{\partial \eta}{\partial x_i} = \frac{1}{K}\frac{\partial u_j(x_j,\eta)}{\partial \eta}$, we get the first-order condition:

$$\frac{\partial u_i(x_i,\eta)}{\partial x_i} = \frac{-1}{K}\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta} \tag{3.2}$$

that the optimal solution must satisfy.

This says that user i should use the network until the marginal benefit from their usage equals the marginal cost that they impose on the other users using the same pipeline. Define a "shadow price" p_e which measures the marginal cost of congestion that an increase in i imposes on other users (this is independent of i):

$$p_e = \frac{-1}{K}\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta}$$

If the user is charged the shadow price p_e for usage, they now solve the following utility maximization problem:

$$\max_{x_i} u_i(x_i,\eta) - p_e x_i$$

The solution to which is given by the first order condition:

$$\frac{-1}{K}\frac{\partial u_i(x_i,\eta)}{\partial \eta} + \frac{\partial u_i(x_i,\eta)}{\partial x_i} = p_e \text{ or} \tag{3.3}$$

$$\frac{-1}{K}\frac{\partial u_i(x_i,\eta)}{\partial \eta} + \frac{\partial u_i(x_i,\eta)}{\partial x_i} = \frac{-1}{K}\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta} \tag{3.4}$$

For large n, the term $\frac{-1}{K}\frac{\partial u_i(x_i,\eta)}{\partial \eta}$ will be negligible compare to the term $\frac{-1}{K}\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta}$, and hence (3.4) is essentially the same as the first-order condition in (3.2).

For example, if consider the case when $u_i(x_i,\eta) = v_i(x_i) - D(\eta)$. Then the social optimum in (3.2) is described by:

$$u_i'(x_i) = \frac{n}{K}D'(\eta)$$

and the individual optimization in (**??**) is described by:

$$u_i'(x_i) = \frac{n+1}{K}D'(\eta)$$

For large n, both are virtually the same. Hence the social optimum coincides with the equilibria if a congestive-sensitive shadow pricing is implemented. The price p_e, "internalizes" the externality by making the user face the costs that she imposes on the other users. Note that in this model each user faces the same price for usage: the sum of the marginal congestion costs that each user imposes on the other users.

Let us reconsider (3.1). $W(K)$ is the maximum welfare given arbitrary capacity K. Differentiating with respect to K:

$$W'(K) = -\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta}\frac{t}{K^2} - x'(K) \tag{3.5}$$

(recall that $t = \sum_{j=1}^{n} x_j = \eta * K$). Using shadow price $p_e = \frac{-1}{K}\sum_{j=1}^{n}\frac{\partial u_j(x_j,\eta)}{\partial \eta}$, we can write this as :

$$W'(K) = p_e\frac{t}{K} - c'(K) \tag{3.6}$$

$W'(K) > 0$ iff $p_e * \sum_{j=1}^{n} x_j > Kc'(K)$. This means that expanding capacity increase welfare if and only if the revenue from the congestion fees $p_e * \sum_{j=1}^{n} x_j$ exceeds the value of capacity $Kc'(K)$, where capacity is valued using the marginal cost of capacity. If the bandwidth usage never fills up capacity (even at no cost for packets), then there is no need to expand capacity.

Congestion-sensitive pricing provides guidance about when to expand capacity. Consider the case with fixed capacity: Packet prices measure the marginal value of the last admitted packet. If the cost of expanding capacity to accommodate one more packet is less than the marginal value of that packet, then it makes economic sense to expand capacity. If expansion costs more, it is not economically worthwhile. As MacKie-Mason and Varian (1995) observes:

> Optimal congestion pricing plays two roles - it efficiently rations access to the network in times of congestion, and it sends the correct signals about capacity expansion. In this framework, all the revenues generated by congestion prices should be used to expand capacity

Why, then, is congestion-sensitive pricing not observed in practice?

A first concern about congestion sensitive pricing is that poor users will be deprived of access if the prices during congestion are too high. However, this is not a problem with pricing itself, but with the distribution of wealth. That certain users have sufficient resources to purchase a base level of services by could be ensured by redistributing resources through vouchers or lump sum grants. In an efficient network, the total costs will be lower, so it will be less costly to meet distributional objectives than the economically inefficient flat-price network.

The main problem with congestion-sensitive pricing is the assumption, implicit in our analysis, that exclusion and metering is costless. These costs would be astronomical if ISPs were required to keep detailed accounts for every packet sent because packets are very small units. The costs of accounting and billing such as back-end calculation of charges, applying and documenting pricing schemes, dealing with overcharges, providing customer service for dealing with bills - all contribute to the overhead. The existence of these costs make congestion-sensitive pricing inferior to flat-rate pricing in practice. Moreover, users may not like tariffs which are inherently unpredictable and charged a posteriori. The difficulty in communicating to consumers the state of the network and bandwidth in network would lead to "bill shocks" (Li, 2015; Helms, 2015; Conto, 2015). These factors render congestion-sensitive pricing inadequate as a solution to congestion.

Chapter 4: Two Sided Pricing - Charging Content Providers

Would allowing ISPs to charge CPs for delivering their traffic relieve congestion?

One reason why the current network capacities are frequently congested is the misalignment between who funds investment in Internet infrastructure and who ultimately captures value. Under *net neutrality* regulations, implemented in many countries, ISPs are not allowed to differentiate between CPs and charge them. Neither do the CPs have any means to express preferences and improve their welfare.[12] The positive externalities of expanding the infrastructure enjoyed by the CPs are not accounted for in the private cost-benefit decisions of the ISPs. This results under-provisioning of network capacity.[13]

The primary model through which ISPs propose to generate revenue from the content side of the market is through paid prioritization of content. Paid prioritization is a form of service differentiation in which ISPs offer CPs tiered services in exchange for fees. This can be done, for example, if the ISP divides its capacity into a premium and ordinary class, and CPs get charged for carrying traffic in the premium class. The added revenue from such services could be used to pay for the building of increased broadband access to more consumers.

However, allowing ISPs to set up price barriers for content threatens content innovation. Priority lanes may make it hard for nascent CPs to compete against established ones. Content-side access fees and price discrimination would thus deter entry, reduce CP surplus and CPs' innovation incentives, especially affecting nascent CPs (Becker et al., 2010).

While nascent CPs and content-side innovation would be threatened, might paid prioritization improve overall welfare through lower consumer prices and higher investments? ISPs investments are driven by the trade-off between softening consumer price competition and increasing revenues from CPs. Specifically, investments are higher in the non-neutral regime because it is easier to extract revenue through appropriate CP pricing. On the other hand, participation of CPs may be reduced in a non-neutral network due to higher prices. The net impact of paid-prioritization on social welfare is determined by which of these two effects is dominant.

[12] In the US, for example, the Federal Communications Commission (FCC) has implemented net neutrality rules that would "prohibit a broadband ISP from discriminating against, or in favour of, any content, application or service." ISPs would be prohibited from: (1) prioritizing traffic and charging differential prices based on the priority status; (2) adopting business models that offer exclusive content or that establish exclusive content based relationships with particular content providers; and (3) charging content providers to access the Internet based on factors other than the bandwidth supplied.

[13] Revenues from online services are growing more than twice as fast as those from Internet access (Page et al., 2010)

A second concern is that the benefits of prioritization to those paying for the preferential treatment are only realized during times of congestion. This creates perverse investment incentives because broadband providers will, in turn, benefit from congestion in the network. Paid-prioritization will deter investment by broadband providers that can benefit from this artificial scarcity. Saavedra (2009), for example, finds that although unregulated regime leads to higher quality investments, ISPs have an incentive to degrade content quality.

The economic literature is curiously divided or non-commital about the issue of desirability of net neutrality regulations. Most papers have approached the problem by using the framework of non-cooperative game theory to find the output of the interactions between selfish actors that are end users, CPs and ISPs. Ma and Misra (2013) have shown that the effect of non-neutrality on consumer welfare depends on ISP competition. In a monopolistic scenario, network neutral regulations benefit consumer but in an oligopolistic situation, the presence of competing price-discriminating ISPs is more desirable (from a purely consumer welfare point of view). Njoroge et al. (2013) finds that when CP-quality heterogeneity is large, the non-neutral network is always welfare superior in a ”walled-gardens” model, while the neutral network is superior in a ”priority lanes” model. Musacchio et al. (2009), on the other hand, argues that although the returns on investment under one-sided or two-sided pricing are comparable, the size of investment and profit depend on the advertising rates and users’ price sensitivity. From an overall social welfare point of view, Njoroge et al. (2013) also concludes that two-sided pricing results in higher social welfare; however, Lee and Wu argued that the zero pricing at the CP-side could be optimal in theory (Lee and Wu, 2009; Armstrong, 2006; Rochet and Tirole, 2003). In addition, Pil Choi and Kim (2010) concludes that the short-run welfare is higher under one-sided neutral regulation.

All things considered, the concerns about anti-competitive effects seem to just about settle the argument against allowing content side discrimination by the ISPs. The basic principle of open, non-discriminatory interconnection has been an important principle of two-way communications networks. Non-discriminatory interconnection is why the Internet evolved very differently from cable television, why it grew into an infrastructure that facilitates the truest expression of the free-market.

Chapter 5: A Market in Congestion Risk

Service prioritization and two-sided pricing are not allowed under the net neutrality principles. While the consequent cross subsidization to CPs has nurtured content innovations at the edge of the Internet, it reduces the investment incentives for the access ISPs to expand capacity. A viable solution to the congestion crisis would not just help ISP raise funds for upgrading their networks, but would also require linking funding to the positive externalities so generated. The problem with the business and pricing models that ISPs operate today is that CPs and users bear all the congestion risk, while ISPs bear the investment risk.

Risk is costly to bear (in utility terms). If we can defray risk through market mechanisms, we can potentially make many people better off without making anyone worse off. This is why risk sharing agreements such as insurance exist: efficient risk markets can unequivocally improve social welfare.

Lessons learned from the financial markets suggest that financial instruments, when well understood and properly utilized, are beneficial to the sharing and controlling of undesired risks through properly structured hedging strategies. The uncertainties arising due to unforeseen congestion in the network can be priced, divided into marketable `chunks` and sold to someone who is willing to bear that risk - in exchange for a fee or a future stream of payments. An insurance based on congestion would mitigate the effects of congestion volatility for Internet businesses and Consumers as well as assure investment cost recovery for ISPs. Just like weather derivatives protect businesses from the vagaries of weather (Stix, 1998)[14] or electricity derivatives protect utilities and generators from fluctuations in electricity spot prices (Deng and Oren, 2006), a class of instruments for risks from Internet congestion, ‘Congestion derivatives’, would provide relief to all the stakeholders in the Internet.

A market in congestion based risks would provide a novel solution to the problem of raising money from content providers while maintaining net neutrality, providing ISPs with both funds (through upfront payments by content providers who buy the contract) and incentives to invest in Internet infrastructure.

This chapter is structured as follows:

- Section 5.1 introduces a simple model of the Internet from Ma and Misra (2013). The model introduces the basic concepts of throughput, demand and congestion level that will be used in the description of congestion-risk sharing contracts

[14]An investor can buy or sell a contract whose value depended entirely on fluctuations in temperature or accumulations of rain, hail or snow. The weather derivatives are a means for an insurer to help provide for future claims by policyholders or a farmer to protect against crop losses. Or the contracts might allow a heating oil supplier to cope with a cash shortfall from a warmer than expected winter by purchasing a heating degree - day floor - a contract that would compensate the company if the temperature failed to fall below 65 degrees as often as expected.

- Section 5.2 describes a 'congestion option,' a risk-sharing contract that gives the owner a right to receive payment, subject to congestion outcomes within a specified period of time.
- Section 5.3 explains how congestion-based risk sharing contracts can be priced. In particular, we look at two alternative pricing methodologies: *rational expectations* and *no-arbitrage*, inspired from insurance pricing and option pricing respectively. We observe that in the absence of an underlying complete market with tradeable assets, the no-arbitrage approach would not be applicable for the pricing of congestion risk contracts. However, insurance-like pricing through rational expectations requires the stochastic process of congestion to be known.
- Finding a stochastic process that governs congestion is an open area of research. Section 5.4 sketches out an approach.
- Finally, section 5.5 looks at possible causes of market failure for congestion risk markets.

5.1 Three-Party Ecosystem Model

This section introduces a simple model of the Internet from Ma and Misra (2013) that will be useful for devising a risk sharing contract of the sort we envisage.

Consider a model of the Internet with three parties: CPs, ISPs, and consumers.[15] [16] Assume that the consumer group is fixed in a targeted geographic region. Let N and M denote the number of CPs and consumers respectively. Each consumer subscribes to an Internet access service via a single ISP.[17] The ISP acts as an intermediary between the consumers and the content providers, providing a fixed bandwidth broadband pipeline which the consumers collectively access. Denote K as the last-mile bottleneck capacity towards the consumers in the region. Given a set N of CPs, a group of M consumers, and a link with capacity K, denote the system as a triple (M, K, N). Denote by λ_i the aggregate throughput

[15]For clarification, companies which provide content, through websites or online services, like Netflix, Youtube, Wikipedia and the like, are CPs. Ordinary users of the Internet, who browses we pages, download files, stream multimedia content, are Consumers. Comcast, AT &T, Verizon or British Telecom (BT) are examples of ISPs.

[16]The distinction between content providers and consumers might seem nebulous. For example, in a skype conversation or email exchange, who is consumer and who is the content provider? However, these are instances of consumer to consumer traffic, which can be ignored for our purposes. Email exchanges between consumers take up trivial volumes. With the development of VoIP, file sharing, and online gaming services, the absolute volume of consumer-to-consumer traffic is increasing, but still expected to constitute a relative decrease of the percentage of total traffic due to explosive growth in Internet video streaming and downloads. The traffic between CPs and the traffic between consumers take up small volumes relative to the volume of traffic from CPs to consumers (see Laffont and Tirole (2001)). Also See Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2012-2017.

[17]The model does not include backbone ISPs because the bottleneck of the Internet is often at the last-mile connection towards the consumers (Courcoubetis et al., 2003)

(the total traffic) rate from CP i to the consumers.

Figure 5.3 depicts the contention at the bottleneck among different flows from the CPs.

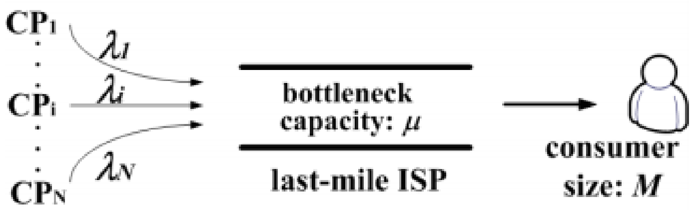

Figure 5.3: Contention at the last-mile bottleneck link. Source: Ma and Misra (2013)

Let $\hat{\theta}_i$ denote the unconstrained throughput demand for a typical user of CP i. For example. the unconstrained throughput for the highest quality Netflix streaming movie is about 5 MB/s, and given an average query page of 20kB and an average query response time of 0.25 s, the unconstrained throughput for a Google search is about 600 kb/s, or just over 1/10 of Netflix.

Define $\hat{\lambda}_i = M\hat{\theta}_i$ as the unconstrained throughput of CP i.

Without contention, CP i's throughput λ_i equals $\hat{\lambda}_i$. However, when the capacity K cannot support the unconstrained throughput from all CPs, i.e., $K < \sum_{i \in N} \hat{\lambda}_i$, a typical user of CP i will obtain a throughput $\theta_i < \hat{\theta}_i$ from CP i and some active users might stop using CP i when θ_i goes below the threshold, e.g., users of streaming content like Netflix. Define a demand function $D_i(\theta_i(\eta))$ that represents the percentage of consumers that still demand content from CP i under the achievable throughput θ_i. The demand is a non-negative, continuous and non-decreasing function of throughout θ defined on the domain of $0, \hat{\theta}$. Since θ is a function of congestion levels η, the demand is in turn a function of congestion η. In particular, $D_i(\eta = 0) = 1$. For example, if demand decays exponentially with congestion, then we can model demand as $D(\theta(\eta) = e^{-\alpha_i \eta}$ where α_i measures the CP i demand's sensitivity to throughput.

Under minimal assumptions about the rate allocation mechanism under congestion,(Ma and Misra, 2013) have shown that any given system (M, K, N) has a unique level of congestion (let us denote it by η) and throughput rates at equilibrium. In other words, for given a set of CPs (characterised by given unconstrained throughput rates and demand as a function of achieved throughput), and a per-user capacity, there is a unique set of achievable equilibrium throughput rates (determined by the rate allocation mechanism). If the congestion level is denoted by η, given the system, there is a unique level of throughput rates $\boldsymbol{\theta}(\boldsymbol{\eta}) = \{\theta_i(\eta) : i \in [1, N]\}$. For example, if CP i's throughput rates decay exponential with increase in congestion η, then throughput can be described by $\theta = \hat{\theta} e^{-\beta_i \eta}$ for some β_i.

5.2 A Congestion Option

In this section, we devise a 'congestion option' in direct analogy to the stock options in a financial market.

An option is a contract which gives the buyer (the owner or holder) the right, but not the obligation, to buy or sell an underlying asset or instrument at a specified strike price on or before a specified date, depending on the form of the option. There are two types of options: calls and puts. The buyer of a call, for example, pays the seller a premium upfront at the beginning of the contract. In return, he gets the right to buy the underlying at an agreed upon 'Strike Price' at the predetermined date. (Practically speaking, if the strike price is more expensive then the actual asset value at the predetermined date, then the buyer wouldn't exercise the option).

On similar lines, a 'Congestion Option' would give the owner a right to receive payment, subject to certain congestion conditions, within a specified period of time. A generic congestion option can be formulated by specifying the following parameters: the contract type (call or put), the contract time (some particular instant in the future or a full period), the underlying index (congestion index in this case), strike level, and payouts.

A congestion call option would give the owner, the right to receive q units of currency (call it 'tick size') for each basis point (0.0001) of congestion exceeding a minimum threshold amount $\tilde{\eta}$ at time T. We can write the payout of the contract as:

$$\tau(\eta(T)) = \begin{cases} 0 & \text{if } \eta(T) \leq \tilde{\eta} \\ q(\eta(T) - \tilde{\eta}) & \text{if } \eta(T) > \tilde{\eta} \end{cases}$$

For illustration of how such a contract works, assume a CP (say Youtube) is concerned that its customer base in a particular geography is vulnerable to congestion. Further assume that its revenue depends on the real time throughput it is able to deliver. This is the case for the revenues through video advertisements. The throughput (hence revenue) that Youtube is able to achieve at any point of time depends on the congestion level at that time. There is risk borne from the uncertainty in how much revenue it will make during those hours.

A contract which promises a compensation if that geography experiences heavy congestion in return for fixed upfront premium would essentially allow Youtube to hedge the risk of losing its revenue (in that geography). On the other side of the market, the upfront premium could be used to raise funds that are pumped into improving the network infrastructure by an ISP such as Comcast, creating a win-win situation for all. [18][19]

Consider CP i, whose total throughput under congestion η is given by $\lambda_i = MD(\theta_i)\theta_i$. Let the per unit traffic profit for CP i be p_i. Thus, CP i's revenue is given by $R(\eta) = p_i MD(\theta_i)\theta_i$, where $\theta_i = \eta(M,K,N)_i$. Let $D(\theta) = e^{-\alpha_i\eta}$ where α_i measures the demand's sensitivity to throughput and $\theta = \hat{\theta}e^{-\beta_i\eta}$. Thus, $R(\eta) = p_i Me^{-(\alpha_i+\beta_i)\eta}$.

With the option bought at premium P and base income y_0, the CP's congestion contingent total income would be:

$$R(\eta) = \begin{cases} y_0 + p_i M\hat{\theta}e^{-(\alpha_i+\beta_i)\eta} - P & \text{if } \eta \leq \tilde{\eta} \\ y_0 + p_i M\hat{\theta}e^{-(\alpha_i+\beta_i)\eta} + q(\eta - \tilde{\eta}) - P & \text{if } \eta > \tilde{\eta} \end{cases}$$

5.3 Pricing of Congestion Options

How should a congestion option be priced?

The congestion option can be conceived as an insurance-like bilateral risk sharing agreement. However, insurance is also a special case of an option. In finance, a call option allows one to buy the right, but not the obligation, to buy a stock at a predetermined price at a future date. The call option acts as an insurance against the scenario where the stock becomes unaffordable and overpriced. Similarly, a put option insures a stock holder against the scenario where the stock loses its value. Thus, it would be useful to look at techniques from both insurance pricing as well as option pricing to find the market value of such a contract.

In the next sections, we apply two pricing approaches that is common in finance literature: *rational expectations* and *no-arbitrage*. In the rational expectations approach, used in insurance pricing, we conceive of congestion option as a risk trade between two parties with different risk preferences. This gives upper and lower bounds on the price of the option. This pricing approach is preferred if the contract is not tied to other assets in the

[18] Compare this hypothetical Youtube-Comcast transaction to the deal which launched the successful 'weather derivatives' movement: A contract between Aquila Energy and Consolidated Edison, which involved ConEd's purchase of electric power from Aquila for the month of August 1996, and a clause which stipulated that Aquila would pay ConEd a rebate if August turned out to be cooler than expected.(Hamisultane et al., 2008; Nicholls, 2004)

[19] As an alternative example: A video-on-demand provider whose primary business model is pay-per-view may decide that it needs to protect itself from potential loss of revenue due to heavy congestion caused by live streaming of weekend football matches. It might buy a congestion call option for the weekends. If the congestion during the weekend is indeed beyond the minimum agreed threshold amount, the video-on-demand provider makes additional revenue which offsets any potential losses. If not, then the premium paid upfront would be made up for by the revenue enjoyed from no congestion. The other side of the market could include a speculator (who gets the desired amount of risk-return in his portfolio), a service provider (which recovers the cost of its investment in infrastructure in case of no congestion), or even a football streaming website (which makes up for its loss of high viewership).

market whose price is known. If this is not the case, incorrect pricing might lead to arbitrage opportunities. If the market is such that there are other assets in the market which are tied to the option, then the *no-arbitrage* principle of option pricing is preferable. As Jensen and Nielsen (1995) observes:

> Theories and models dealing with price formation in finanical markets are divided into (at least) two markedly different types. One type of models attempting to explain levels of asset prices, risk premium etc. in an absolute manner in terms of the so called fundamentals. A crucial model of this type includes the well known rational expectation model equating stock prices to the discounted value of expected future dividends. Another type of models has a more modest scope, namely to explain in a relative manner some asset prices in terms of other, given and observable prices.

5.3.1 Congestion Option As A Bilateral Risk Sharing Agreement

This section uses the rational expectations model, inspired from the insurance market, to answer the following questions:

1. Under what conditions would such a congestion option enhance welfare?
2. What is the fair valuation of such a contract?
3. What would be the equilibrium price for the contract in a competitive market?

For the rest of the discussion we assume that the risk-sharing contract is between a CP affected by congestion and an insurer (which can be an ISP or an insurance company or other such entity). Further assume that the traffic from CP i is not big enough to influence congestion level η, or in other words, CP i is a 'congestion-taker'.[20]

Buying the option decreases the overall for the content provider when congestion is low and increases it when congestion is high. If the CP is risk averse , then such an option would be attractive. There are several reasons why a CP might be risk averse.[21] Nondiversified ownership, liquidity limitations or costly financial distress all push the CP towards risk aversion. Even if maximizing profits is the preferred objective of the CP, entrusting control to a risk-averse manager whose wages are linked to performance may induce the CP to behave risk-aversely as revenues are subject to uncertainty. Empirically, it has been observed that firms not only pursue profit frrom risky business but also avoid risk by hedge and insurance. Oil producers hedge oil prices(Jin and Jorion, 2006) and corporations buy currency derivatives (Géczy et al., 1997). So we assume that the CP's utility v is a concave function of its profits, that is $v'(R) > 0$ and $v''(R) < 0$)

[20]This assumption is analogous to the assumption that firms in a perfectly competitive economy experience a common price level; every firm is a 'price taker': no individual firm can influence the price on its own. Relaxing this assumption leads to moral hazard problems discussed later

[21]Firm behaviour under uncertainty is one of the unsettled areas of economics (see Choudhary and Levine (2009), for instance)

Similarly, the insurer's or ISPs utility function u, may be a rather complicated matter. If the ISP is risk-seeking or risk neutral, then it can definitely benefit from taking on some of a risk averse CP's profit risk. But it is worth noting that even if the insurer is risk neutral or risk averse, there are mutually beneficial risk transfer if the cost of risk to the CP is more than the cost of the risk to ISP.

Consider a simple congestion option, designed for binary outcomes where η takes just two values : uncongested 0 and congested 1. Further assume that there's no throughput under the state of congestion $\hat{\theta}e^{-(\alpha_i+\beta_i)} \approx 0$, and $\hat{\eta} = 0$. Let τ denote the payoff of the contract:

$$\tau(\eta) = \begin{cases} 0 & \text{if } \eta = 0 \\ q & \text{if } \eta = 1 \end{cases}$$

where q is the payout in the congestion event.

For a content provider with fixed income y_0 and congestion dependent profits, the revenue $R(\eta)$ with the contract is:

$$R(\eta) = \begin{cases} y_0 + p_i M\hat{\theta} - P(q) & \text{if } \eta = 0 \\ y_0 + q - P(q) & \text{if } \eta = 1 \end{cases}$$

where $P(q)$ is the price of the contract.

This contract can be seen as an insurance with loss function $L = p_i M\hat{\theta}$, cover q and the probability of loss $\pi_1 = P(\eta = 1)$ (with $\pi_0 = P(\eta = 0) = 1 - \pi_1$). The premium $P = P(q)$ is a function of cover provided.

Let $V = E[v]$, so $V(\pi_1, q, P(q)) = \pi_1 v(y_0 + p_i M\hat{\theta} - P(q)) + (1-\pi_1)v(y_0 + q - P(q))$. The maximum amount $P_{max}(q)$ that the CP is willing to pay for a contract providing cover q is given by:

$$V(\pi_1, q, P_{max}(q)) = V(\pi_1, 0, 0), \text{ or}$$
$$\pi_0 v(y_0 + p_i M\hat{\theta} - P_{max}(q)) + \pi_1 v(y_0 + q - P_{max}(q)) = \pi_0 v(y_0 + p_i M\hat{\theta}) + \pi_1 v(y_0)$$

What about the other side of the market?

If we assume that the insurer has utility function u, initial wealth k then any contract $(q, P(q))$ such that

$$\pi_0 u(k + P(q)) + \pi_1 u(k + P(q) - q) \geq u(k)$$

would be beneficial to the insurer. At $P(q) = P_{min}(q)$:

$$\pi_0 u(k + P_{min}(q)) + \pi_1 u(k + P_{min}(q) - q) = u(k)$$

The only requirement for the contract $(q, P(q))$ to be mutually beneficial is that:

$$P_{max}(q) \geq P(q) \geq P_{min}(q)$$

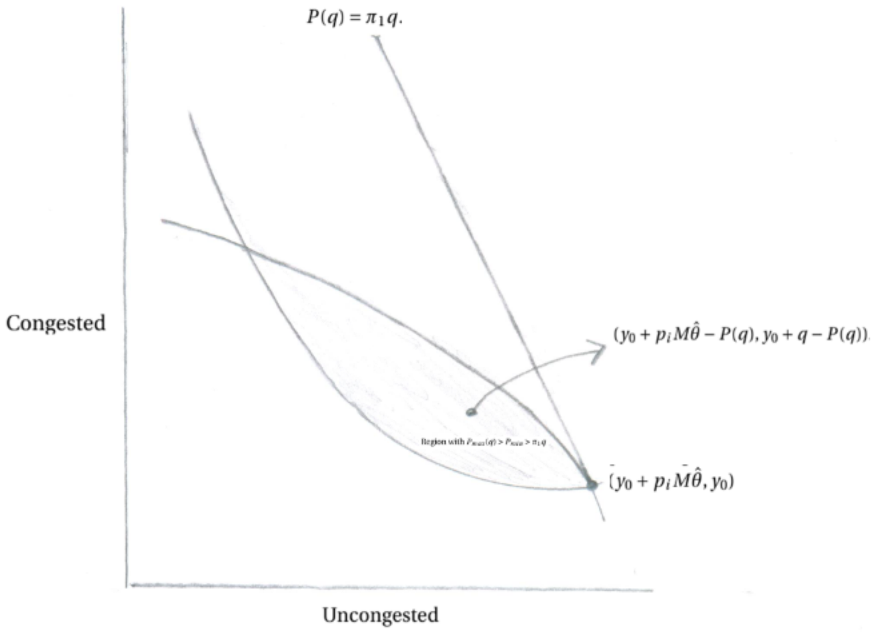

Figure 5.4: The state space diagrams. The x-axis represents income in the state of no congestion. The y-axis represents the state of congestion. The feasible contracts like between the two indifference curves

In the state space diagram (Figure 5.4), the set of contracts extend along a region from initial revenue $(y_0 + p_i M\hat{\theta}, y_0)$ to $(y_0 + p_i M\hat{\theta} - P(q), y_0 + q - P(q))$. For example, if the price function of the contract $P(q) = mq$ for some m, then the set of contracts extend along a line with slope $-\frac{1-m}{m}$. Any contract $(q, P(q))$ which lies in the region but above the indifference curve passing through the initial revenue $(Y_0 + p_i M\hat{\theta}, y_0)$ would be beneficial for the CP. For these contracts, $V(\pi_1, q, P(q)) > V(\pi_1, 0, 0)$.

If the insurer is risk-neutral ($u'' = 0$), $P_{min}(q) = \pi_1 q$. In this case, the only welfare enhancing contracts lie between the region bounded by the content provider's indifference curve and the line with slope $-\frac{1-\pi_1}{\pi_1}$ passing through the initial revenue $(y_0 + p_i M\hat{\theta}, y_0)$.

If the insurer is risk-averse, but less risk averse than the CP, then $P_{max}(q) > P_{min} > \pi_1 q$

This answers our first question: a congestion option offered within the region lying above the indifference curve for the CP and below the indifference curve for the insurer does indeed improve the welfare for all parties involved in the transaction. If the insurer is risk-neutral, its indifference curve is just its the line where $P(q) = \pi_1 q$.

Let us answer the second question: the fair value of a congestion option. If the valuation is actuarially fair, then the premium should be equal to the expected value of the insurance payout (τ); $P(q) = E[\tau] = \pi_1 q$. Note that contract would be beneficial even if the pricing is actuarially unfair, or if insurance seller makes non-zero profit, as long as expected utility maximizing optimal contract lies above the indifference curve passing through the intitial revenue. In this case the price of the contract is not equal to the expected value of the payoff, and $P(q^*) > \pi_1 q^*$. The extra margin can incorporate insurer interests such as higher profits or sufficiently protective solvency margins or risk aversion.

To calculate the price when the contract is traded in a free market, we need to make use of the notion of a competitive equilibrium. Assume that at equilibrium:
(1) There is sufficient competition in the supply side so that the contract seller makes no utility gain
(2) A risk averse content provider maximizes its expected utility.

Assumption (1) implies $P(q) = P_{min}$. If the insurer is risk neutral,$P(q) = P_{min} = \pi_1 q^*$.[22]

Maximizing expected utility $V(\pi_1, q, P(q)) = \pi_0 v(y_0 + p_i M\hat{\theta} - P(q)) + \pi_1 v(y_0 + q - P(q))$ requires:

$$\frac{\partial V(q^*)}{\partial q} = 0$$

$$\pi_0 v(y_0 + p_i M\hat{\theta} - P(q^*))(-P'(q^*) + \pi_1 V v(y_0 + q^* - P(q^*))(1 - P'(q^*)) = 0$$

Rearranging,

$$\frac{\pi_0 v(Y_0 + p_i M\hat{\theta} - P(q^*))}{\pi_1 v(Y_0 + q^* - P(q^*))} = \frac{1 - P(q^*)}{P(q^*)}$$

For this to be a maxima, the V must be concave function of q:

$$\frac{\partial^2 V}{\partial q^2} = V_{qq} < 0$$

because $v'' < 0$[23]

[22]Since the premium equal to the expected value of insurance payout, this pricing is actuarially fair.
[23]This can be worked out by differentiating V twice with respect to q.

For example, if $P(q) = \pi_1 q$, this reduces to

$$v(Y_0 + p_i M\hat{\theta} - P(q^*)) = v(Y_0 + q^* - P(q^*))$$
$$\implies q^* = p_i M\hat{\theta} \text{ (since the utility function is monotonous } v'(R) > 0).$$

In a competitive equilibrium under fair pricing, the CP would prefer to cover his full revenue loss $p_i M\hat{\theta}$ from congestion, and the premium would be equal to the expected revenue loss due to congestion $P(q) = \pi_1 p_i M\hat{\theta}$. Note that if there is a significant time lag between the buying the contract and the payout, then the price would be the *discounted value* of expected payout, with the discount rate being the risk-free interest rate of the market.

It is worth noting that our market in congestion options differs from the insurance market in several aspects:

1. While the holder of an insurance contract has to prove that they have suffered a financial loss as a result of congestion, no such demonstration is required from the holder of a congestion option. A congestion option owner receives the payout from the contract based on the actual outcome of congestion, regardless of how it affects them. A market participant doesn't have to be sensitive to congestion, as these can be bought for mere speculation or diversification of a portfolio.

2. The other difference is that insurance contracts are designed to protect the holder from extreme events with very low probability. They do not work very well with uncertainties arising from normal fluctuations in the underlying and having a range of outcomes. Congestion derivatives, on the other hand, would be useful under any degree on uncertainty in the congestion levels.

3. A third advantage of congestion derivatives is that they allows two participants with different risk preferences to efficiently share risk. An insurance market, on the other hand, is generally restricted to a risk averse insurer and a risk neutral insurance company.

5.3.2 Pricing of Securities Linked to Assets

The previous section derives feasible price levels in a bilateral risk sharing contract in an absolute manner through a rational expectations model, where price are equal the discounted expected value of the payouts. This model, however, ignores the other asset information and observable prices available in the market. When the option are connected to another asset via fundamental underlying linkages (common stochastic processes, for example), then the pricing must take into account information about its prices (see Varian (1987) for a good exposition).

To see this, let us return to the problem of valuing a pure option whose payoff, governed by congestion, is given by:

$$\tau(\eta) = \begin{cases} 0 & \text{if } \eta = 0 \\ q & \text{if } \eta = 1 \end{cases}$$

Assume that there's a lag of one time period between buying the option and its outcome, and that the riskless rate of return in the market is r (A risk-free bond B bought for 1 unit of currency in time period zero pays r units in period one). Let the price of the option be $P(q)$,

Consider an asset A, whose price at period i is given by A_i, and whose price at period 1 is governed by the outcome of congestion:

$$A_1 = \begin{cases} A_1(0) & \text{if } \eta = 0 \\ A_1(1) & \text{if } \eta = 1 \end{cases}$$

Construct a portfolio that has x shares of the the asset A and B bonds that pay off $B(1+r)$ in period one. Then this portfolio has a return of:

$$xA_1 + B = \begin{cases} xA_1(0) + B(1+r) & \text{if } \eta = 0 \\ xA_1(1) + B(1+r) & \text{if } \eta = 1 \end{cases}$$

Choose x and B so as to create the same return pattern as the pure option:

$$x^* A_1(0) + B^*(1+r) = 0$$
$$x^* A_1(1) + B^*(1+r) = q$$

The solution to these equations is given by:

$$x^* = \frac{q}{A_1(1) - A_1(0)}$$
$$B^* = \frac{-qA_1(0)}{(A_1(1) - A_1(0))(1+r)}$$

Since this portfolio x^*, B^* has the same returns as the pure option, it must have the same price. This is because of the *no-arbitrage principle*, which basically says that "There are no free lunches" in a market at equilibrium. If the price of the option ever deviated from the price of the portfolio that generates the same pattern of payoffs, then there would be a sure way of making money - just sell the option and buy the portfolio, or vice versa, depending on which is worth more - without bearing any risk. Equilibrium conditions require that all

such opportunities have been exploited.
Thus, the price of the pure option must be given by:

$$
\begin{aligned}
P(q) &= x^* A_0 + B^* = \\
&= \frac{q A_0}{A_1(1) - A_1(0)} + \frac{-q A_1(0)/(1+r)}{A_1(1) - A_1(0)} \\
&= q * \frac{A_0 - A_1(0)/(1+r)}{A_1(1) - A_1(0)}
\end{aligned}
$$

If $\pi^* = \frac{A_0 - A_1(1)}{A_1(1) - A_1(0)}$, then $P(q) = \pi^* q/(1+r)$. Call π^* the risk-neutral probability.

This result is different from our earlier result. In the rational expectations model, we saw that in the absence of any underlying asset linked to the congestion process, the price of the option contract is given by $P(q) \geq \pi_1 q$. If the insurer is risk-neutral, $P(q) = \pi_1 q/1 + r$, the price is equal to the discounted value of the expected payoff under the *actual* probability of congestion.

However, if the option is linked to an asset through common dynamics (for example, the dynamics of congestion), then the actual probability of congestion event $\pi_1 = P(\eta = 1)$ has no role to play in the pricing. The price depends on the risk neutral probability $\pi^* = \frac{(1+r)(A_0 - A_1(1))}{A_1(1) - A_1(0)}$, which depends on *how much* the linked asset values change as the state of congestion changes. Instead of the discounted value of the expected payoff under the *actual* probability of congestion, the price is given by the discounted value under the *risk-neutral* probability.

The *completeness* of the market is crucial for the existence of a unique risk-neutral probability measure, and hence unique prices. According to Embrechts (2000):

> *If our market is such that we have sufficiently many basic building blocks in the market so that new assets can be represented as linear combinations of these building blocks, and these building blocks have a unique price then the market is termed complete. If not, the market is incomplete. In the former cases (completeness) prices are unique whereas in the second case (typical in insurance) without further information on investor specific preferences, only bounds on prices can be given.*

If the stochastic process followed by the asset values is known, then the precise value of the option, or any other derivative linked with the asset, can be calculated using only the no-arbitrage principle. For example, the Black-Scholes model for option pricing assumes that the log returns of the stock price is an infinitesimal random walk with drift; more precisely, a geometric Brownian motion (Black and Scholes, 1973). This is found to aptly describe the movement of stock prices, and any security whose payoff is linked to stock price movements can be priced uniquely using the no-arbitrage principle. Similarly, the Ornstein-Uhlenbeck process (which describes the velocity of a massive Brownian particle under the influence of friction) has been suggested to describe temperature movements for pricing temperature-based weather derivatives (Alaton et al., 2002).

5.4 A Stochastic Model of Local Congestion

Consider again the general Congestion Call Option. A congestion call option would give the owner, the right to receive q units of currency (call it 'tick size') for each basis point of congestion exceeding a minimum threshold amount $\tilde{\eta}$ at time T. The payoff function is:

$$\tau = \begin{cases} 0 & \text{if } \eta \leq \tilde{\eta} \\ q(\eta(T) - \tilde{\eta}) & \text{if } \eta(T) > \tilde{\eta} \end{cases}$$

Since the underlying index of a congestion derivative $\eta(T)$ is non-tradeable (unlike stocks in a stock market), the assumptions of completeness and no-arbitrage required for pricing like in the Black Scholes model are not applicable.

In the absence of a stochastic model of congestion, the simplest way to price such a contract is through historical analysis of how a contract would have performed in previous time periods. The historical payout of the derivative is computed to find the expectated value. (In the context of pricing weather derivatives, this is called Burn Analysis (Clarke et al., 2012)). The attractiveness of this approach is its simplicity. However, the method can be accused of being unreliable, since it assumes that the future would be similar to the past, as well as myopic, because it ignores the specific dynamics, factors and relationships that characterize congestion movements.

However, finding an appropriate stochastic model that describes the process $\eta(T)$ would be a significant step towards rationally pricing congestion based securities. An appropriate measure of congestion could be extracted from observed broadband speeds (see Figure 5.5, for example). The measure can be continuous or discrete time. Depending on the network topology, it could pertain to a particular locale or an average of locales . A good model would also be consistent with spatial and temporal correlation between congestion levels at various nodes and the structure of the network. Once the stochastic process governing congestion is determined, then the prices can be computed at through Monte Carlo Simulations. Monte Carlo simulation generates random paths of the congestion till time T given the information at time s, each of which results in a payoff for the option (possibly zero). The idea is to generate a lot of trajectories of the process and then approximate the expected value with the arithmetic mean. The expected value will need to discounted at the risk free rate r to yield a price for the option.

Example of A Stochastic Model

Any credible model stochastic model of congestion would require market data on congestion patterns. Since the scope of this thesis is limited, we sketch an analytical approach

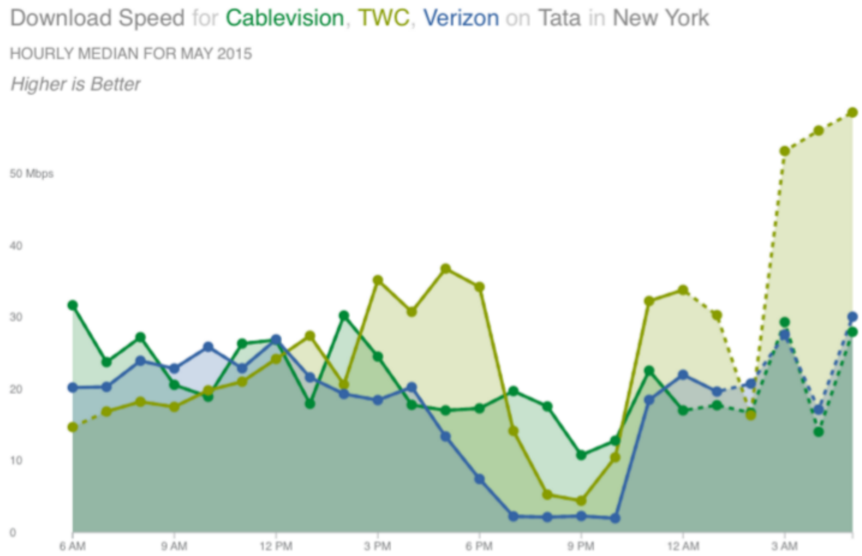

Figure 5.5: Source: Anderson (2015)

through which a stochastic model of congestion could be worked out. This approach combines the methodology used to arrive at a stochastic process for temperature movements in Hamisultane et al. (2008) with the three party ecosystem model of Ma and Misra (2013).

The real time broadband speeds themselves would not serve as an appropriate measure of congestion. This is because broadband speeds are affected by a number of factors other than congestion - by the speed of the computer, use of microfilters , electrical or wireless interference, and distance from the telephone exchange (Broadband, 2015). However, broadband speeds in absence of congestion could give an idea of individual bandwidth demand patterns.

Take monopolistic ISP providing a last mile broadband pipeline to n homogeneous consumers. The n users compete for bandwidth in the single broadband pipeline of width K provided by the ISP. Each user i expresses the amount of bandwidth it wants to receive by submitting a 'bid' b_i (The demand for user i). Depending on the congestion in the network, however, user i may or may not receive the bandwidth it wishes to receive.

If $b_i(t)$ be the stochastic process which governs a typical consumer's demand, one can expect it to possess the following properties:

- The demand would have vary during a day, with peaks usually during evening, after-work hours. The demand pattern on weekdays would be different from the demand pattern during weekdays. However, the demand pattern of one week as a whole would look similar to the demand pattern of another week. It should be possible to incorporate about bandwidth demand movements using some sine function, with the form:

$$sin(\omega t + \phi) \tag{5.7}$$

where t denotes the time, measured in hours. We let $t = 1,2,3...$ Since we expect the period of oscillation to be one week, we have $\omega = \frac{2\pi}{7*24}$. We have introduced the phase ϕ since the weekly maximum or minimum might not be during the start or middle of the week respectively.

- We also expect a long term trend in congestion behaviour. This might be very weak, but it will be useful to incorporate it. To capture this trend we assume as a first approximation that the trend is linear:

$$A + Bt \tag{5.8}$$

- But individual bandwidth demand is not deterministic, and to obtain a more realistic model we have to add some sort of noise or stochastic element. One choice is the standard Wiener process, $(W_t, t \geq 0)$, with quadratic variation $\sigma^2 \in \mathbb{R}_+$
- It might be reasonable to now allow congestion levels to deviate from the deterministic mean value for more than short periods of time. In other words, we would want the individual bandwidth demand to exhibit a *mean-reverting property.*

Summing up, a stochastic model for a typical bandwidth demand pattern at time t, $b_i(t)$, would have the form

$$db_t = a(A + Bt + Csin(\omega t + \phi) - b_t)dt + \sigma_t dW_t \tag{5.9}$$

where a, A, B, C, ϕ are parameters to be computed to fit the curve best with actual data, a being the speed of mean reversion. The solution to this process is called an Ornstein-Uhlenbeck process (also suggested to model temperature for pricing weather derivatives(Hamisultane et al., 2008)).

But our ultimate goal is to build a function that measures congestion. For this we make two reasonable assumptions :

- Assumption 1: The congestion level in the broadband pipeline at any point of time is a function of the bids and the width of the pipeline. Let congestion level in the network at any time by the variable Θ. Then $\eta = f(\mathbf{b}, K)$, where $\mathbf{b}$ is the random vector $(b_1, b_2, ..., b_n)$. Moreover, f is non-decreasing in $\mathbf{b}$, thus increasing any single bid value while keeping the other bid values constant would not decrease congestion. For the rest of the discussion we assume that the width of the pipeline is fixed. Thus $\eta = f(\mathbf{b})$.
- Assumption 2: The congestion level depends on time insofar as the bid values depend on time. The same bid values at different points of time correspond to the same level of congestion. Thus, $\eta(t) = f(\mathbf{b(t)})$, where $\mathbf{b(t)} = (b_1(t), b_2(t), ..., b_n(t))$

A function satisfying the assumptions given above would be:

$$\eta = \begin{cases} 0 & \text{if } \sum b_i \leq K \\ \frac{\sum b_i}{K} - 1 & \text{if } \sum b_i \geq K \end{cases}$$

While the congestion process arrived at here incorporates historical data as well as random noise, it doesn't take into account known future events which will affect congestion. For example, a forthcoming big sporting event or network pipeline refurbishment can significantly affect congestion. Therefore, when we want to model congestion during such a period, the model must be adjusted to take these into account. The parameters: A, B, C, a, ϕ can be calibrated to make such adjustments.

5.5 Caveats

Efficient risk markets can unequivocally improve social welfare. There is thus an exceedingly strong economic case for a market in congestion risk. However, markets don't exist for all kinds of risk. For example, there are no any markets for risks of bad decisions, or low earnings, or getting fired from a job. A waiter whose income is dependent on cash tips from customers would find it difficult to find someone willing to insure him against fluctuations in income. Could a market in congestion risk fail because of similar reasons?

Moral Hazard: Throughout this chapter we have assumed that the traffic from a content provider is not big enough to influence congestion levels. In practice, some CPs might have market power and be able to influence the congestion outcome in their favour. The quasi-monopolistic nature of CPs is not a trivial concern: for example, Netflix alone accounts for more than a third of peak traffic in the United States. A buyer of congestion option might have incentives to influence the the congestion outcome in its favour. However, unlike in most insurance markets, behaviour of market participants can easily be tracked by monitoring bandwidth usage and congestion levels. Regulatory oversight to monitor the behaviour of market participants can also mitigate moral hazard issues.

Adverse Selection: When buyers and sellers have access to different information, traders with better private information about the value of a contract will selectively participate in trades which benefit them the most (at the expense of the other trader). For example, in job markets a worker may know his effort costs before an employer makes a wage offer. Similarly, an ISP might have private information, like planned refurbishment in network lines, which might give it an unfair advantage in a risk sharing contract. In theory, adverse selection can lead to rapid unravelling of insurance markets (Akerlof (1995), for example, examines how the quality of goods or contracts traded in a market can degrade in the presence of information asymmetry between buyers and sellers, leaving only bad quality goods or unprofitable deals behind). Signalling and screening mechanisms can however help mitigate adverse selection. Indeed, empirical evidence for presence of adverse selection in many insurance markets is weak (Cawley and Philipson, 1996; Chiappori and Salanie, 2000), suggesting that the screening mechanisms employed at the time of underwriting insurance contracts are usually quite effective towards resolving informational asymmetries.

Non-diversifiable risk: Besides information assymetry problems, another reason for the failure might be the presence of systematic or non-diversifiable risks. For example, few would be willing to buy the risk of a climate catastrophe, because everyone faces the same risk simultaneously. Risk markets work well when investors can diversify or hedge their risk from one contract from another uncorrelated or negatively correlated contract. A concern can be raised that congestion options would be unattractive because risks from adverse congestion outcomes cannot be diversified. However, such concerns only pose a serious threat to insurance markets only when the underlying risks are not well understood or are unquantified. If the underlying systematic risks can be quantified, then they can be priced in within the contracts. The experience of weather and crop insurance markets is paradigmatic: both markets are found to work well despite the widespread adverse weather effects of drought, flood and freeze. Further research towards understanding the stochastic process of congestion would help in the understanding and quantification of the systematic risks involved in congestion.

CHAPTER 6: CONCLUSION

Congestion arises because the current economic arrangements that govern the Internet do not allow price signals to allocate risks and resources efficiently. Prices are an important way to communicate information throughout a market economy. The absence of congestion-sensitive pricing leads to overuse of bandwidth resources. However, implementation costs and consumer discontent make congestion pricing unattractive in practice. Similarly, net neutrality regulations prohibit Internet service providers directly charging content providers because of the potentially harmful effects on innovation overall welfare. We have shown that under net neutrality regulations, a separate market for risks based on congestion outcomes can provide an alternative channel through which costs and risks of congestion are efficiently distributed. Thus, market mechanisms, like the ones discussed in this thesis, have a powerful role to play toward solving the Internet's congestion problem.

References

Akerlof, G. (1995). *The market for lemons: Quality uncertainty and the market mechanism*. Springer.

Alaton, P., Djehiche, B., and Stillberger, D. (2002). On modelling and pricing weather derivatives. *Applied mathematical finance*, 9(1):1–20.

Anderson, C. (2015). New Opportunities for Test Deployment and Continued Analysis of Interconnection Performance measurement lab. `https://www.measurementlab.net/blog/interconnection_and_measurement_update/`. Accessed: 2015-12-02.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics*, 37(3):668–691.

Becker, G. S., Carlton, D. W., and Sider, H. S. (2010). Net neutrality and consumer welfare. *Journal of Competition Law and Economics*, 6(3):497–519.

Bingemann, M. (2015). 'Netflix effect' is slowing the nation's broadband internet speed the australian business review. `http://www.theaustralian.com.au/business/technology/netflix-effect-is-slowing-the-nations-broadband-internet-speed/news-story/1ca3e44dd35ca8e4f7cdfcc16aafaec4`. Accessed: 2015-12-02.

Black, F. and Scholes, M. (1973). The pricing of options and corporate liabilities. *The journal of political economy*, pages 637–654.

Bohn, R., Braun, H.-W., Claffy, K. C., and Wolff, S. (1994). Mitigating the coming internet crunch: Multiple service levels via precedence. *Journal of High Speed Networks*, 3(4):335–349.

Broadband, T. (2015). Understanding Broadband Speed. `http://www.thinkbroadband.com/guide/broadband-speed.html`. Accessed: 2015-12-02.

Cawley, J. and Philipson, T. (1996). An empirical examination of information barriers to trade in insurance. Technical report, National Bureau of Economic Research.

Chiappori, P.-A. and Salanie, B. (2000). Testing for asymmetric information in insurance markets. *Journal of political Economy*, 108(1):56–78.

Choi, J. P., Jeon, D.-S., and Kim, B.-C. (2015). Net neutrality, business models, and internet interconnection. *American Economic Journal: Microeconomics*, 7(3):104–141.

Choudhary, M. A. and Levine, P. (2009). Risk-averse firms and employment dynamics. *Oxford Economic Papers*, page gpp036.

Clarke, D., Mahul, O., Rao, K. N., and Verma, N. (2012). Weather based crop insurance in india. *World Bank Policy Research Working Paper*, (5985).

Conto, R. (2015). Personal Correspondence.

Courcoubetis, C., Weber, R., and Coe, M. (2003). Pricing communication networks: Economics. *Technology and Modelling, John Willey & Sons, Ltd.*

Deng, S. and Oren, S. S. (2006). Electricity derivatives and risk management. *Energy*, 31(6):940–953.

Embrechts, P. (2000). Actuarial versus financial pricing of insurance. *The Journal of Risk Finance*, 1(4):17–26.

Géczy, C., Minton, B. A., and Schrand, C. (1997). Why firms use currency derivatives. *The Journal of Finance*, 52(4):1323–1354.

Hamisultane, H. et al. (2008). Which method for pricing weather derivatives? Technical report.

Helms, K. S. (2015). Personal Correspondence.

Jang, H., Lee, H., and Yi, Y. (2010). On the interaction between isp revenue sharing and network neutrality. In *Proceedings of the ACM CoNEXT Student Workshop*, page 19. ACM.

Jensen, B. A. and Nielsen, J. A. (1995). *Pricing by" no Arbitrage".* Handelshøjskolen i København, Institut for Finansiering.

Jin, Y. and Jorion, P. (2006). Firm value and hedging: Evidence from us oil and gas producers. *The Journal of Finance*, 61(2):893–919.

Klinker, E. (2013). Solving the Internet's Congestion Problem. `http://www.internetlivestats.com/internet-users/`. Accessed: 2015-12-02.

Laffont, J.-J. and Tirole, J. (2001). *Competition in telecommunications.* MIT press.

Lee, R. S. and Wu, T. (2009). Subsidizing creativity through network design: Zero-pricing and net neutrality. *The Journal of Economic Perspectives*, 23(3):61–76.

Li, T. (2015). Personal Correspondence.

Lyons, D. (2016). Usage-based pricing encourages Netflix to become better corporate netizen techpolicydaily.com. `http://www.techpolicydaily.com/internet/usage-based-pricing-encourages-netflix-to-become-better-corporate-netizen/`. Accessed: 2015-12-02.

Ma, R. T. and Misra, V. (2013). The public option: a nonregulatory alternative to network neutrality. *IEEE/ACM Transactions on Networking (TON)*, 21(6):1866–1879.

MacKie-Mason, J. K. and Varian, H. R. (1995). Pricing congestible network resources. *Selected Areas in Communications, IEEE Journal on*, 13(7):1141–1149.

MacKie-Mason, J. K., Varian, H. R., et al. (1993). Pricing the internet. Technical report.

Masnick, M. (2014). Level3 Proves That Verizon Is Absolutely To Blame For Netflix Congestion techdirt.com. https://www.techdirt.com/articles/20140718/06533327927/level3-proves-that-verizon-is-absolutely-to-blame-netflix-congestion-using-verizon shtml. Accessed: 2015-12-02.

Metz, C. (2016). The Counterintuitive Tech Behind NetflixâĂŹs Worldwide Launch wired. http://www.wired.com/2016/01/the-counterintuitive-tech-behind-netflixs-worldwide-launch/. Accessed: 2015-12-02.

Musacchio, J., Schwartz, G., and Walrand, J. (2009). A two-sided market analysis of provider investment incentives with an application to the net-neutrality issue. *Review of Network Economics*, 8(1).

Nicholls, M. (2004). Confounding the forecasts. *Environmental Finance*, 6(1):22–23.

Nisan, N., Roughgarden, T., Tardos, E., and Vazirani, V. V. (2007). *Algorithmic game theory*, volume 1. Cambridge University Press Cambridge.

Njoroge, P., Ozdaglar, A., Stier-Moses, N. E., and Weintraub, G. Y. (2013). Investment in two-sided markets and the net neutrality debate. *Review of Network Economics*, 12(4):355–402.

Page, M., Rossi, L., and Rand, C. (2010). A viable future model for the internet. *AT Kearney*.

Pil Choi, J. and Kim, B.-C. (2010). Net neutrality and investment incentives. *The RAND Journal of Economics*, 41(3):446–471.

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.

Saavedra, C. (2009). Bargaining power and the net neutrality debate. Technical report, working paper, 2010. Available at http://sites. google. com/site/claudiasaavedra.

Singh, R., Yuksel, M., Kalyanaraman, S., and Ravichandran, T. (2000). A comparative evaluation of internet pricing models: Smart market and dynamic capacity contracting. In *Proceedings of Workshop on Information Technologies and Systems (WITS)*, volume 199.

Solsman, J. E. (2014). Net congestion no problem, except for one big 'but': Netflix cnet. http://www.cnet.com/news/net-congestion-no-problem-save-for-one-big-exception-netflix/. Accessed: 2015-12-02.

Stats, I. L. (2016). Internet Live Stats. http://www.internetlivestats.com/internet-users/. Accessed: 2016-01-01.

Stewart, M. (2012). Sky broadband users hit by network congestion alphr. http://www.alphr.com/news/373300/sky-broadband-users-hit-by-network-congestion. Accessed: 2015-12-02.

Stix, G. (1998). A calculus of risk. *Scientific American*, 278(5):92–97.

Varian, H. R. (1987). The arbitrage principle in financial economics. *The Journal of Economic Perspectives*, 1(2):55–72.