The Tragedy of the Last Mile: Economic Solutions to Congestion in Broadband Networks*

Jacob B. Malone† Aviv Nevo‡ Jonathan W. Williams§

Preliminary and Incomplete

Abstract

The growth of the Internet has constrained broadband networks, forcing service providers to search for solutions. We study the effectiveness of several such solutions. We develop a (dynamic) model of daily usage during peak and off peak periods, and estimate consumers price and congestion sensitivity using high frequency usage data. Using the model estimates, we calculate the welfare changes associated with different economic and technological solutions for reducing congestion, including peak-use pricing, throttling connectivity speeds, and local-cache technologies. We find that peak-load pricing combined with local-cache technology is the most effective way to shift activity from peak to non-peak periods.

Keywords: broadband, congestion, nonlinear pricing

JEL Codes: L11, L13, L96.

*We are grateful to the North American Internet Service Providers that provided the data used in this paper. We gratefully acknowledge financial support from NSF (grant SES-1324717), and Jonathan Williams from CableLabs and the NET Institute (www.netinst.org).

†CableLabs, j.malone@cablelabs.com.

‡University of Pennsylvania, anevo@upenn.edu.

§University of North Carolina at Chapel Hill, jonwms@unc.edu.

1 Introduction

The use of the Internet and the demand for online content, especially over-the-top (OTTV) video, is soaring. Internet Service Providers (ISPs) are struggling to keep network capacity in line with this demand. An industry estimate places private broadband investment around \$1.3 trillion between 1996 and 2013, or about \$75 billion per year.[1] Historically, broadband investment has been financed by private firms, but its importance is now leading some public funding to support further investment and additional public policy tools are be used to encourage investment.[2]

Investment in higher capacity and faster networks is clearly needed to deal with the increased demand. However, Nevo et al. (2016) find that there is a wedge between the social return on investments and firms ability to recoup their costs. Furthermore, network investment does not effectively deal with a key feature of most networks: demand is significantly higher during peak-demand periods. Building enough capacity to meet demand during these periods leaves the capacity idle during non-peak periods. Both these suggest that an effective approach should combine investments with economic solutions to reduce the demand during peak periods. In this paper, we explore the effectiveness of such solutions.

We start by estimating demand for online content. We use data on hourly Internet usage and network conditions made available to us by a North American ISP. This provider offers plans with three-part tariffs – consumers pay a monthly fee, receive a monthly allowance and pay per gigabyte (GB) if the allowance is exceeded – which makes the usage decision dynamic within a billing period. This allows us to use (shadow) price variation and variation in network congestion to estimate the impact of congestion on demand, as well as consumers' willingness to pay for a network with no congestion. We then use the estimates to explore the effects of various economic solutions to reduce overall demand and congestion during peak-demand periods. In particular, we explore the effect of throttling speed, peak-use pricing, and local-cache technology.

Before estimating the model, we explore our rich data and demonstrate several patterns that inform some of the questions of interest. First, there is a very consistent within

[1] See USTelecom's estimates at `http://www.ustelecom.org/broadband-industry-stats/investment/historical-broadband-provider-capex` and page 15 of the FCC's 2015 Broadband Progress Report found at `https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2015-broadband-progress-report`.

[2] For example, as a part of the Charter/Time Warner Cable merger review, broadband investment to modernize and expand the network was a condition for approval. (State of New York Public Service Commission's Case 15-M–0388 on "Joint Petition of Charter Communications and Time Warner Cable for Approval", released on January 8, 2016). In another example, the recent debate over the FCC's 2015 Open Internet order, also known as net neutrality, and whether it should be repealed, has focused on whether it slowed investment in broadband networks.

day usage pattern: usage is high during evening and significantly lower most other parts of the day. This pattern is consistent across days and across consumers with different total usage. Second, measures of network performance are correlated with network utilization, with performance decreasing during peak demand periods. This suggests that any policy that shifts demand from peak to non-peak times could improve network performance. Third, most of the usage is due to video, which at least in principle could be downloaded during non-peak times with the right incentives and technology in place. Finally, there is a large amount of heterogeneity in usage across consumers, which should be accounted for in the modeling approach.

Our focus is congestion at the node, which is a network device that connects a group of subscribers to the rest of the operator's network. A node is a common place for bottlenecks to occur and is an essential part of the portion of the network commonly referred to as the "last mile". Buffering video streams, websites failing to load, and being disconnected from an online video game are common examples of how congestion might affect a consumer. ISPs constantly invest in the network by splitting existing nodes and adding new nodes. This is usually done once average utilization exceeds certain thresholds. When a node is split, its subscribers are distributed across two new nodes, and congestion should decrease. We observe several node splits and use these events to compare before-and-after congestion and subscriber usage.[3] After a split, average daily usage increases by 7% and packet loss, our measure of congestion, drops by 27%. This suggests that there is an elasticity of usage with respect to congestion.

To quantify the consumers willingness to pay for a less congested network we develop and estimate a model of Internet usage. The model extends the model of Nevo et al. (2016) in two ways. First, we allow network congestion to impact the effective speed of the network and therefore impact plan choice and usage. Second, we let consumers make both peak and off-peak usage decisions (and not just a daily decision). This allows us to explore the impact of peak-use pricing, as well as other strategies to reduce usage.

We estimate this finite-horizon dynamic model of usage by extending to panel data the methods proposed by Ackerberg (2010), Fox et al. (2011), and Fox et al. (2016), and applied by Nevo et al. (2016). Estimation proceeds in two steps. First, we solve the dynamic problem once for a large number of types. Next, the solution to these dynamic problems is used to estimate the distribution of types by computing for each household

[3] To measure congestion we rely of two measures commonly used by the Federal Communications Commission (FCC): latency, which measure how long it takes requests to move across the Internet and packet loss, which is roughly, the percentage of requests that fail to make it to their destination.See the FCC's 2015 Measuring Broadband Report at `https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-2015`.

in the data the likelihood that their usage were generated by each type. This results in a distribution across types for each household. Aggregating across the households yields a distribution of types in the population. The estimated marginal and joint distributions illustrate the strength of the flexibility built into our estimation approach.

We find that most consumers have a low willingness to pay for faster advertised speeds as well as for increased allowances. On the other hand, on average consumers are willing to pay over 20% more per month to eliminate congestion during peak-demand periods. Turning to our policy counterfactuals, we find that throttling speeds has little effect on usage and actually increases it slightly. The reason usage increases is that consumers have a low willingness to pay for speed. Combined with a low probability of exceeding the allowance leads to higher usage. Next, we explore peak-use pricing and find that it too has little effect on usage. This is likely the case because consumers have limited ability to shift usage from peak to non-peak periods. Our final counterfactual introduces a way to shift usage by introducing a local-cache technology that allows consumers to download during non-peak times and consume at anytime like a DVR technology for OTTV content. We find that such technology is effective in shifting usage to non-peak periods.

This paper, like our previous work Nevo et al. (2016), Malone et al. (2016), and Malone et al. (2014), uses high-frequency data to study subscriber behavior. However, this paper differs in several important ways. Malone et al. (2016), and Malone et al. (2014) are purely descriptive and examine different issues. Nevo et al. (2016) also estimates demand and examines the effectiveness of usage based pricing in reducing usage and the economic viability of building fiber to the home networks. This paper differs in three important ways. First, as we detail above, the demand model is different. Second, the way we estimate the model is different. Previously we only used aggregate moments, while here we use the individual usage data and the panel structure. Finally, we examine a different set of questions and counterfactuals.

The paper also relates to Varian (2002) and Edell and Varaiya (2002), who run experiments where consumers face different prices for varying allowances and speeds. Goolsbee and Klenow (2006) estimate the benefit to residential broadband; Hitte and Tambe (2007) show Internet usage increases by roughly 22 hours per month when broadband is introduced. Other related papers are Lambrecht et al. (2007), Dutz et al. (2009), Rosston et al. (2013), and Greenstein and McDevitt (2011).

Finally, the question of peak load pricing also comes up in other contexts, real time pricing of utilities being a prime example. Recent work in the this area includes Wolak (2006, 2010, and 2016), Strapp et al (2007), Ito (2014) and Anderson et al (2017). For more complete survey see Bowker (2010) and Faruqui and Sergici (2010).

2 Preliminary Analysis

In this section we describe our main data sources and provide some preliminary analysis that motivates some of the modeling that follows. Specifically, we show that there are clear intra-day usage patterns with peaks in the evenings and lower usage at other times. These patterns are consistent across days of the week and subscribers. Second, we show that the network performance is correlated with overall usage. Third, we provide direct evidence of the effect of a reduction in congestion on usage. Finally, we show the composition of traffic and in particular the importance of video.

2.1 Data and Descriptive Statistics

2.1.1 Data

The main dataset we use comes from a North American ISP. The provider offers several plans with features that include a maximum download speed, an access fee, usage allowance, and overage price per GB for data in excess of the usage allowance.[4] Usage in GBs is recorded for both uploads and downloads, but for billing purposes, and consequently our purposes, the direction of the traffic is ignored. For each subscriber, we observe usage and details of network conditions each hour for February 1st through December 31st of 2015. The usage information comes from what is known as Internet Protocol Detail Records (IPDR). These data report hourly counts of downstream and upstream bytes, packets passed, and packets dropped/delayed by each cable modem. The IPDR data also record a cable modem's node. We combine these data with a second data source on hourly utilization by node. We know the plan chosen by the subscriber.

The sample includes 46,667 subscribers, and a total of over 330 million subscriber-day-hour observations. The metropolitan area where the subscribers are drawn from has demographic characteristics that are similar to the overall US population. Average income in the MSA is within 10% of the national average and the demographic composition is just slightly less diverse. Like many markets for residential broadband, our ISP competes with another ISP offering substantially slower services, particularly in more rural parts of the market. Therefore, we expect the insights from our analysis to have external validity in other North American markets.

In addition to the data we use for our main analysis, we discuss statistics from complementary data from another ISP during the the same period. This data set is national in scope and includes information from a deep-packet inspection (DPI) platform, which

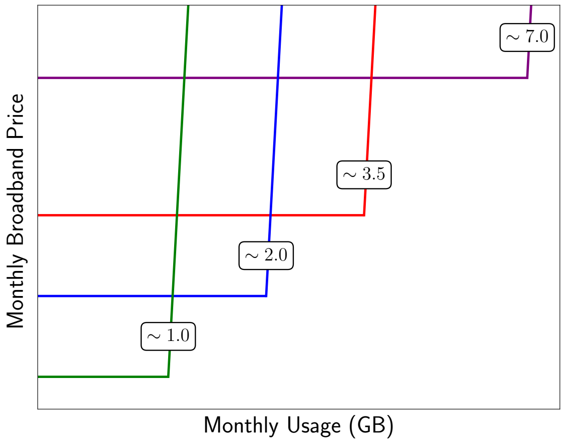
[4] Subscribers are not on long-term contracts, only incurring a disconnection fee if service is canceled.

provides insight into the types of traffic (OTTV, gaming, web browsing, etc) generated by each user. However, the operator has not implemented UBP and has an overbuilt network, both substantially limiting our ability to use it to infer demand. However, the high-level descriptive statistics of the composition of traffic are helpful in explaining the findings from our model, and in providing external validity as to the representativeness of the patterns observed in the our data used in the analysis.

2.1.2 Internet Plans

The ISP sells Internet access via a menu of plans with more expensive plans including both faster access speeds and larger usage allowances. Overages are charged for usage in excess of the allowance. The approximate relationship between monthly usage (GB) and monthly price ($) across plans is shown in Figure 1. The average subscriber pays $58.89 per month for a 22 Mbps downstream connection with a 267 GB usage allowance. The maximum offered speeds and allowances are consistent with those offered in North America, but few consumers choose them (as we have observed in the data of other ISPs with similar offerings).

Figure 1: Internet Plan Features



Note: This figure represents the approximate relative relationship between monthly usage and price for the ISP's menu of plans. Since this ISP has implemented usage-based pricing, there is a set usage allowance for each plan and usage in excess of the allowance is billed. The box label that intersects each plan's line represents the approximate relative differences in speeds.

Consumers on more expensive Internet plans use more data on average. In Table 1, we present the distribution of daily usage for each of the plans, and the distribution of consumers across plans. Most notable is that over 90% of subscriber-day observations are from Tiers 1 and 2, as most subscribers find the larger allowances and speed to not be worth the cost. The distribution of usage for more expensive plans stochastically dominate lesser tiers. Median (average) usage on the highest tier is over thirteen (six) times greater than the lowest tier, and the standard deviation is over three times greater. Thus, consumers with greater and more variable usage select more expensive plans, similar to the findings of Lambrecht et al. (2007).

Table 1: *Daily Usage Distributions by Internet Plan Tier*

	Tier 1	*Tier 2*	*Tier 3*	*Tier 4*	*All*
Mean	1.4 GB	3.4 GB	5.4 GB	8.2 GB	2.3 GB
Std. Dev.	2.9	5.0	7.3	10.4	4.5
25^{th} %tile	0.0	0.3	0.6	1.3	0.1
Median	0.4	1.5	3.1	5.3	0.6
75^{th} %tile	1.5	4.7	7.6	11.4	2.7
90^{th} %tile	4.1	9.0	13.6	19.4	6.7
95^{th} %tile	6.3	12.5	18.5	26.1	10.2
99^{th} %tile	12.8	22.3	32.0	46.2	20.3
N	8,539,830	2,910,234	1,117,680	320,085	12,887,829

Note: This table reports daily usage statistics for the four Internet service plans and entire sample for which the unit of observation is the *subscriber-day.*

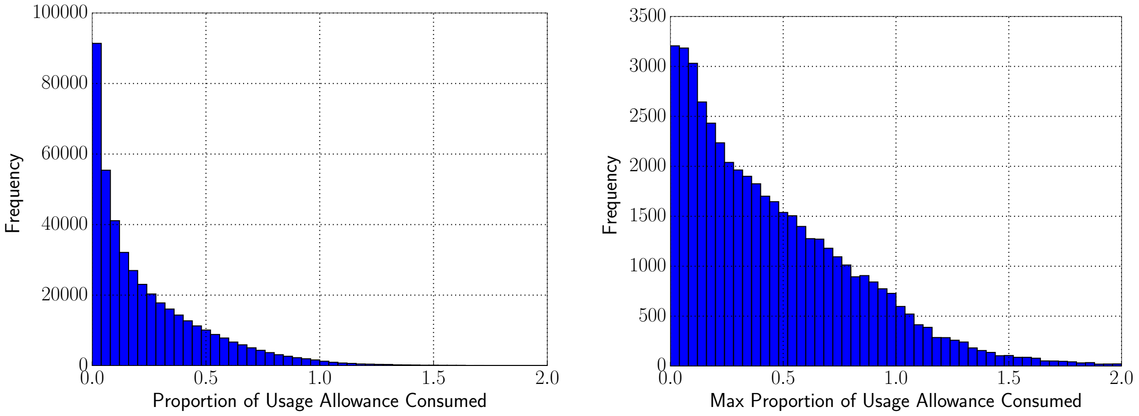
2.2 Temporal Patterns in Usage

In this subsection we display patterns in usage over days and within days. First, we study the behavior of usage during the billing cycle. We start by looking at the fraction of observations that exceed the allowance. This gives us insight into the importance of the dynamics to the consumers. We show that a significant fraction of consumers get near the allowance at some point during the 10 month period. We then show direct evidence that usage responds to the shadow price implied by the dynamic optimization problem. Second, we show patterns of usage within the day. We find a very regular pattern of peak usage during the evening and much lower usage at other times.

2.2.1 Inter-Day Usage

Exceeding the usage allowance is fairly infrequent in our sample, as only about 2.5% of subscriber-month observations have usage in excess of the allowance. The distribution of the ratio of usage to the usage allowance at the subscriber-month unit of observation is presented in Figure 2(a). Most individuals, particularly those on less-expensive plans use only a small amount of their allowance. However, there is considerable variability in usage from month to month. Figure 2(b) provides a histogram of the maximum proportion of the monthly usage allowance used by each customer over the eleven month sample. Approximately 14% of customers exceed their usage allowance during the panel, and the average of this maximum usage is over 70%. Observing most consumers making marginal decisions during the panel is helpful for identification purposes.

Figure 2: *Distribution of Proportion of Allowance Used*



(a) All Billing Cycles

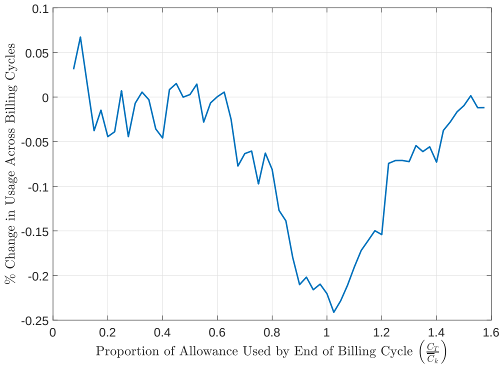
(b) Max Billing Cycle

Note: This figure presents two figures related to the distribution of the proportion of the allowance used by consumer's each month. In panel (a), we present the distribution of this proportion for all customer billing cycles resulting in 11 observations for each customer. In panel (b), we present the distribution of the maximum of this proportion for each consumer across billing cycles, resulting in 1 observation for each customer.

A natural question to ask is whether consumers respond to the shadow price variation. Nevo et al. (2016) explore this question using both within-month and across-month variation. Here we repeat their cross-month analysis. Specifically, subscribers encounter a change in the shadow price when their usage allowance is refreshed at the beginning of a new billing cycle. A forward-looking subscriber near the allowance at the end of a billing cycle knows that the shadow price decreases at the beginning of the next billing cycle. Conversely, a subscriber well below the allowance likely experiences an increase in the shadow price as the new billing cycle begins.

Figure 3 summarizes responses to this end-of-billing-cycle price variation. Specifically, for each subscriber, we calculate the percentage change in usage from the three days before the last day of the billing cycle to the first three day of the next billing cycle. Then we calculate the mean percentage change for groups of subscribers that used various fractions

Figure 3: *Across Month Dynamics*



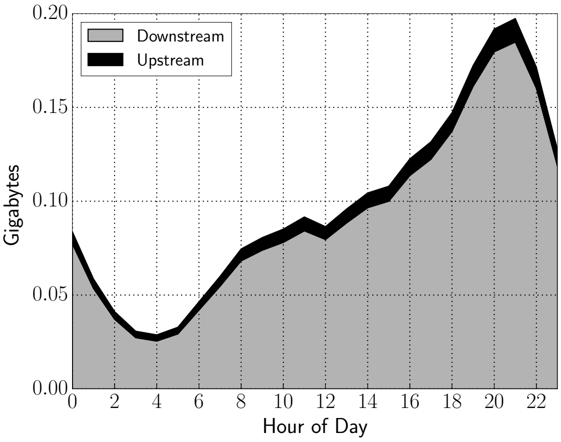
Note: This figure presents how the percentage change in usage from the last three days of a billing cycle to the first three days of the next varies with the proportion of the allowance consumed by a subscriber at the end of the billing cycle.

of the allowance by the end of the month. Like Nevo et al. (2016), we find subscribers facing a price increase at the beginning of the next month consume relatively more at the end of the current month, while those expecting a price decrease consume relatively less. We observe little change in usage for those well above the allowance in the current month. This provides support for the hypothesis that subscribers are forward looking.

2.2.2 Intra-day Patterns in Usage

Temporal patterns in usage play an important role for understanding the potential for more efficient use of broadband networks. Figure 4 presents average daily usage for each hour in both the upstream (e.g. uploading a file to icloud) and downstream (e.g. streaming movie from Netflix). The proportion of downstream traffic is approximately 90% at every hour of the day. This directional disparity is almost exclusively due to OTTV and web browsing being heavily asymmetric and constituting the majority of traffic at all hours. Usage follows a cyclical pattern of maximum usage around 9PM (0.2 GBs) and minimum usage around 4AM (0.03 GBs). This pattern is nearly identical to what is found in Malone et al. (2014) with IPDR data from 2012. Throughout this analysis, we will refer to 12PM–12AM as *peak hours*, i.e. the 12 hours when the network is most highly utilized, and the rest of the day as *off-peak hours*. Approximately 70% of usage occurs during peak hours with the 9PM hour alone accounting for over 8% of daily usage. We find that these average temporal patterns do not differ substantially by the day of the week.

Figure 4: *Temporal Usage Patterns*



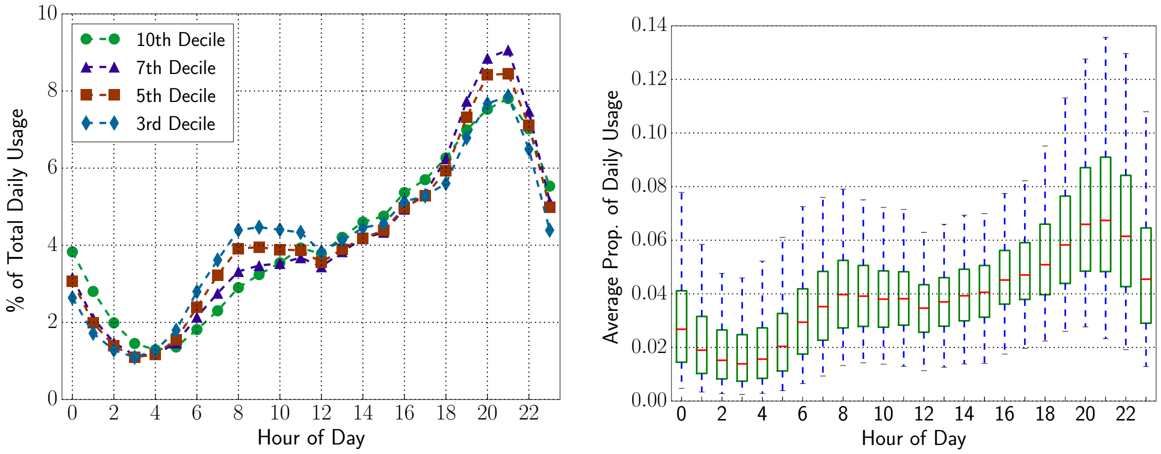
Note: This figure presents statistics on how usage is distributed throughout the day.

We find that the pattern in daily usage does not consistently relate to the level of a consumer's overall usage, despite substantial heterogeneity in temporal patterns across consumers. To see this, we calculate the proportion of total usage for each consumer during each hour of the day over the entire panel. Figure 5(a) presents the mean proportions for different deciles of users where consumers are assigned to deciles based on their total usage. The heaviest-usage consumers (10^{th} decile) have only a slightly flatter profile throughout the day, revealing a very weak correlation between the volume and timing of usage. Yet, the absence of a strong relationship between the volume and timing of usage hides substantial heterogeneity in the timing of usage across consumers within any given decile.

For each hour of the day, Figure 5(b) presents the distribution across consumers of the proportion of usage during that hour. For example, during the 9PM hour 50% (95%) of people have average usage that is less than 6.5% (13.7%) of their average daily usage over our panel (line within the box represents median). The box and whiskers capture the interquartile range (25^{th} and 75^{th}) and the 5^{th} and 95^{th} quantiles, respectively. The dispersion at every hour is indicative of substantially different temporal usage patterns across consumers, albeit not correlated with the overall level of the consumer's usage. In Section 3 we discuss how we account for this important source of heterogeneity in our model.

Together, Figures 4 and 5 demonstrate a clear pattern in usage across times in the day. This usage pattern implies that either at peak-demand times the network will be extremely congested or there is a large amount of excess capacity for the majority of the

Figure 5: *Statistics of Usage as a Percentage of Daily Total*



(a) Daily Usage %s for Total Usage Deciles

(b) Heterogeneity in Daily Usage %

Note: This figure presents two figures related to how temporal patterns in usage varies by consumer. In panel (a), we report hourly percentages for deciles 3, 5, 7, and 10, where the deciles are calculated using each consumer's total usage across the entire panel. That is, the 10^{th} decile includes consumers in the top 10% of all consumers in terms of average monthly usage. Each series sums to 100%. In panel (b), we report variation in the temporal profile across all users. Specifically, for each user we calculate the proportion of their overall traffic used during each hour of the day. Panel (b) reports the heterogeneity in these proportions across consumers during each hour.

day.

2.3 Network Congestion and Usage

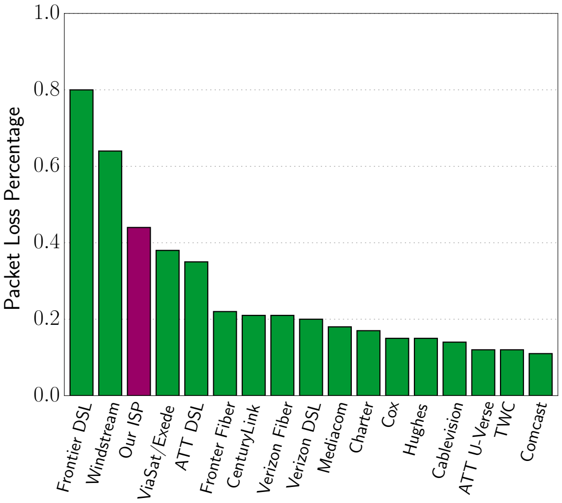
In this section we demonstrate that network congestion impacts usage. We first show the link between congestion and packet loss, a measure of connection quality. We then provide direct evidence on the effect of congestion on usage.

2.3.1 Network Congestion and Packet Loss

Network congestion occurs when subscriber demand exceeds some capacity constraint on the network. During congested periods, subscribers may find that websites fail to load or online video buffers multiple times. There are two ways to measure congestion in our data. One is through hourly average node utilization. The node being the primary bottleneck in the "last mile" of an ISP's network. The second being the hourly proportion of packets dropped/delayed, which we, and others, refer to as *packet loss*.

Hourly packet loss has two advantages over node utilization as a measure of congestion.

Figure 6: *Industry Statistics on Packet Loss from FCC's 2015 Report*



Note: This Figure is a reproduction from the FCC's 2015 Measuring Broadband America Fixed Report (https://www.fcc.gov/reports-research/reports/measuring-broadband-america/measuring-broadband-america-2015), which we have modified to include statistics from our ISP's sample. From the FCC report, it is not clear exactly how their statistics are calculated, but our personal experience with SamKnows data suggests the statistics are the average of hourly packet loss percentages from a specific test conducted by the modem. We calculate the average for "our ISP" according to this methodology. If we aggregate to a daily level, the average is 0.9%.

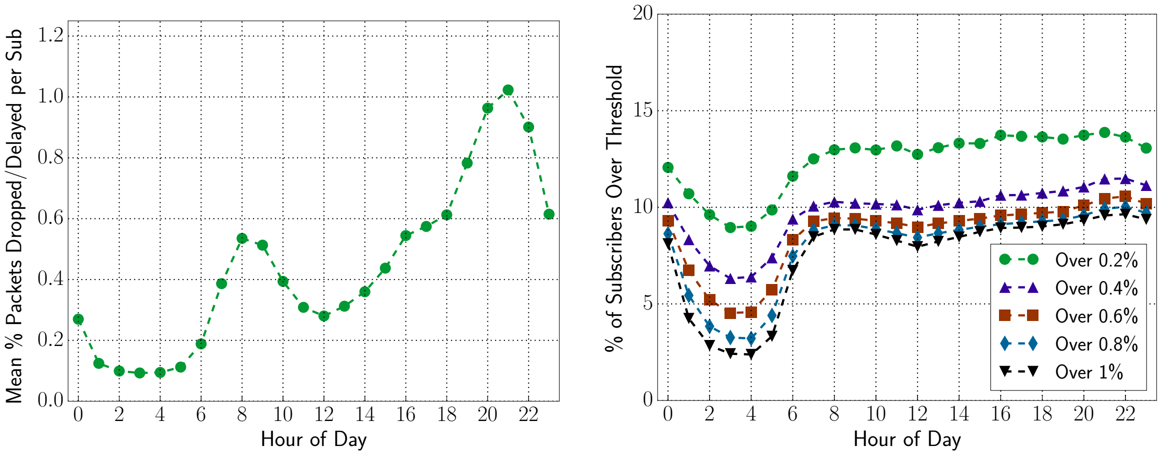
First, packet loss is an individual measure instead of an aggregate one. Even when a node is highly utilized, some subscribers may have a normal experience over the hour. Packet loss occurs when data is undeliverable to a subscriber because current network delivery queues are full. Packet loss is more likely to occur when nodes are highly utilized, which we observe in our data. Second, we only observe node utilization at the hourly level, which may not be granular enough to accurately reflect a subscriber's experience within an hour. However, the performance of the network at the instant the subscriber sends and receives packets will be reflected in subscriber-specific hourly packet loss measures.

The network of the ISP in our data is congested relative to typical US networks. Figure 6, presents statistics on average hourly packet loss from our data and from FCC data for other ISPs. This particular measure of network performance would rate our ISP as the third worst across all types of networks in the FCC data (DSL, cable, fiber, and satellite).

The distribution of hourly packet loss is highly skewed. For example, in panel (a) of Figure 7, which plots average packet loss during each hour, average packet loss is around

1% at 9PM. However, from panel (b), we find over 90% experience less than 1% packet loss. Panel (b) of Figure 7 better captures the right-tail of the packet loss distribution through the percentage of subscribers that are over various packet loss thresholds by hour. Notice in the early morning, when packet loss is lowest, about 3% of subscribers still experience about 1% packet loss on average, compared to the day's maximum of 10% during peak hours. Interestingly, after 8AM the percentage of subscribers exceeding each threshold remain fairly constant over the remainder of the day.

Figure 7: *Average Hourly Subscriber Packet Loss*



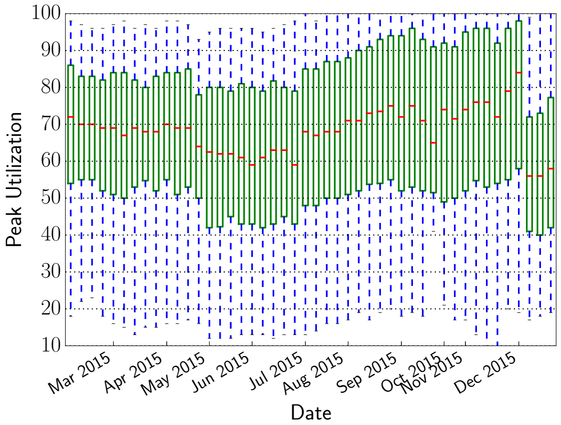
(a) Average Packet Loss

(b) Variation in Packet Loss

Note: These figures report statistics of average hourly packet loss. For each subscriber in the sample, we average hourly packet loss across the panel to generate these figures. In panel (a), we report average hourly packet loss for all subscribers. In panel (b), reports the percentage of average packet loss that is over various thresholds. For each subscriber in the sample, we average hourly packet loss across the panel. The percentage of hourly observations over 0.2%, 0.4%, 0.6%, 0.8%, and 1% are shown in the figure.

Therefore, the majority of people experience little packet loss over the day, but in some cases, packet loss is very severe. The effects of packet loss on customer experience can be variable, too. For example, when watching a streaming video, 0.5% of packet loss may be acceptable for the video to finish. However, if someone is browsing a website, dropping a single packet could be the difference in a website failing to load correctly. This is important from a modeling standpoint, as we provide a flexible framework to estimate a rich distribution of tastes, which accounts for heterogeneity in the types of content the individual prefers to consume.

Figure 8: *Weekly Node Utilization Statistics*



Note: Figure presents the weekly variation in peak utilization of network nodes. The green box is the IQR in each week, the red line is the median, and the blue dashed lines extend to the 5th and 95th percentiles.

2.3.2 Direct Evidence of the Impact of the Network on Congestion and Usage

In Figure 8, we plot the distribution of node utilization for each week in our panel. We see an overall trend of increased utilization with distinct drops in May, September, and December when the ISP improved node capacity. These changes are also noticeable in how median peak utilization varies. The dashed whiskers represent the 5th and 95th percentiles of peak usage, where even during these network events the variation within a week is unaffected.

One way an ISP can alleviate congestion on a node is to perform a node split, for example, by splitting its subscribers across two new nodes.[5] When such a change is made, the network state for the affected subscribers should be improved since there are half as many subscribers using the same node. If subscriber behavior is responsive to such changes in network quality, we would expect an increase in usage. Note that the increase in usage could come from a change in the subscriber himself, or bandwidth adaptive applications becoming more responsive.

There are 5 distinct node splits in the data, whereby a group of subscribers is clearly split over two new nodes. We summarize the changes after these splits in node utilization, packet loss and usage in Table 2. We do see improvements in the average network state with decreases in both utilization and packet loss. Maximum hourly node utilization

[5] This is just one option available to an ISP – an ISP can use other hardware, software, and licensing methods to change the capacity of and bandwidth made available to a node.

Table 2: *Changes in Node Utilization, Packet Loss and Usage After Node Split*

	Before	*After*	*Diff*	*% Change*
Hourly Utilization	49%	34%	-15%	-31%
Max Hourly Utilization	87%	62%	-25%	-29%
Hourly Packet Loss	0.11%	0.08%	-0.03%	-27%
Max Hourly Packet Loss	1.0%	0.61%	-0.39%	-39%
Off-Peak Usage	0.75 GB	0.74 GB	-0.01 GB	-1.3%
Peak Usage	1.80 GB	1.99 GB	0.19 GB	10.5%
Total Daily Usage	2.55 GB	2.73 GB	0.18 GB	7.1%

Note: This table reports how the averages of node utilization and packet loss compare before and after the node split. 7 days of data is taken from before and after the node split date to calculate means. These averages are at the node level of observation and are weighted by the number of people on the node.

falls by 29% and maximum hourly packet loss falls by 39%. Over this same period, we find a 7.1% increase in daily usage. Peak usage increases by 10.5%, while off-peak usage decreases 1.3%. This suggests that there is some degree of unmet demand prior to the node split that is now able to be realized, and weak evidence of intra-day substitution to avoid congestion during peak hours.

2.4 Composition of Usage and External Validity

Our main data source, discussed above and used for our analysis, does not provide information into the types of online activities, only the volume and timing. We do, however, have data on the composition of usage from a another ISP's network. The data is from over 500,000 customers and national in scope, and from the same period, February – December 2015. We do not use these data in our analysis, only to demonstrate the patterns in the data used for the analysis, particularly temporal ones, are representative, and to provide further insight into usage patterns and rationalize some of the predictions from our model estimates.

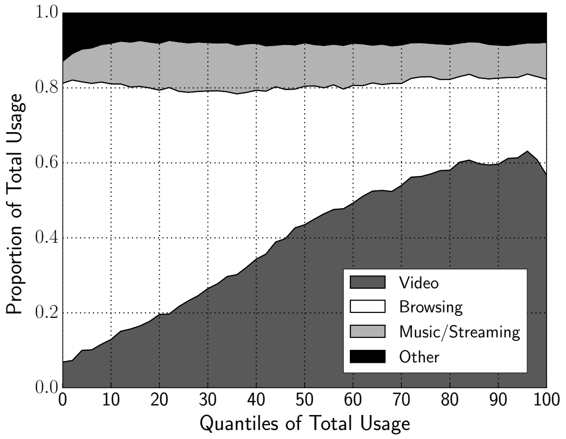
Table 3 presents the percentage of usage for each of several categories. We find that video, music, and streaming collectively account for over 65% of overall traffic, while browsing represents nearly 27%. Perhaps surprisingly, since they previously represented important sources of growth in Internet traffic, combined gaming and sharing represent less than 4% of traffic, while all other sources represent a negligible share. Thus, for the remaining descriptive statistics below we use only four categories: browsing, video, music/streaming, and other.

Table 3: *Percent Usage by Application*

Groups	*Description (Examples)*	*% of* *All Usage*
Administration	System administrative tasks (STUN, ICMP)	1.19
Backup	Online storage (Dropbox, SkyDrive)	0.58
Browsing	General web browsing (HTTP, Facebook)	26.70
CDN	Content delivery networks (Akamai, Level3)	2.95
Gaming	Online gaming (Xbox Live, Clash of Clans)	3.06
Music	Streaming music services (Spotify, Pandora)	3.40
Sharing	File sharing protocols (BitTorrent, FTP)	0.20
Streaming	Generic media streams (RTMP, Plex)	6.26
Tunneling	Security and remote access (SSH, ESP)	0.07
Video	Video streaming services (Netflix, YouTube)	55.47
Other	Anything not included in above groups	0.13

Figure 9 presents the composition of total usage for each quantile of user, where the quantiles are defined based on average monthly usage over the sample. For example, the user with the median average-monthly usage has traffic that is approximately 42% video, 28% browsing, 10% music/streaming, and the remaining 10% of traffic from all other sources. Interestingly, there is a nearly monotonically increasing pattern between the proportion of video and a consumer's overall usage. For high-usage consumers, the greater proportion of video is associated with a lesser proportion of browsing, other proportions remain largely unchanged.

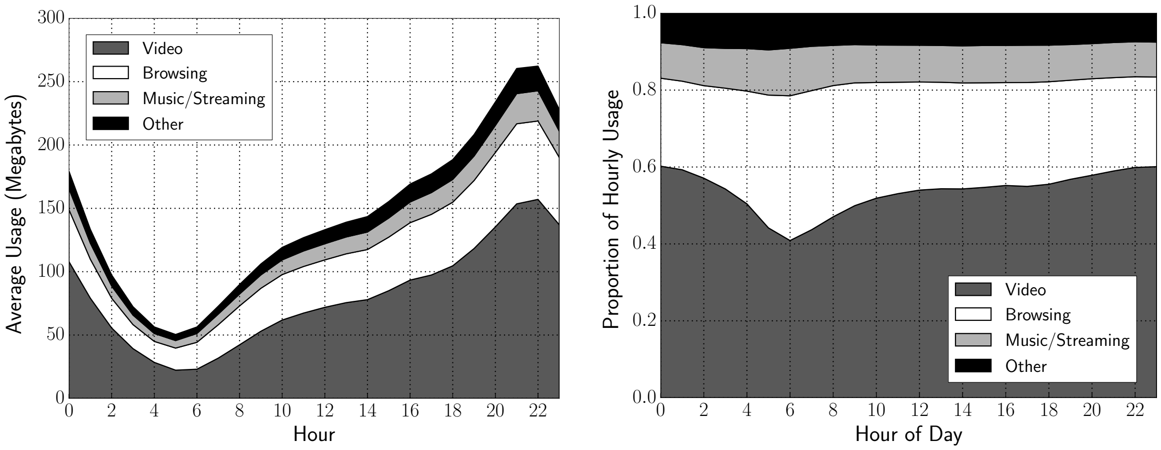
Figure 9: *Data: Monthly Usage by Quantile and Traffic Type*



Note: The average proportion by hour for certain deciles of users where the deciles are based on total usage (e.g., the tenth decile is the top 10% of consumers).

Figure 10(a) and 10(b) present average usage by hour and traffic type, and the proportion that each traffic type accounts for at each hour, respectively. The overall temporal

Figure 10: *Data: Hourly Usage by Group*



(a) Deciles of Daily Usage %s by Hour

(b) Distribution of Daily Usage %s by Hour

Note: This figure presents two figures related to how daily usage is proportionally distributed across the day. In panel (a), we report the average proportion by hour for certain deciles of users where the deciles are based on total usage (e.g., the tenth decile is the top 10% of consumers). In pane (b), we report the distribution across all consumers of these proportions at each hour.

pattern in average usage is nearly identical to the pattern from the data above that is used for our analysis. Video is the most peak-intensive activity, which is not surprising given that most OTTV services require the user to download the movie at the same time as viewing it.

These statistics provide a external validity of the data we use for the analysis, in terms of its representativeness of US broadband usage patterns. Peak usage is only slightly higher in the DPI data, which is consistent with unlimited usage allowances and a less-congested network.

3 Model

To quantify the consumers willingness to pay for a less congested network we develop and estimate a model of online content consumption. We build on the the model of Nevo et al. (2016) and extend it in two ways. Like Nevo et al. (2016), the consumer makes a series of usage decisions on an optimally chosen plan over a finite horizon. We explicitly incorporate network congestion into the model. Additionally, we disaggregate the daily usage decision into daily peak and off-peak usage decisions. These additions to the model permit a more flexible framework and allow us to study richer counterfactual scenarios.

3.1 Subscriber Utility From Content

Subscribers derive utility from consumption of content and a numeraire good. Each day of a billing cycle, $t = 1....T$, a subscriber chooses the amount of content to consume during peak and off-peak hours on their chosen service plan, $k = 1, ..., K$. Plans are characterized by a provisioned speed content is delivered in the absence of congestion, s_k, by a usage allowance, $\overline{C}_k$, by a fixed fee F_k that pays for all usage up to the allowance, and by an overage price, p_k, per GB of usage in excess of the allowance. The menu of plans, and the characteristics of each, are fixed.[6] The provisioned speed is impacted by the state of the network, ψ, which changes daily due to variation in congestion and periodic network upgrades. We assume this evolution follows a first-order Markov process, G_ψ.

Utility from content is additively separable over all days in the billing cycle, and across billing cycles.[7] Let consumption of content during peak and off-peak hours be denoted by c^p and c^{op}, respectively, and the consumption of the numeraire good is given by y. The utility for a subscriber of type h on plan k is given by

$$u_{hk}(c^p, c^{op}, \psi, \upsilon) = \upsilon_1 \left(\frac{(c^{op} + c^p)^{1-\alpha_h}}{1 - \alpha_h} \right) - (c^{op})^2 \left(\frac{\upsilon_2 \kappa_h}{\ln(s_k)} \right) - (c^p)^2 \left(\frac{\kappa_h}{\ln(\psi s_k)} \right) + y.$$

The first term captures the subscriber's utility from consuming the content. Marginal utility is declining, as we expect the first of any activity (email, web browsing, video, etc.) to bring higher marginal utility than subsequent usage. The convexity of the utility function is also quite flexible, nesting everything between log ($\alpha_h \to 1$) and linear ($\alpha_h = 0$). This leads to a straightforward link between α_h and the price elasticity of demand, such that α_h is the elasticity with respect to the entire cost associated with consuming content. Uncertainty in utility from consumption of content is introduced by a time-varying shock, υ_1, which is realized at the beginning of each day before either consumption decision is made.

The second and third terms of the utility function capture the consumer's cost of consuming content during off-peak and peak hours, respectively, for a consumer of type h. During peak hours the marginal cost of usage is given by $c^p \left(\frac{2\kappa_h}{\ln(\psi s_k)} \right)$ such that increasing Internet usage comes at the cost of alternative activities with greater value. The consumer-type specific parameter, $\kappa_h > 0$, interacts with the plan's provisioned speed, s_k and the state of the network, ψ, to determine the marginal cost of consuming the content, capturing the consumer's preference for speed and the opportunity cost of consuming the

[6] Plans were changed months prior to our sample, but unchanged during our sample, and the ISP had no plans to change them in the months after our sample ends.

[7] Nevo et al. (2016) provide evidence that content is likely not substitutable across days and billing cycles.

content. Importantly, for any finite speed, this specification implies that each subscriber type has a satiation point even in the absence of overage charges. The multiplicative specification with the network state, ψ, and provisioned speed, s_k, captures the proportional rationing of bandwidth used by the ISP when the network is congested. During off-peak hours, the marginal cost of consuming content is $c^{op}\left(\frac{2\upsilon_2\kappa_h}{\ln(s_k)}\right)$. Congestion is a lesser concern so ψ is omitted and provisioned speeds are realized, but consumers may experience different costs due to their opportunity cost of consuming content during off-peak hours relative to peak hours. To capture this, we scale κ_h by a shock υ_2, which is realized before making either daily-usage decision.

This specification of the benefit from consuming content assumes that the value derived from content is similar across the day, i.e., enters the utility function additively, and that differences in the utility derived from content arises on the cost side. For example, the value from watching a movie on Netflix, cost considerations aside, is the same during peak and off-peak hours. This assumption is reasonable for the vast majority of content, particularly OTTV, which currently constitutes two-thirds of usage and continues to grow. There are activities like email or conference calls for which satisfaction from participating in the activity is time dependent, but these represent a very small fraction of overall usage. To flexibly model variation in usage across days, and within days, we assume that υ_1 and υ_2 are independent normal random variables with means equal to μ_{1h} and μ_{2h}, respectively, and a common coefficient of variation, ρ_h.

3.2 Optimal Usage

The vector of parameters, $(\alpha_h, \kappa_h, \mu_{1h}, \mu_{2h}, \rho_h)$, describes a subscriber of type h. A consumer solves a finite-horizon dynamic programming problem within each billing cycle.[8] There are a total of T days in each cycle, and the consumer must make two decisions each day (t), usage during off-peak and peak hours, c_t^{op} and c_t^p, respectively. We assume the consumer observes realizations of υ_{1t} and υ_{2t}, and knows the distribution of potential network states during peak hours, dG_ψ, before choosing c_t^{op} on day t. Further, we assume ψ_t is known (or is costless to discover) when c_t^p is chosen later in the day. Therefore,

[8] The observability of the network state and our focus on the approximately 95% of consumers enrolled on a single plan the entire sample period simplifies the characterization of optimal usage by eliminating inter-billing cycle dependency.

conditional on choosing plan k, this subscriber's problem is

$$\max_{\{c_t^p, c_t^{op}\}_{t=1,\ldots,T}} \quad \sum_{t=1}^{T} E_{(\psi,\upsilon)}\left[u_{hk}(c_t^p, c_t^{op}, \psi, \upsilon)\right]$$

$$s.t. \;\; F_k + p_k \times Max\{C_T - \overline{C}_k, 0\} + Y_T \leq I, \quad C_T = \sum_{j=1}^{T}(c_t^{op} + c_t^p), \quad Y_T = \sum_{j=1}^{T} y_t.$$

We do not discount future utility since we model daily decisions, over a finite and short horizon. Uncertainty involves the realizations of υ_t and ψ_t. We assume that wealth, I, is large enough so that it does not constrain consumption of content.

We solve the consumer's problem recursively. This requires solving a series of intra-day optimization problems nested within a larger non-stationary (due to the finite horizon) inter-day optimization problem. During peak hours on the last day of the billing cycle (T), the consumer solves a static optimization problem, conditional on usage during off-peak hours (c_T^{op}), the state of the network (ψ_T), and preference shocks (υ_T). Depending on the values of c_T^{op}, ψ_T, and υ_{1T}, the consumer will either consume a satiation level of utility such that $\frac{\partial u_{hk}(c_T^{op}, c^p, \psi_T, \upsilon_{\mathbb{T}})}{\partial c^p} = 0$, the remaining portion of their allowance such that $\overline{C}_{kT} = 0$, or incur overages such that $\frac{\partial u_{hk}(c_T^{op}, c^p, \psi_T, \upsilon_{\mathbb{T}})}{\partial c^p} = p_k$. Denote this optimal level of consumption, or the policy function on day T during peak hours, as $c_{hkT}^p(c_T^{op}, C_{T-1}, \psi_T, \upsilon_T)$ or more compactly, c_{hkT}^{p*}.

Given the optimal policy during peak hours on day T, the optimal policy for off-peak usage, $c_{hkT}^{op}(C_{T-1}, \psi_{T-1}, \upsilon_T)$ or c_{hkT}^{op*}, satisfies

$$c_{hkT}^{op*} = \underset{c^{op}}{argmax} \int\limits_{\psi} \Bigg[\upsilon_{1T} \frac{\left(c^{op} + c_{hkT}^{p*}\right)^{1-\alpha_h}}{1-\alpha_h} - (c^{op})^2 \left(\frac{\upsilon_{2T}\kappa_h}{\ln(s_k)}\right)$$
$$- (c_{hkT}^{p*})^2 \left(\frac{\kappa_h}{\ln(\psi s_k)}\right) - p_k \mathcal{O}_{tk}(c^{op} + c_{hkT}^{p*}) \Bigg] dG_{\psi}(\psi|\psi_{T-1}),$$

where the expectation is only over ψ (which impacts the current speed and the optimal peak-usage policy) because υ_{1T} and υ_{2T} are known when c_{hkT}^{op} is chosen. The expected value from following the optimal policies during off-peak and peak hours on day T conditional on entering that day at state, (C_{T-1}, ψ_{T-1}), equals

$$E_{(\psi,\upsilon)}\left[V_{hkT}(C_{T-1},\psi_{T-1})\right] = \int\limits_{\psi}\left[\int\limits_{\upsilon} V_{hkT}(C_{T-1},\psi_{T-1}\psi,\upsilon)dG_{\upsilon}^{h}(\upsilon)\right]dG_{\psi}(\psi|\psi_{T-1}),$$

where $V_{hkT}(C_{T-1},\psi_{T-1},\psi,\upsilon)$ is the value associated with following the optimal policies for a particular realization of the network state (ψ) and preference shocks (υ).

Optimal policies are defined similarly for any day $t < T$. The optimal peak-usage policy on day t, $c_{hkt}^{p}(c^{op},C_{t-1},\psi_t,\upsilon_t)$ or c_{hkt}^{p*}, satisfies

$$c_{hkt}^{p*} = \underset{c^p}{argmax}\left[\upsilon_{1t}\frac{\left(c_{hkt}^{op*}+c^p\right)^{1-\alpha_h}}{1-\alpha_h} - (c^p)^2\left(\frac{\kappa_h}{\ln(\psi_t s_k)}\right)\right.$$
$$\left. + \beta E_{(\psi,\upsilon)}\left[V_{hkt}(C_{t-1}+c_{hkt}^{op*}+c^p,\psi_{t-1})\right]\right].$$

Similarly, the optimal policy for off-peak hours on day $t < T$ is

$$c_{hkt}^{op*} = \underset{c^{op}}{argmax}\int\limits_{\psi}\left[\upsilon_{1t}\frac{\left(c^{op}+c_{hkt}^{p*}\right)^{1-\alpha_h}}{1-\alpha_h} - (c^{op})^2\left(\frac{\upsilon_{2t}\kappa_h}{\ln(s_k)}\right) - (c_{hkt}^{p*})^2\left(\frac{\kappa_h}{\ln(\psi s_k)}\right)\right.$$
$$\left. + \beta\int\limits_{\upsilon} V_{hk(t+1)}(C_{t-1}+c^{op}+c_{hkt}^{p*},\psi,\upsilon)dG_{\upsilon}^{h}(\upsilon)\right]dG_{\psi}(\psi|\psi_{t-1}).$$

These state-dependent policy functions are stored along with the value functions when the model is solved for each type, h, on every plan, k. This permits a comparison of usage and utility for each type to identify that type's optimal plan. Our econometric approach discussed in Section 4 only requires solving the model once for each type.

3.3 Plan Choice

We assume consumers select plans to maximize expected utility, before observing any utility shocks, and remain on that plan during our sample. More precisely, we assume that the subscriber selects one of the offered plans, $k \in \{1,...,K\}$, or no plan, $k = 0$, such that

$$k_h^* = \underset{k\in\{0,1,...,K\}}{argmax}\left\{\int\limits_{\psi_1}\Big(E_{(\psi,\upsilon)}\left[V_{hk1}(C_1=0,\psi_1)\right]\Big)\pi_{\psi_1} - F_k\right\}.$$

The optimal plan, k_h^*, maximizes expected utility for the subscriber, net of the plan's fixed fee (F_k), where the expectation is taken over initial states, ψ_1, and the probability weights, π_{ψ_1}, are equal to the stationary distribution of the markov process for the network state, dG_ψ. The outside option is normalized to have a utility of zero. Note, that we assume that there is no error, so consumers choose the plan that is optimal. Similar to Nevo et al. (2016), (potentially weak) tests of optimal plan choice reveal that it is extremely rare to observe a subscriber whose usage decisions are such that switching to an alternative plan would yield lower total costs at no slower speeds. The weakness of this optimality test is due to the positive correlation between speed and usage allowances of the offered plans (see Figure 1). Our assumptions on plan choice are easily relaxed in theory, but introduce a substantial additional computational burden. Given the infrequency of both clear ex-post mistakes in choosing a plan and switching of plans in our sample, there is little evidence to infer the assumption is incorrect. This optimality assumption results in a one-to-one correspondence between plans (k) and consumer types (h). The usage policy functions for a consumer type (h) on the optimal plan, $c_{hk_h^*t}^{op*}$ and $c_{hk_h^*t}^{p*}$, serve as the basis for our fixed-grid maximum-likelihood estimation procedure.

4 Estimation

The goal of the estimation is to recover the distribution of types. There are two ways to view this distribution. The first is to assume that each household belongs exactly to one type and with enough data that type will be revealed. We then aggregate households to estimate the distribution of types in the population. Alternatively, we could assume that each household is itself a mixture of types, either because the household consists of several members or because a given individual's behavior is best described as a mixture of types. In this case even with a very large data the distribution of types for each household will not be degenerate. Since ultimately the aggregate distribution of types is of interest, we take a flexible approach and do not try to distinguish between these two views.

Our estimation approach is a panel-data modification of Ackerberg (2010), Fox et al. (2016), and Nevo et al. (2016). Like the previous literature, we solve the problem (once) for a large number of types, where a types is defined by a vector of the parameters. We then estimate the distribution across types by matching patterns from the data to those predicted by the behavior of the types. Where we differ from previous papers is that we explicitly take account of the panel structure of the data.

Specifically, we proceed in two steps. First, we solve the model (i.e., the dynamic programming problem) for a large number of candidate consumer types (h), each characterized by a vector of parameters, $(\alpha_h, \kappa_h, \mu_{1h}, \mu_{2h}, \rho_h)$. From the solution, we store the policy functions on each candidate type's optimal plan, which give peak and off-peak usage at each state. Second, for each consumer on a particular plan (k) in our data, we calculate the likelihood that the observed sequence of peak and off-peak usage decisions is generated by each of the H_k candidate types that optimally choose that plan. The derivation of the likelihood comes from the density of usage decisions conditional on the observed states, (t, C, ψ), that arises due to variation in the (stochastic) unobserved states, (υ_1, υ_2). The relative likelihood values over the H_k candidate types for each consumer is then taken as a posterior refinement of the consumer's type, from a uniform prior over the candidate types that optimally chose the same plan as the consumer. We then calculate an estimate of the density of consumer types in the population by aggregating candidate type weights across consumers on each plan, and then across plans while accounting for each plan's market share. We discuss the details of the approach below.

4.1 Estimation

For each individual on plan k, $i = 1....I_k$, our data includes a series of data, $m = 1.....M$, which captures usage at a daily frequency on the chosen plan.This includes both off-peak usage, $(c^{op}_{i1}, ..., c^{op}_{iM})$, and peak usage, $(c^{p}_{i1}, ..., c^{p}_{iM})$. In addition, for date m, we observe the consumer's state: days into the current billing cycle (t_m), cumulative usage up until day t_m of the billing cycle (C_{t_m-1}), and the network state entering the day (ψ_{t_m-1}) and during peak hours during that day (ψ_{t_m-1}).

We solve the model for 7,776 (6^5) candidate types of consumers, corresponding to a fixed grid with six points of support for each of the five parameters, $(\alpha_h, \kappa_h, \upsilon_{1h}, \upsilon_{2h}, \rho_h)$. The solution to the model yields state-dependent policy functions for peak and off-peak usage on each type's optimal plan, $c^*_{hkt_{op}}(C_{t-1}, \psi_{t-1}, \upsilon_t)$ and $c^*_{hkt_p}(c_{t_{op}}, C_{t-1}, \psi_t, \upsilon_t)$, respectively. Additionally, the parametric structure of the model (distribution of υ), yields probabilities associated with any realization of the policy functions, $\mathbb{P}\left[c^{op}_{hkt}(C_{t-1}, \psi_{t-1}, \upsilon_t) = c\right]$ and $\mathbb{P}\left[c^{p*}_{hkt}(c_{t_{op}}, C_{t-1}, \psi_t, \upsilon_t) = c\right]$, conditional on the observed states. These probabilities, together with the Markov process for ψ (dG_ψ) recovered from the data, are the foundation for constructing the likelihood that each candidate type (h) generates a particular sequence of usage observed in the data.

Specifically, consider all candidate types that optimally choose plan k, $h = 1, ..., H_k$. Further, consider one consumer's sequence of usage on plan k, $(c^{op}_{i1}, ..., c^{op}_{iM})$, and $(c^{p}_{i1}, ..., c^{p}_{iM})$, and the values of the associated observed states (t_m, ψ_m, C_{t_m}). The likelihood that candidate type h generated this sequence of data, conditional on the individuals observed states, equals

$$\mathscr{L}_{ih} = \prod_{m=1}^{M} \mathbb{P}\Big[c^{op}_{hkt}(C_{m-1}, \psi_{m-1}, \upsilon_m) = c^{op}_{ikm}\Big] \times \mathbb{P}\Big[c^{p}_{hkt}(c^{op}_{m}, C_{m-1}, \psi_m, \upsilon_m) = c^{p}_{ikm}\Big] \times dG_{\psi}(\psi_m|\psi_{m-1}).$$

The likelihood for individual i is calculated for all candidate types that optimally choose plan k, the individual's optimal plan, yielding $\mathscr{L}_{i1}, ..., \mathscr{L}_{iH_k}$.

Using Baye's rule, the posterior probability that individual i is of type h is then straightforward to calculate. Specifically, the selection of plan k by individual i reveals only that it belongs to the set of H_k candidate types that optimally choose plan k. For this reason, we assume a uniform prior over the H_k candidate types, which yields a posterior probability proportional to the likelihood,

$$\mathbb{P}_{ih} = \frac{\mathscr{L}_{ih}}{\sum\limits_{j=1}^{H_k} \mathscr{L}_{ij}}$$

To form estimates of the posterior probabilities, $\widehat{\mathbb{P}}_{ih}$, we calculate probabilities from the solution to the model via simulation over υ, i.e., $\widetilde{\mathbb{P}}\Big[c^{op}_{hkt}(C_{m-1}, \psi_{m-1}, \upsilon_m) = c^{op}_{ikm}\Big]$ and $\widetilde{\mathbb{P}}\Big[c^{p}_{hkt}(c^{op}_{m}, C_{m-1}, \psi_m, \upsilon_m) = c^{p}_{ikm}\Big]$, respectively, and use the same estimate of the Markov process for the network state, $\widehat{dG_{\psi}}\,(\psi_m|\psi_{m-1})$, that is used to solve the model.

The estimate of the density for any candidate type h among the population of consumers then equals

$$\widehat{\mathbb{P}}_h = \mathbb{S}_k \times \left(\frac{1}{I_k}\sum_{i=1}^{I_k} \widehat{\mathbb{P}}_{ih}\right).$$

where S_k is the proportion of individuals selecting plan k. The summation is only over $i = 1, .., I_k$ because each candidate type receives positive density on (at most) its optimal plan. Standard errors on the distribution of types in the population are calculated via a block-resampling procedure. Specifically, a block of dependent data in our application is characterized by the entirety of each individual's usage series. We repeatedly $(r = 1, ..., R)$ perform the estimation using samples created through sampling with replacement, calculating the standard errors directly from the distribution of estimates.

4.2 Identification

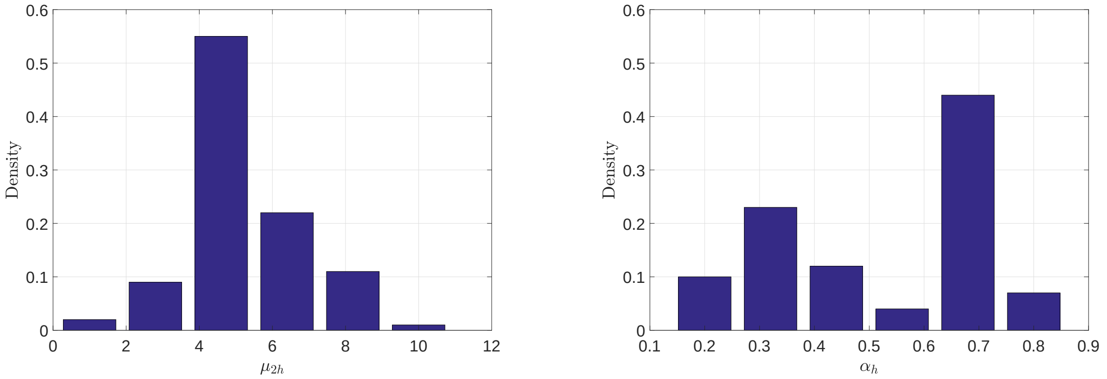
Our approach to estimation makes explicit the two main sources of identifying variation in our data: consumer's plan choice and subsequent usage on these plans. We rely on both, along with the structure of the model, to identify the distribution of tastes in the population. However, in contrast to Nevo et al. (2016), we rely on the richness of our panel to exploit within-consumer variation to identify this distribution.

Recall, that our objective is to estimate the distribution of types. Our choice of the grid is equivalent to placing a uniform distribution over types on this grid and zero elsewhere. An individual's choice of plan, along with the assumption that this choice is optimal, refines this distribution since it reveals the subset of candidate types to which this consumer belongs. That is, among the support we consider for the vector describing a type, $(\alpha_h, \kappa_h, \mu_{1h}, \mu_{2h}, \rho_h)$, plan choice assigns each type to a subset of this space corresponding to its optimal plan (or the outside option of no plan). However, like Nevo et al (2016) demonstrate, this source of variation is relatively weak without temporal or cross-sectional variation in the offered plans that would identify the distribution of tastes in the population. Yet, the segmentation of the types by plans is not entirely uninformative and does reveal some information about taste parameters. For example, types with a preference for large allowances and/or high speeds, and an accompanying willingness to pay for these features, self select into more-expensive tiers. Thus, the subset of candidate types assigned to each plan are not homogeneous across plans, creating at least some weak association between a consumer's type and their plan choice.

A more informative source of variation is a consumer's usage conditional on plan choice, and in particular the responsiveness of usage to congestion and the possibility of overage fees. For a given consumer, the plan choice places a uniform distribution over the subset of candidate types that would optimally select his/her chosen plan. To refine this distribution, we rely on high-frequency variation in usage, and the covariation between usage and the observable states of the model, (t, C_t, ψ_t), which determine the tradeoffs faced by the consumer.

Optimal usage is revealing about the individual's preferences. For example, fixing the shadow price of usage at zero, the responsiveness of c_t^p $(c \to c')$ to variation in the the network state $(\psi \to \tilde{\psi})$ identifies the marginal cost of usage, κ_h. Similarly, variation in the shadow price due to cumulative usage, C_t, reveals the curvature of the marginal benefit function determined by α_h. Note a positive shadow price of usage shifts upward the total marginal cost of usage, and if C_t is below the $\overline{C}_k$, it will also increases its slope. For a fixed network state (ψ), the mean and variance of c^p identify the parameters of the υ_1 distribution. Finally, the distribution of υ_2 that determines the relative costs of the marginal cost of usage at different times of the day is identified by the relative intensity of peak and off-peak usage.

Figure 11: *Marginal Distribution of μ_2 and α*



(a) Marginal of μ_2

(b) Marginal of α

Note: The figures are the estimated marginal distributions for μ_2 and α, respectively. Within each panel, the sum of all five bars in each figure total 732, the total number of types.

5 Results

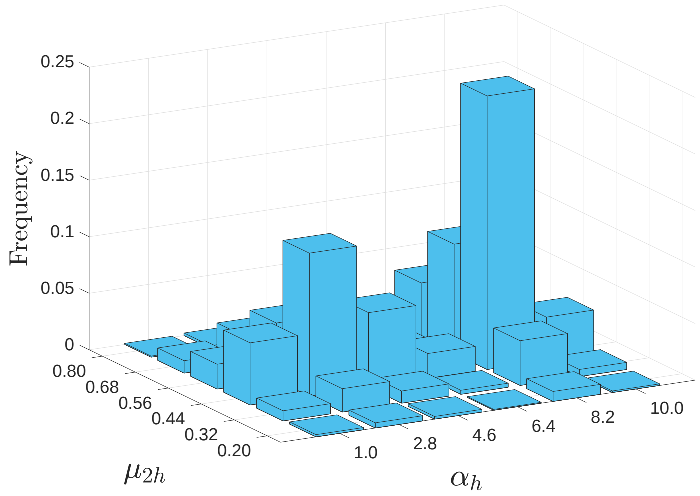
We present our estimation results in three parts. First, we summarize our estimates of the type distribution. Next, we discuss the implications of our estimates of preferences for measuring the value of investing and important policy issues. Finally, we discuss counterfactual exercises that explore the value of economic and technological solutions to reducing congestion.

5.1 Type Distribution Estimates

We estimate a weight greater than 0.01% for 732 types. That is, 732 different types (h) were chosen from the 7,776 candidate types considered. On each plan, the mass is quite concentrated with the most common representing over 22% of the plan's total mass, and the top two types over 30%. Most of the types with positive weight, 298, come from the most-expensive plan, despite it only representing about 2.5% of the sample. This is intuitive since that plan has the most heterogeneity, types selecting the plan for its high speed and its usage allowance, or both.

In Figure 11, we present the marginal distributions for two of the parameters to demonstrate the flexibility of the econometric approach. Figure 11 (a) and (b) give the marginal distribution for μ_2 and α, respectively. The vast majority of the mass for both distributions (and the other three parameters) is not on the boundary of the chosen support for the parameters, and we find that increasing the range of the support results

Figure 12: *Joint distribution of μ_{2h} and α_h*



Note: Joint histogram of μ_{2h} and α_h given by frequency counts.

in virtually no redistribution of mass outside the chosen support.[9] Thus, the range of the support chosen for the discrete distribution of candidate types does not constrain optimization over the type weights. The μ_2 parameter that determines the intra-day allocation of usage, along with the network state, has an intuitive shape and distribution. In our data, most consumers have peak usage that averages approximately five times off-peak usage. The mean of μ_2 is about five and the vast majority of types are tightly distributed around that mean, while the heterogeneity in temporal usage patterns is also captured with some types with $\mu_2 = 1$ (flat profile on average) receiving positive weight. The marginal distribution of α is perhaps the most interesting with a clearly bimodal shape. The two modes represent consumer types with substantially different elasticities for content, which has important implications for each of the policy issues and counterfactual scenarios we discuss below.

As would be expected given the quite different marginal distributions, the joint distributions are quite irregular. Figure 12 gives the joint distribution of μ_{2h} and α_h. The multi-peaked nature of the α_h distribution is still clearly visible, but there are non-negligible correlations between the two parameters. This demonstrates the importance of the flexibility of our estimation approach, which allows for free correlations between each pair of parameters rather than the normality and lack of covariance often assumed

[9] For some of the parameters, e.g., μ_2, the support is naturally bounded so it is only necessary to enrich the support in one direction.

in structural econometric applications.

5.2 Policy Implications

Directly from the estimates, we calculate the distribution of willingness to pay for connectivity speed and usage allowances. For connectivity speed, we calculate the value associated with increasing the provisioned speed of each type's optimal plan by 1 Mb/s, conditional on subscribers realizing those speeds (i.e., $\psi = 1$). We simply resolve the model for each type, in the absence of congestion, after increasing the speed of their optimal plan by a fixed amount and compare it to the currently-offered plans. This approximates the change in expected utility at the beginning of a billing cycle with respect to speed (s_k),

$$\frac{\partial E_{(\psi,\upsilon)}\left[V_{hk1}(C_1 = 0, \psi = 1)\right]}{\partial s_{h_{k^*}}}.$$

on each type's optimal plan, h_{k^*}. We find that a 1 Mb/s increase in speed on each type's optimal plan is valued at and average of \$0.64. Yet, this value drops off fairly quickly, as a 10 Mb/s increase is valued at less than \$3.50.

The fairly low preference for increased connectivity speed is interesting for a number of reasons. The definition of broadband service now requires a speed of at least 25 Mb/s, which is above the average speed of customers in our data. Thus, the FCC is well ahead of the average consumer's preferences with respect to what is needed for most commonly used applications. Additionally, regulatory authorities have begun to make approval of mergers and acquisitions conditional on substantial investment in networks. A recent example is Altice's acquisition of Cablevision, which was approved by the New York State Public Service Commission conditional on Altice making \$243 million of investment to increase broadband speeds up to 300 Mb/s by 2017, along with expanding its network at a cost of \$40 million to subsidize access for underserved areas.[10]

Our results suggest that these required network investments are likely to have quite mixed returns. In particular, it is important to distinguish between the speed and capacity of a network, as a network capable of delivering fast speeds but without adequate capacity is quite undesirable (as our descriptive analysis demonstrates). Therefore, we expect the subsidies to cover the costs of access to be quite valuable as access at even modest speeds is highly valuable, but investment to dramatically increase overall speeds would likely not do much to penetration and have quite low value to existing customers. For example, an HD movie from most OTTV services requires approximately 5 Mb/s, and recent encod-

[10] This decision can be found in the Commission Documents section of the Commissions web site at www.dps.ny.gov and entering Case Number 15-M-0647 in the input box labeled "Search for Case/Matter Number".

ing and compression technologies continue to reduce requirements, which suggests that a home with 100 Mb/s (speed currently offered by Cablevision) could already stream 20 movies simultaneously. Thus, a more efficient allocation of these resources might be to ensure that consumers are consistently achieving provisioned speeds by increasing capacity, reducing the frequency of congestion, rather than pushing the top-end capabilities of the networks.

Similarly to preferences for speed, our model yields an estimate of the value of one more GB added to the usage allowance on each type's optimal plan. This value equals

$$\frac{\partial E_{(\psi,\upsilon)}\left[V_{hk1}(C_1=0,\psi=1)\right]}{\partial \overline{C}_{h_{k^*}}},$$

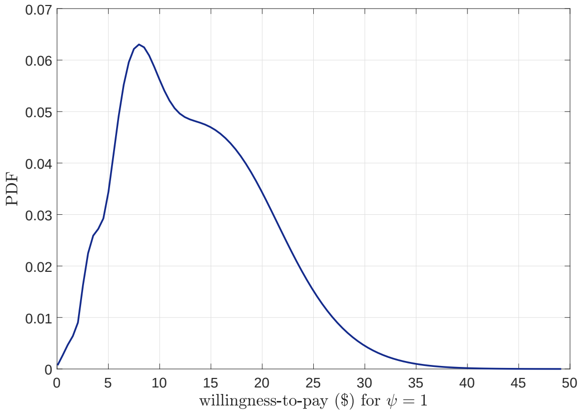
which is the change in expected utility at the beginning of the billing cycle from one more GB of usage allowance, in the absence of congestion. We find the median consumer values an additional GB at only \$0.07, while the average is \$0.11 due to the extreme right tail of users. The relatively low value, particularly as compared to Nevo et al (2016), suggests that current usage allowances are well ahead of most consumer's preferences. Additionally, the highly skewed nature of this valuation suggests that usage-based pricing (as implemented by this ISP) is impacting only the heaviest of broadband users.[11]

Thus, if the median consumer were to watch an additional typical HD movie each billing cycle, approximately 2 GBs, the value to the consumer in expectation would be about \$0.22. This is quite small, because even if you substantially decrease the usage allowances, the marginal value doesn't increase too rapidly since most users are far from the allowance in a typical month. Thus, the inframarginal cost of most of the bytes generated by OTTV services is a tiny fraction of the marginal cost. This estimate is particularly useful for the recent policy discussion around zero-rating policies by ISPs, i.e., the policy of not counting usage from the ISP's streaming service against usage allowances while other OTTV services are. Our estimates suggest that the price differential introduced by these policies is quite small for almost all users. Or, at the current ratio of allowances to usage, zero-rating policies by ISPs have very little impact on disadvantaging competing OTTV services.

Both of the calculations above, willingness-to-pay for greater speed and larger usage allowances, are done so purposely with congestion absent. That is, it is important to understand willingness-to-pay for plan features in its absence. However, the volatility and rate of growth in demand makes investment difficult to target and time, making con-

[11] The plan features of the usage-based pricing by this ISP are similar to those implemented by other North American ISPs, and so these estimates are not much different for the industry as a whole.

Figure 13: *CDF of willngness-to-pay (\$) for* $\psi = 1$



Note: This figure presents the pdf of the willingness-to-pay to eliminate congestion from the network for the entire billing cycle in our population of subscribers.

gestion an ever-present feature of many networks. Thus, it is also of interest to perform such similar calculations in the presence of congestion, as the results may more accurately reflect the plight of some consumers. Figure 13 presents the pdf of the willingness-to-pay to eliminate congestion from the network for the entire billing cycle. This is the additional surplus generated by purchasing a "fastlane" that is not subject to congestion (i.e., realized speeds always equal maximum advertised speeds), in contrast to the conditions currently observed on the network. We find the average (median) consumer will pay approximately \$14.77 (\$12.94), which is substantially more than increasing advertised speeds in the absence of congestion. This further reinforces our findings that delivering reliable service, even at modest speeds, delivers substantial value relative to costly investments to offer speeds that very few consumers value.

5.3 Economic and Technological Solutions to Congestion

Constant investment in broadband networks has been the industry's primary response to keep up with Internet usage, which has grown recently in excess of 40% annually.[12] However, as our descriptive analysis in Section 2 demonstrates, networks are built for peak demand and are vastly under-utilized the majority of the day. This suggests that economic and technological solutions can serve a complementary role to maximize the value of these investments. We explore three such potential solutions: throttling of connectivity speeds, peak-use pricing, and local-cache technologies.

[12] The primary economic solution employed by ISPs in recent years has been the implementation of simple three-part tariffs to discourage low-value usage. However, Malone et al. (2014) show these tariffs discourage usage at all times, including off-peak hours when usage is nearly costless.

Ideally for the counterfactual computation below we would resolve the dynamic problem for each consumer faced with the new pricing, throttled speed or ability to locally cache. We would then compute the congestion and resolve the consumer problem, which would imply a new congestion, and so forth. This involves iterating between a dynamic problem and an equilibrium computation of congestion, which computationally is not feasible. For the calculations below we focus on the equilibrium computation of congestion and simplify the dynamics when solving the counterfactuals.[13] Specifically, for the computation below compare usage and improvement in consumer surplus under each of these scenarios to a baseline simulation from the model where we simplify the subscriber's problem to the decision of how to allocate usage within a single day. This eliminates inter-day dynamics that arise due to usage-based pricing, much like the decision regarding usage on the last day of the each billing cycle. Specifically, for each subscriber, optimal usage in the baseline specification solves

$$\begin{aligned} 0 &= \upsilon_1 \left(c^{op} + c^{p}\right)^{-\alpha} - \frac{2c^{op}\upsilon_2\kappa_h}{\ln(s_k)} \\ 0 &= \upsilon_1 \left(c^{op} + c^{p}\right)^{-\alpha} - \frac{2c^{p}\kappa_h \ln(\psi s_k) - (c^{p})^2\kappa_h \frac{1}{\psi}\frac{\partial \psi}{\partial c^{p}}}{\ln^2(\psi s_k)} \end{aligned}$$

given a particular realization of υ. Note that each subscriber accounts for the effect of their usage during the peak period on the network state, ψ. To calculate average usage and consumer surplus, we simulate different realizations of υ for all subscribers, resolve the system of equations, and calculate averages from these solutions. Each of the counterfactuals then involve altering these first-order conditions to account for differential prices introduced by each of the alternatives to resolve congestion.

Throttling Speed

One way to deal with congestion is to slow down during peak times the consumers who have exceeded their allowance. This is the way that allowance is implemented on some cellular networks, such as T-Mobile. To simulate the effect of such a policy, we slow consumers down during peak usage periods after exceeding their usage allowance rather than incur overage fees. We assume the throttled speed is 7 Mb/s, the upper limit of what is required to stream from most OTTV services in HD. The results are in Table 4. The effect is a bit counterintuitive, as total usage during peak and off-peak hours increases. However, the majority of consumers have a low valuation on speed so they much rather

[13] In an earlier version of this paper we went to the other extreme, where we resolved the new dynamic problems but held congestion fixed rather than solved for the equilibrium value. Interestingly, qualitatively the direction of the counterfactual results was the same.

Table 4: *Counterfactual: Throttling*

		Throttling	
Usage and Surplus	*Baseline*	*7 Mb/s* (s_k)	*3 Mb/s* (s_k)
Daily Usage (GB)	2.6	2.9	2.8
Peak Usage (GB)	1.9	2.0	1.9
Off-Peak Usage (GB)	0.7	0.9	0.9
Consumer Surplus	68.31	72.40	73.02
Revenue	59.75	57.09	58.18

prefer the penalty of a low speed than having to pay the per GB charge. There additional usage improves consumer welfare, while there is only a slight decrease in ISP revenue due to the absence of overage charges from those that opted into throttling.

Peak-Use Pricing

We consider a simple form of peak-use pricing where off-peak usage is not counted against the allowance, while peak usage is counted fully, and the allowance is decreased by a given amount. The results are in Table 5. The first column provides a baseline where congestion is absent from the network, while the second and third columns consider a 25% and 50% reduction in the baseline allowance when only peak usage is counted. Interestingly, and consistent with the results from the analysis of node splits, we find that while peak usage responds to a higher price whether it be in the form of overages or congestion, off-peak usage is largely unchanged. That is, the intra-day elasticity of usage is quite small. Thus, simply raising the price of peak usage is not a particularly attractive alternative, and it simply results in a small transfer from consumers to the ISP. However, we find that the WTP for an extra GB is almost three times higher once the allowance is reduced by 50%. Even though the low elasticity makes the usage response modest, the reduction in the allowance and peak use pricing provides strong incentive to content providers to make the traffic associated with their applications transferrable to off-peak hours. One way to do that is through introduction of local-caching technologies.

Local-Cache Technology

OTTV services accounts for a disproportionate share of peak-use traffic. Yet, it is a passive activity for which arrival of the content and actual consumption or viewing of the content need not coincide temporally, unlike web browsing. This presents an opportunity to potentially shift the timing of the downloads to off-peak hours when the network is under-utilized, while the viewing still occurs during peak hours. One approach

Table 5: *Counterfactual: Peak-Use Pricing*

	Peak-Use Pricing		
Usage and Surplus	*Baseline*	*25% Reduction* ($\overline{C}_k$)	*50% Reduction* ($\overline{C}_k$)
Daily Usage (GB)	2.6	2.7	2.4
Peak Usage (GB)	1.9	1.8	1.7
Off-Peak Usage (GB)	0.7	0.9	0.7
Consumer Surplus	68.31	69.06	65.24
Revenue	59.75	58.11	60.22

Table 6: *Counterfactual: Local-Caching Technology*

	Local-Caching Technology		
Usage and Surplus	*Baseline*	*25% Reduction* (μ_2)	*50% Reduction* (μ_2)
Daily Usage (GB)	2.6	2.8	3.2
Peak Usage (GB)	1.9	1.8	1.7
Off-Peak Usage (GB)	0.7	1.0	1.3
Consumer Surplus ($)	68.31	75.88	81.70
Revenue ($)	59.75	60.97	63.01

to detach the two activities is local caching of content. The ASAP feature on Amazon's Fire streaming device predictively caches content using machine-learning algorithms that exploit past-viewing behaviors. The ASAP feature only loads a small portion of each piece of content to improve quality, faster startup and higher resolution, but the technology can be easily adapted to give the user more control over how much of each piece of content to cache. Envision an application for OTTV services that is analogous to Tivo's functionality for traditional pay TV. Therefore, such solutions can be implemented at very low cost.

Such a technology would be expected to decrease the effort associated with downloading OTTV during off-peak hours (envision a phone app for Netflix that allows the user choose what and when to cache), but not change the utility from consuming it since that can still be done during peak hours. In our model, this is similar to decreasing the mean of the shock to the cost of off-peak usage λ_2. In Table 6, we provide estimates of the effect, relative to the baseline with no congestion, if local-cache technologies were to reduce the cost of off-peak usage by 50% for all consumer types. This of course may substantially understate its effect, as the heaviest of users consume substantially more OTTV services and would benefit substantially from such technologies. Despite this potential downward bias, we find that consumers benefit substantially from such technology, but only increase peak-use slightly, a cost the ISP could surely recapture through re-optimization of prices.

6 Conclusion

We estimate demand for residential broadband using an 11-month panel of hourly subscriber usage and network conditions. The key feature of our model of demand is the inclusion of network congestion as a determinant of subscriber's inter- and intra-day usage decisions. We estimate the model by extending the methods proposed by Ackerberg (2010), Fox et al. (2011), and Fox et al. (2016), and applied by Nevo et al. (2016). We find that the flexibility these methods provide in estimating heterogeneity is important to explain the data. Our findings show a low willingness-to-pay for larger usage allowances and faster advertised speeds, indicating diminishing returns to further investment aimed at delivering gigabit and beyond speeds to consumers. Yet, we find consumers are willing to pay, on average, just under $15 to eliminate congestion.

Next, we use the model estimates to explore several economic solutions to eliminate congestion and more efficiently use network capacity throughout the day. We find that throttling of speeds when usage exceeds a consumer's allowance actually increases both total and peak usage, because many consumers view throttling as less costly than incurring overage fees. Thus, strategies employed on wireless networks are likely to be unsuccessful on residential broadband networks. Our results suggest that peak-use pricing has potential, but (in isolation) is ineffective at shifting demand towards non-peak periods because consumers have little ability to shift usage within a day. We then introduce local-cache technology and find that it will be effective in reducing demand during peak-use periods by allowing consumers to divert demand to off-peak periods. We envision this diversion being mostly downloads of video content during off-peak periods that will be viewed during peak periods. Together, our estimates suggest that the combination of peak-use pricing and local cache technology should be particularly effective in reducing congestion.

The ability to divert downloads to off-peak periods is likely to make OTTV an even more attractive option, encouraging further adoption possibly at the expense of traditional pay-tv services sold by the ISP. This, together with the cost asymmetry between OTTV services, which place substantial demands on broadband networks, and pay-tv services, for which the costs on the network are fixed, creates complex incentives for the ISP. An interesting area for future research, that can directly contribute to ongoing policy debates, is to jointly estimate demand for video and broadband services, and then use these estimates to explore incentives for ISPs to foreclose OTTV providers.

References

Dutz, Mark, Jonathan Orszag and Robert Willig (2009). "The Substantial Consumer Benefits of Broadband Connectivity for US Households." *Internet Intervention Alliance Working Paper.*

Edell, Richard and Pravin Varaiya (2002). *Providing Internet Access: What We Learn from INDEX*, volume Broadband: Should We Regulate High-Speed Internet Access? Brookings Institution.

Goolsbee, Austan and Peter Klenow (2006). "Valuing Products by the Time Spent Using Them: An Application to the Internet." *American Economic Review P&P*, 96(2): 108–113.

Greenstein, Shane and Ryan McDevitt (2011). "The Broadband Bonus: Estimating Broadband Internet's Economic Value." *Telecommunications Policy*, 35(7): 617–632.

Hitte, Loran and Prasanna Tambe (2007). "Broadband Adoption and Content Consumption." *Information Economics and Policy*, 74(6): 1637–1673.

Lambrecht, Anja, Katja Seim and Bernd Skiera (2007). "Does Uncertainty Matter? Consumer Behavior Under Three-Part Tariffs." *Marketing Science*, 26(5): 698–710.

Malone, Jacob, Aviv Nevo and Jonathan Williams (2016). "A Snapshot of the Current State of Residential Broadband Networks." *NET Institute Working Paper No. 15-06.*

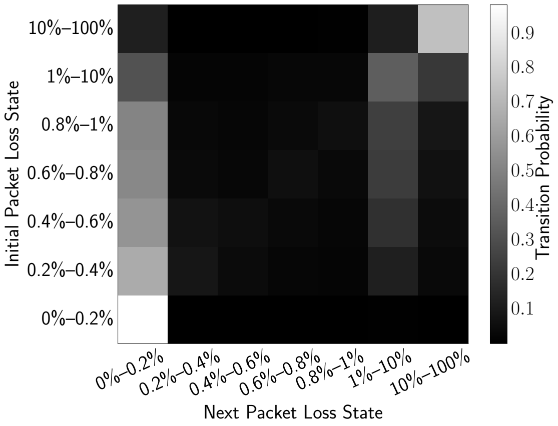
Malone, Jacob, John Turner and Jonathan Williams (2014). "Do Three-Part Tariffs Improve Efficiency in Residential Broadband Networks?" *Telecommunications Policy*, 38(11): 1035–1045.

Nevo, Aviv, John Turner and Jonathan Williams (2016). "Usage-Based Pricing and Demand for Residential Broadband." *Econometrica*, 84(2): 411–443.

Rosston, Gregory, Scott Savage and Bradley Wimmer (2013). "Effect of Network Unbundling on Retail Price: Evidence from the Telecommunications Act of 1996." *Journal of Law and Economics*, 56(2): 487–519.

Varian, Hal (2002). *The Demand for Bandwidth: Evidence from the INDEX Experiment*, volume Broadband: Should We Regulate High-Speed Internet Access? Brookings Institution.

Figure 14: *Heatmap of Peak Packet Loss Transitions*



Note:

Appendix

Descriptive Statistics

In Table 7 and Figure 14, packet loss is split into seven bins that are used to study how persistent packet loss is day-to-day; the values in the heat map are the same as in the table. From these transition probabilities, there are a couple of notable takeaways. First, if a subscriber's peak packet loss is poor one day, there is a high probability it will be better the next day. Second, if a subscriber does end up in the worst packet loss state, they are most likely to be in a poor state the next day. Third, the vast majority of subscribers experience low packet loss and will experience low packet loss tomorrow.

For the model, we use the transition matrix in Table 7 to estimate the frequencies of transition between packet loss, or network congestion, states. Below in the model discussion, this will be G_{ψ}. This matrix will be used to solve the model. For the estimation procedure, all we need are *day-hour* observations of daily consumption and the observed peak packet loss state for each account in the sample.

Results

The marginal distribution for each of the parameters is summarized in 8.

Table 7: *Transition Matrix of Peak Packet Loss*

	Next State						
Initial State	*0–0.2*	*0.2–0.4*	*0.4–0.6*	*0.6–0.8*	*0.8–1*	*1–10*	*10–100*
0–0.2	0.984	0.002	0.001	0.001	0.001	0.006	0.004
0.2–0.4	0.662	0.086	0.044	0.027	0.021	0.124	0.037
0.4–0.6	0.570	0.074	0.055	0.037	0.027	0.186	0.051
0.6–0.8	0.526	0.041	0.031	0.062	0.039	0.235	0.066
0.8–1	0.511	0.032	0.026	0.042	0.059	0.244	0.087
1–10	0.316	0.023	0.020	0.029	0.029	0.364	0.218
10–100	0.122	0.004	0.003	0.005	0.005	0.119	0.741

Note: This table reports probabilities of *peak hour-day* packet loss transitions at the subscriber level of observation. Each bin is of the form $(x\%, y\%]$ and represent a range of packet loss. The first bin includes 0% packet loss, too.

Table 8: *Descriptive Statistics for Types*

	Mean	*Median*	*Mode*
α_h	0.52	0.56	0.68
κ_h	3.41	4.03	4.03
μ_{1h}	0.93	0.80	0.80
μ_{2h}	5.21	4.60	4.60
ρ_h	0.37	0.48	0.48

Note: This table reports descriptive statistics of the type distribution: mean, median, and mode.