The Tragedy of the Last Mile: Congestion Externalities in Broadband Networks[*]

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

Preliminary and Incomplete

Abstract

Internet usage, mostly driven by increased demand for video, has been increasing at a fast rate. This growth has substantially constrained broadband networks, slowing down service, degrading quality and forcing service providers to search for solutions. We study the effectiveness of several such solutions. We utilize high frequency household-level data to estimate a (dynamic) model of daily usage during peak and off peak periods. In our setting, demand is dynamic because the plans are three parts tariffs: consumers pay a monthly fee, in exchange get a monthly allowance and pay a per GB fee for anything above that allowance. The dynamics generate variation in the shadow price of usage over the month as consumers are more or less likely to reach the allowance constraint. We use this shadow price variation, as well as variation in congestion due to network upgrades, to estimate consumers price and congestion sensitivity. We estimate the model allowing for a flexible distribution of heterogeneity. 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.

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. Jim Metcalf, Andre Watson, and Craig Westwood provided expert IT support for this project. We gratefully acknowledge financial support from NSF (grant SES-1324717), and Jonathan Williams from CableLabs and the NET Institute (www.netinst.org).

[†] University of Georgia, jbmalone@uga.edu.

[‡] 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 the 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 local governments to pursue municipal broadband and other public funding to support further investment and competition.

In this paper, we aim to understand the impact of congestion on consumer demand and to measure the effectiveness of several solutions to dealing with congestion. We estimate demand by using data on high frequency usage in a setting with three-part-tariffs and variation in network congestion. Our results are of particular importance to any public policy debate that evaluates the value created by broadband investment.

At the heart of the paper is a unique data set made available by a North American ISP. These data include hourly observations of Internet usage and network conditions for roughly 45,000 subscribers from February 2015 through December 2015. At the daily level, we are able to uniquely map an account to a cable modem and active Internet plan. For each Internet plan, we observe the price, advertised speeds, usage allowance, and overage fees. All data tiers charge for data overages at the same per GB rate. The average subscriber in our data uses 2.3 gigabytes (GB) per day, pays \$58.89 for a 22 megabit per second (Mbps) downstream connection, and a 267 GB monthly usage allowance.

Congestion can happen in several places along the network. 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 typically only provides a fixed amount of shared bandwidth to subscribers. Node congestion occurs when demand pushes or exceeds the node's limitations. 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.

A common way ISPs invest in the core network to improve capacity and lower congestion for a group of subscribers is by splitting nodes. A node is a common place for bottlenecks to occur and are what commonly demarcate local, "last mile" networks.

[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`.

When a node is split, its subscribers are distributed evenly across two new nodes, where network conditions should be improved. Many operators target nodes to be split once average utilization exceeds certain thresholds. We observe 5 node splits in our data and use these events to compare before-and-after congestion and subscriber usage. After a split, average daily usage increases by 7% and packet loss, our measure of congestion, drops by 27%. This suggests there is value to consumers from a less congested network.

Our model of subscriber Internet consumption builds on the framework of Nevo et al. (2016) with two notable differences First, we include network congestion and its impact on plan choice and consumption. Second, in our model consumers make bth peak and off-peak consumption decisions (and not just a daily consumption decision). Similarly, our estimation relies on variation in prices and speeds across plans and (shadow) price variation across the billing cycle that is created by usage-based pricing. We also utilize variation in a subscriber's observed packet loss to estimate the effect of congestion.

The price variation arising from usage-based pricing is a result of its three-part tariff structure: a subscriber pays a fixed fee each month, and if the associated usage allowance is exceeded, she is charged at a per GB rate thereafter. While overage fees are only assessed if a subscriber exceeds the usage allowance, a forward-looking subscriber understands today's consumption marginally increases the likelihood of exceeding the usage allowance before the end of the billing cycle – this is a function of how many days remain in the billing cycle and what fraction of the usage allowance has been used previously in the cycle. We incorporate these dynamics in our model similar to Nevo et al. (2016) by allowing consumers to make daily consumption decisions across a billing cycle.

We also use variation in network congestion to identify a subscriber's sensitivity to poor network states. Our hourly data contain *packet loss*, or the percentage of total packets requested that are either dropped or delayed, at the subscriber level, which we use to proxy for network congestion. As mentioned previously, packet loss is one statistic the FCC uses to benchmark network performance across ISPs.

We estimate this finite horizon, dynamic choice model by solving the dynamic problem once for a large number of types. The solution to these dynamic problems is then used to estimate the distribution of types over our sample by minimizing the error between observed and optimal behavior across types. In general, the estimated marginal and joint distributions illustrate the strength of the flexibility built into our estimation approach.

These demand estimates are used to measure the welfare implications of several ways proposed to reduce congestion.

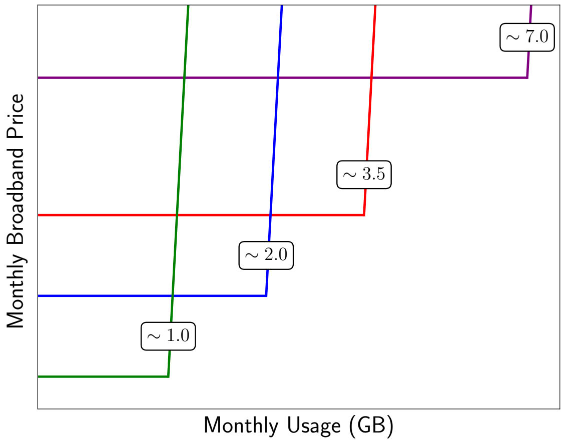
This paper is most closely related to a literature that studies the demand of residential broadband. Recent examples are Nevo et al. (2016), Malone et al. (2016), and Malone et al. (2014) that use similar high-frequency data to study subscriber behavior. However, this literature dates back to the early 2000s with 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).

2 Data

The data for our analysis are a representative sample of 46,667 North American broadband subscribers. The metropolitan area where the subscribers are drawn from have 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. The data include hourly subscriber usage and details of network conditions for February 1^{st} through December 31^{st} of 2015, and are constructed from three primary sources. The first source is Internet Protocol Detail Records (IPDR), which report hourly counts of downstream and upstream bytes, packets passed, and packets dropped/delayed by each cable modem IPDR also record a cable modem's node, a device that connects a set of customers to the rest of the ISP's network. The second data source is average hourly utilization by node. The last data are billing records by customer, where service plan details (e.g., speed, usage allowance, and prices) are included. These data sets are linked by a consumer's account number. Using the account number to link the data sources, rather than a MAC address, permits tracking consumers across hardware changes within the sample.

In addition to the data we use for our main analysis, we also discuss statistics from complementary data, from another ISP for the same period, February–December 2015. This data is national in scope and includes information from a deep-packet inspection (DPI) platform, which 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.

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.

2.1 Sample, Internet Plans, and Monthly Usage

Our panel of data includes a total of over 330 million *subscriber-day-hour* observations. For each observation, we observe downstream/upstream bytes and the total number of packets passed and dropped/delayed. At a daily frequency, we observe each consumer's plan and the mapping of consumers to nodes in the network.

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).

Consumers on more expensive Internet plans use more data on average. In Table 1, we present the distribution of daily 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 or (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.

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 (of the *subscriber-day* usage distribution) for the four Internet service plans and entire sample.

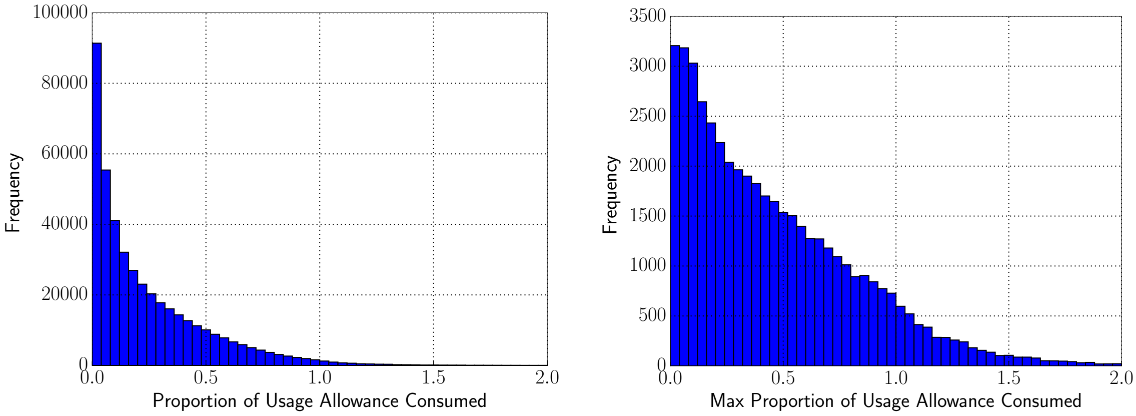
Thus, consumers with greater and more variable usage select more expensive plans, similar to the findings of Lambrecht and Seim (????).

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. This rate of overages is notably lower than the approximately 10% rate reported in Nevo et al. (2016), and is largely due to the recent substantial increase in allowances. 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.

2.2 Temporal Patterns in Usage

Temporal patterns in usage play a crucial role for understanding the potential for more efficient use of broadband networks. Figure 3 presents statistics on how usage varies by time of day. Panel (a) of Figure 3 displays 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

Figure 2: *Distribution of Proportion of Allowance Used*

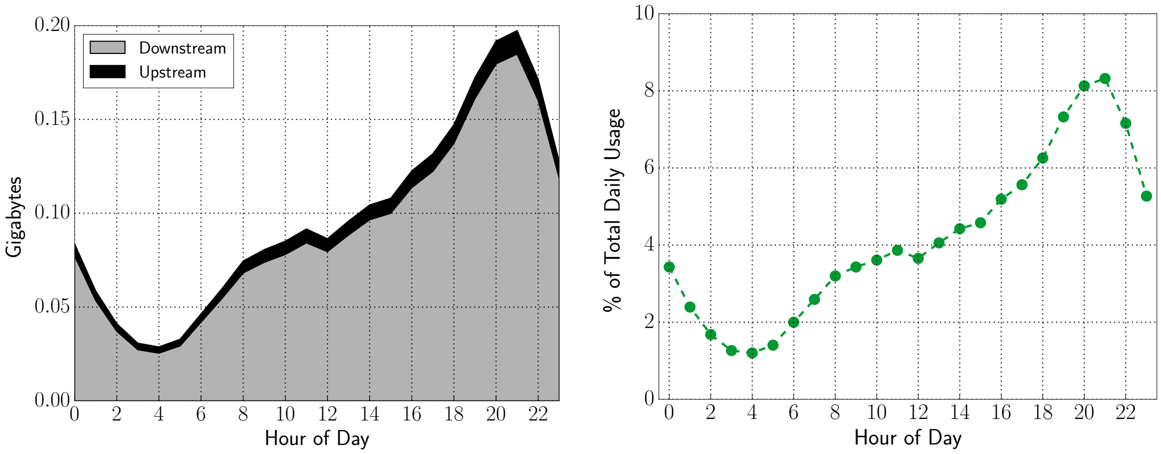


(a) Hourly Percentages

(b) Hourly Percentages by Decile

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.

Figure 3: *Temporal Usage Patterns*

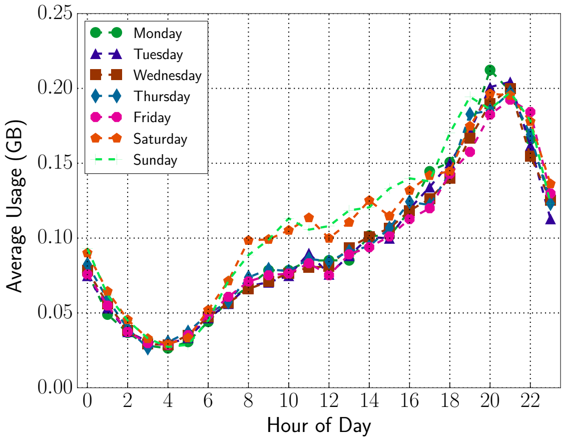


(a) Daily Usage by Hour

(b) Daily Usage %s by Hour

Note: This figure presents statistics on how usage is distributed throughout the day. In panel (a), daily usage in the This figure presents two figures related to how daily usage is proportionally distributed across the day. In panel (a), average hourly usage in gigabytes are reported for each hour aggregated across all subscribers. In panel (b), we report the proportion of overall usage during each hour across all users (series sums to 100%).

Figure 4: *Day-of-Week Variation in Temporal Usage*



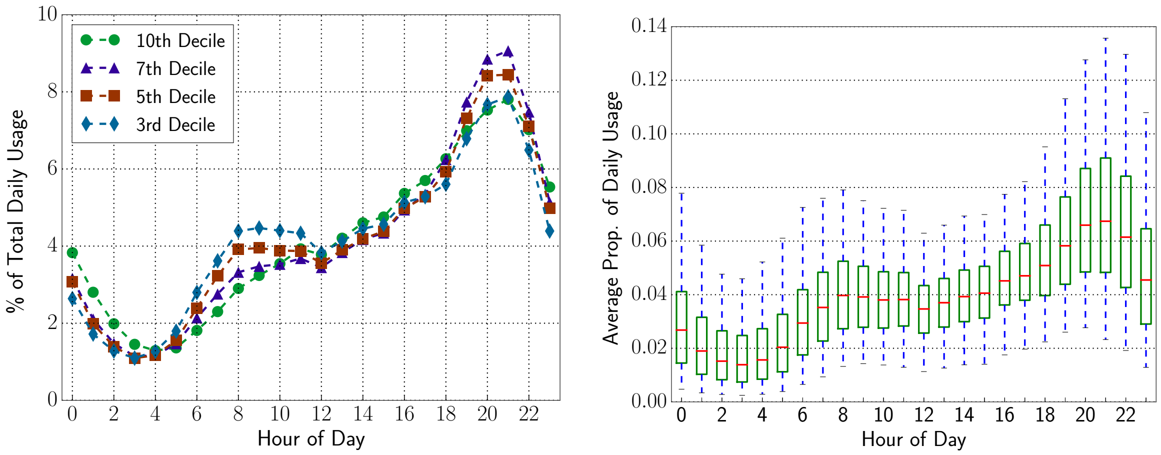
Note: This figure presents average daily usage in gigabytes for each day of the week.

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. Panel (b) of Figure 3 presents the proportion of daily usage by each hour, showing that the peak hour (9PM) alone represents just over 8% of daily usage.

We find that these average temporal patterns do not differ substantially by the day of the week. Figure 4 presents average usage by hour for each day of the week. There is a small intuitive difference between average usage on weekdays and weekends during off-peak hours. From 8am until roughly 3pm, usage is slightly higher during the weekends when individuals are more likely to be home and using the Internet. In our modeling and analysis, we ignore this small difference, because for our purposes the difference is small and occurs during off-peak hours, thereby having a minimal impact on network costs.

Perhaps most interestingly, 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

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



(a) Hourly Percentages

(b) Hourly Percentages by Decile

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.

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 3, 4, and 5 show a strong and consistent pattern in usage across the day. This pattern suggets ISPs must invest enough in their network to meet demand or the network will become poor and unreliable. One unique feature of our data is we observe numerous periods of excess demand placed on the network that result in congestion. Additionally, we also observe the ISP make substantial investments to increase the capacity and improve the quality of their core network. The behavioral response of subscribers to variation in congestion is of primary interest to our analysis. We next discuss measures of congestion, and behavioral responses to congestion–mitigation efforts by the ISP.

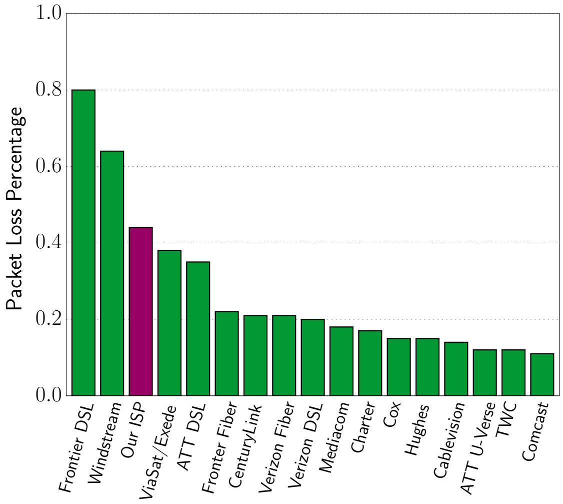
2.3 Network Congestion and Packet Loss

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*.

One advantage of hourly packet loss over node utilization is that 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. Depending on if the data are dropped or delayed, the subscriber's computer may have to request the data again, further increasing the time of delivery. Packet loss is more likely to occur when nodes are highly utilized, which we observe in our data.[2] We only observe node utilization at the hourly level, too, which may not be granular enough to

[2] The two measures, hourly utilization and packet loss, have a correlation coefficient equal to 0.164.

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%.

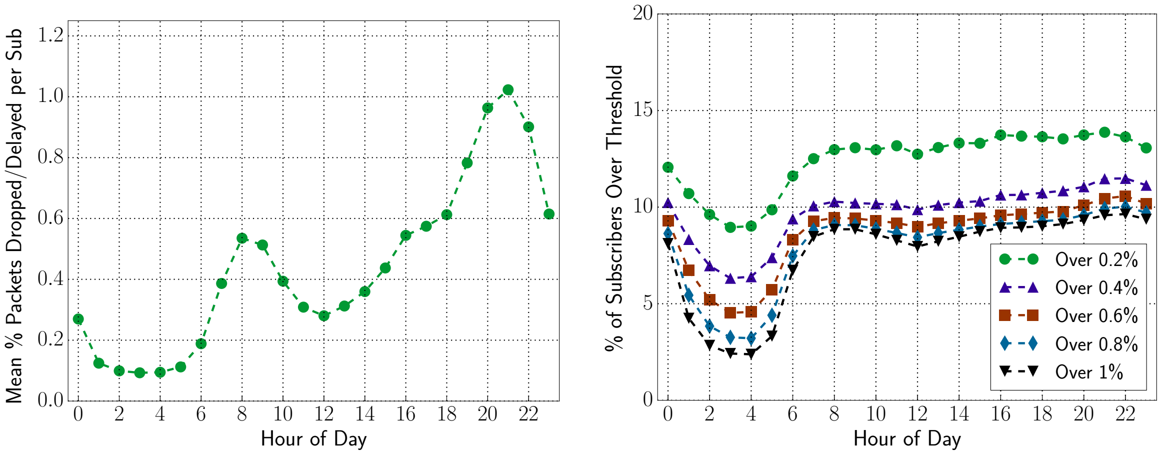
accurately reflect a subscriber's experience within an hour. However, the performance of the network at the instant the subscriber sends and receives packets will be reflected in subscriber-specific hourly packet loss measures.

As part of the FCC's efforts to monitor the quality of broadband networks, it produces an annual report on the state of broadband networks titled "Measuring Broadband America Fixed Broadband Report". In the 2015 version, the FCC includes analysis of data from special (SamKnows) modems, which are distributed across numerous ISP networks. In addition to serving their normal function, these modems conduct hourly tests to measure the performance of the network. One of these tests seeks to measure packet loss by performing an FTP transfer to designated servers located throughout the US. Figure 8 of their report, which we have reproduced in Figure 6, presents statistics on packet loss from each ISP's network. While it is not clear how the values in their Figure 8 are calculated, our experience working with identical data suggests the reported statistics are average hourly packet loss, where the average is taken across subscribers and hours of the day. In our reproduction the FCC figure, we calculate the same statistic for our ISP. 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). If you calculate the overall percentage of daily packets lost, rather than the average of hourly percentages, our ISP would be the worst at around 0.9% of all packets lost. By either defintion, our sample represents a great opportunity to study the importance of congestion in broadband networks.

We now provide a more detailed analysis of the measures of congestion we have in our sample. Most packets are passed during peak hours when usage is highest. This relationship follows from how requests are made online. Whenever a subscriber requests a website, a file to download, or a video to stream on the Internet, packets are sent between the subscriber's computer and the item's location. We also observe the highest frequency of dropped/delayed packets during peak hours in panel (b). The reasoning is twofold. First, there are more packets passed during these hours. Second, peak hours are when node utilization is highest.

At the subscriber level of observation, the distribution of hourly packet loss is highly skewed. For example, in panel (a) of Figure 7, average packet loss is around 1% at 9PM. However, from panel (b), we find over 85% experience less than 0.2% packet loss. 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,

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.

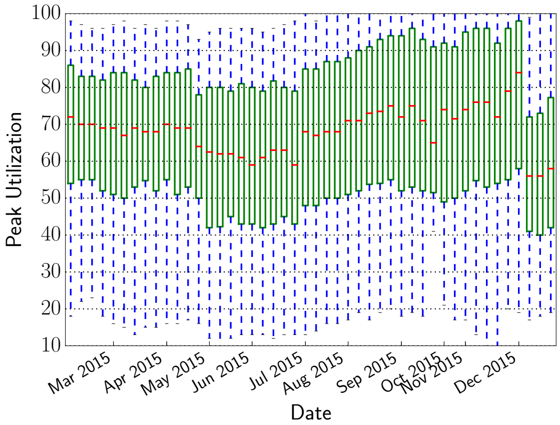
dropping a single packet could be the difference in a website failing to load correctly. This is important from a modeling standpoint, as we provide a flexible framework to estimate a rich distribution of tastes, which accounts for heterogeneity in the types of content the individual prefers to consume.

Panel (b) of Figure 7 better captures the right-tail of the packet loss distribution. In this Figure, we present the 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.

2.4 Evolution of Network Quality

Since our data span a ten-month window, we observe changes in the overall quality of the network that, given the correlation between node utilization and packet loss, would improve packet loss and the network state. In panel (a) of Figure 8, the weighted average of peak node utilization is plotted for each week in our panel. Not only is there variation

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.

across the year, but there are distinct drops in May, September, and December where the ISP improved node capacity. These changes are also noticeable in how median peak utilization varies in panel (b). The dashed whiskers in panel (b) 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*. 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. An example of a node split is for the operator to take a node and split its subscribers across two new nodes. 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. Changes in network conditions are summarized in Table 2 and subscriber usage in Table 3. We do see improvements in the average network state with decreases in both utilization and packet loss. Maximum hourly node utilization 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

Table 2: *Changes in Node Utilization and Packet Loss 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%

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.

Table 3: *Changes in Daily Usage After Node Split*

	Before	*After*	*Diff*	*% Change*
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 subscriber behavior changed around a node split. 7 days of daily usage is taken from before and after the node split date to calculate means. This table summarizes usage for 2,627 subscribers over 5 node splits.

node split that is now able to be realized, and weak evidence of intra-day substitution to avoid congestion during peak hours.

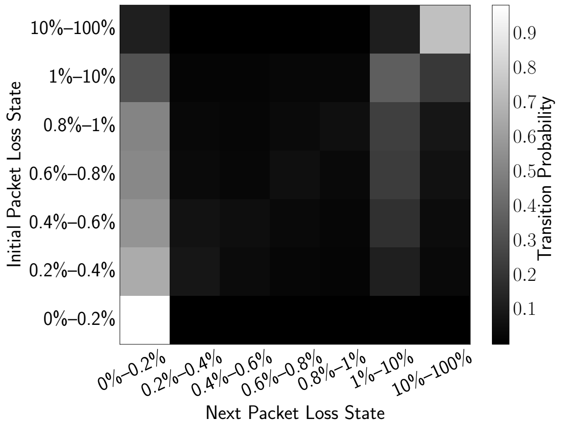
In Table 4 and Figure 9, packet loss is split into seven bins that are used to study how persistent packet loss is day-to-day; the values in the heat map are the same as in the table. From these transition probabilities, there are a couple of notable takeaways. First, if a subscriber's peak packet loss 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 4 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.

2.5 Composition of Usage and External Validity

The data discussed above, and used for our analysis, does not provide insight into the types of online activities for different types of users, only the volume and timing.

Figure 9: *Heatmap of Peak Packet Loss Transitions*



Note:

Table 4: *Transition Matrix of Peak Packet Loss*

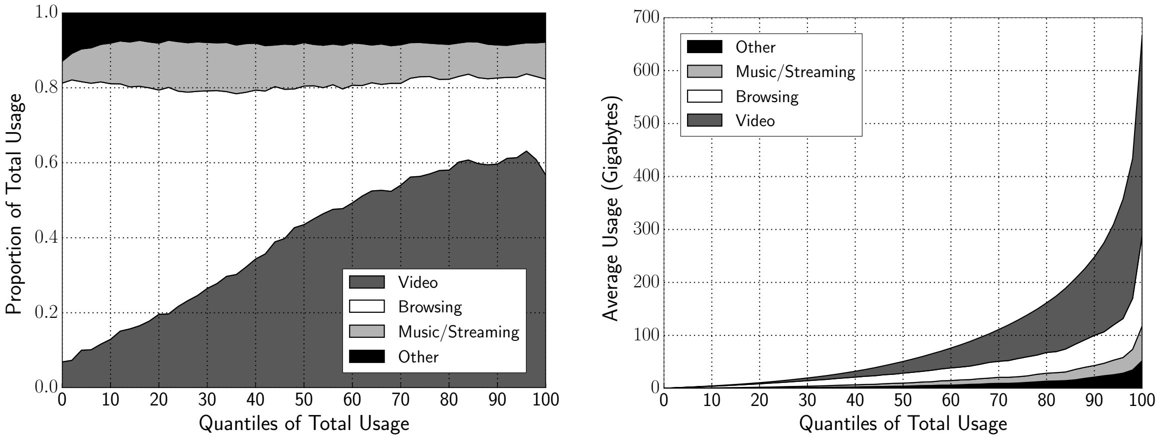
	Next State						
Initial State	*0–0.2*	*0.2–0.4*	*0.4–0.6*	*0.6–0.8*	*0.8–1*	*1–10*	*10–100*
0–0.2	0.984	0.002	0.001	0.001	0.001	0.006	0.004
0.2–0.4	0.662	0.086	0.044	0.027	0.021	0.124	0.037
0.4–0.6	0.570	0.074	0.055	0.037	0.027	0.186	0.051
0.6–0.8	0.526	0.041	0.031	0.062	0.039	0.235	0.066
0.8–1	0.511	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 5: *DPI Application Groupings*

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 10: *Data: Monthly Usage by Quantile and Traffic Type*



(a) Deciles of Daily Usage %s by Hour

(b) Distribution of Daily Usage %s by Hour

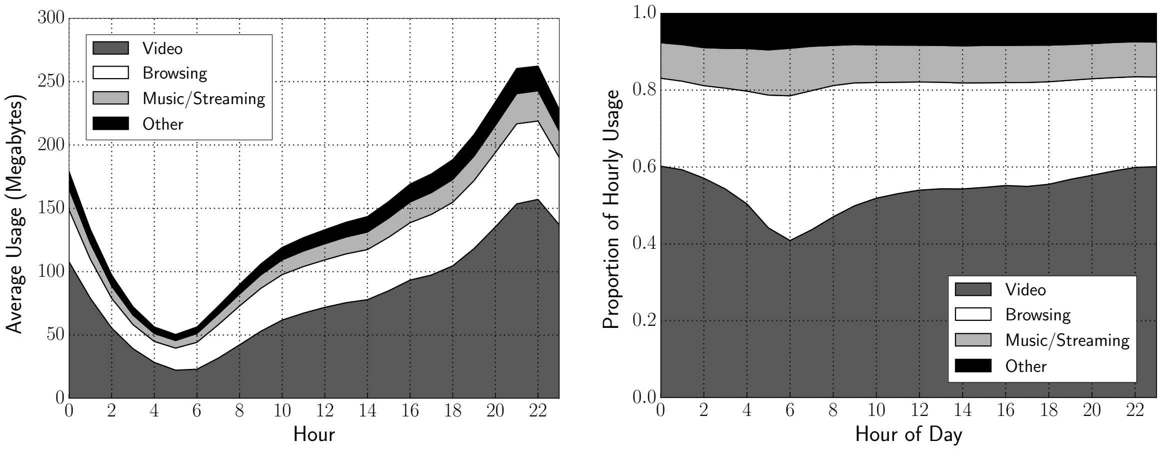
Note: This figure presents two figures related to how daily usage is proportionally distributed across the day. In panel (a), we report the average proportion by hour for certain deciles of users where the deciles are based on total usage (e.g., the tenth decile is the top 10% of consumers). In pane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

We now provide descriptive statistics 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 this 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 5 presents the way we group traffic from the DPI platform, and reports the overall percentage of usage for each of the 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.

Figure 10(a) 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

Figure 11: *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.

greater proportion of video is associated with a lesser proportion of browsing, other proportions remain largely unchanged. Figure 10(b) presents the overall level of usage for each of the quantiles, again broken down by traffic type. The median usage is around 45 GBs, corresponding to about 19 GBs of video each month, or roughly 10 full-length movies. For the 99^{th} percentile user, monthly usage averages over 700 GBs, corresponding to nearly 400 GBs of video, or roughly 200 full-length movies.

Figure 11(a) and 11(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 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 strong source of external validity for the data for our 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. Additionally, despite the astounding magnitude of the usage by some consumers, the patterns in the DPI data present an opportunity of sorts for simple economic and technological solutions to congested networks. The vast proportion of traffic for the highest-usage consumers is comprised of video, which is a passive activity.

Specifically, other than live TV which represents a tiny share of OTTV service traffic, video content need not be downloaded at the same time that it is consumed. Thus, technological and economic solutions to assist in moving video downloads to off-peak hours may introduce efficiencies and improved welfare. Exploring these solutions will be the focus of our counterfactual exercises.

3 Model

Our model builds on the model of Nevo et al. (2016). Like Nevo et al. (2016), the consumer makes a series of usage decisions on an optimally chosen plan over a finite horizon. However, there are a few substantial and important differences between the models. We explicitly incorporate network congestion into the model, which captures a salient feature of our data. Additionally, we further disaggregate the daily usage decision into daily peak and off-peak usage decisions. These additions to the model permit a more flexible framework to measure the return to investing in broadband networks and studying richer and more-interesting counterfactual scenarios.[3]

3.1 Subscriber Utility From Content

Subscribers derive utility from consumption of content. 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.[4] 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.[5] Let consumption of content during peak and off-peak hours be denoted

[3] Given that our focus is on the role of congestion, as mentioned above, we limit our sample to only subscribers who never switched plans over the duration of the panel. This does not affect our analysis because service plans were upgraded shortly before our sample such that 90% of subscribers made no changes during our period of observation. Allowing for plan switching introduces a dependency across billing cycles in the dynamic problem, and by not modeling it the computational burden of solving the model is reduced substantially.

[4] 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.

[5] Nevo et al. (2016) show that there is only weak evidence to suggest content is transferred across days and billing cycles.

by c_p and c_{op}, respectively. The utility for a subscriber of type h on plan k is given by

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

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. We assume that υ_1 is independently and identically distributed according to an exponential distribution with parameter, λ_{1h}, for each consumer type, h.

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 $c_p \frac{\kappa_h}{\ln(\psi s_k)}$ 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 and the state of the network to determine the marginal cost of consuming the content, capturing the consumer's preference for speed and the opportunity cost of the consuming the content. Importantly, for any finite speed, this specification implies that each subscriber type has a satiation point even in the absence of overage charges.[6] 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 $c_{op} \frac{\upsilon_2 \kappa_h}{\ln(s_k)}$. 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 shock is iid and distributed as a truncated exponential distribution with parameter λ_{2h}.[7]

[6] Our specification of this cost departs slightly from that of Nevo et al. (2016), by not including an additive fixed value, κ_1 in their model, that does not interact with speed. However, this parameter is only weakly identified, as the limiting case with unbounded speed that would fully reveal this cost does not occur in our data.

[7] Only the relative changes in peak vs. off-peak costs, and its variability, is identified conditional on the network state and a realization of υ_1. This makes it unnecessary to have a stochastic element to costs during peak and off-peak hours.

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.

3.2 Optimal Usage

Like Nevo et al. (2015), 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, $G_{\psi_t|\psi_{t-1}}$, 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. For a consumer on plan k, we denote the amount of his unused usage allowance on day t as $\overline{C}_{kt} \equiv Max\{\overline{C}_k - C_{t-1}, 0\}$, where C_{t-1} is cumulative usage up until that day. Similarly, denote day-t overages as $\mathcal{O}_{tk}(c_{t_{op}} + c_{t_p}) \equiv Max\{c_{t_{op}} + c_{t_p} - \overline{C}_{k(t-1)}, 0\}$.

A recursive solution to the dynamic program is straightforward, but 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 $\frac{\partial u_{hk}(c_{T_{op}}, c_{T_p}, \psi_T, \upsilon_{\mathbb{T}})}{\partial c_{T_p}} = 0$, the remaining portion of their allowance such that $\overline{C}_{kT} = 0$, or incur overages such that $\frac{\partial u_{hk}(c_{T_{op}}, c_{T_p}, \psi_T, \upsilon_{\mathbb{T}})}{\partial c_{T_p}} = p_k$. Denote this optimal level of consumption, or the policy function on day T during peak hours, as $c^*_{hkT_p}(c_{T_{op}}, C_{T-1}, \psi_T, \upsilon_T)$.

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

[8] The observability of the network state and our focus on the approximately 90% of consumers enrolled on a single plan the entire sample period simplifies the characterization of optimal usage by eliminating inter-billing cycle dependency.

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

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

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

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, $c^*_{hkt_p}(c_{hkt_{op}}, C_{t-1}, \psi_t, \upsilon_t)$, satisfies

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

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

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

These state-dependent integrated policy functions, along with various transformations of the policy functions, are stored along with the value functions when the model is solved for each customer 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

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

The optimal plan, k_h^*, maximizes expected utility for the subscriber given the current state of the network and optimal usage decisions, $E\left[V_{hk1}(C_1 = 0, \psi_1)\right]$, net of the plan's fixed access fee, F_k. 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. (2015), (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 a 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 conditional integrated usage policy functions for a consumer type (h) on their optimal plan (k_h^*), $c^*_{hk_h^*t_{op}}$ and $c^*_{hk_h^*t_p}$, serve directly as the basis for our estimation procedure. The goal of the procedure being to assign each consumer in our data a type from the model by comparing observed behavior to that predicted for each type. This differs from Nevo et al (2016)'s application of the fixed-grid random-effects (FGRE) methodology of Fox et al (2015), which aggregates policy functions across types (h) to match unconditional population moments. This aggregation procedure requires linearity in the type-specific weights to be computationally feasible (a linear objective function). Our panel data adaption of the Fox et al (2015) methodology, which we discuss in Section 4, circumvents this linearity restriction along with other advantages.

4 Estimation

Our estimation approach is a panel-data modification of Fox et al. (2015), which Nevo et al (2016b) refer to as a fixed-grid fixed-effect (FGFE) estimator. The approach exploits the richness of panel data to build upon the application by Nevo et al (2016a) of the fixed-grid random-effects (FGRE) approach of Fox et al. (2015). In contrast to the FGRE approach, our FGFE approach permits identification of each subscriber's type, rather than just the distribution of types, and also allows consideration of moments from the model that are not non-linear in the type-specific population weights. This is advantageous for identification of the model and consideration of richer counterfactual exercises where knowledge of an individual's type is useful rather than just the distribution of types. In applications where observable consumer characteristics are available, recovering each consumer's type permits the fixed parameter vector characterizing them to be projected on these characteristics. We discuss these advantages in greater detail below.

4.1 Econometric Objective Function

For each individual, $i = 1....I$, our data includes a series of data, $m = 1.....M$, which captures usage at a daily frequency on an optimally chosen plan.[9] This includes both off-peak usage, $\left(c_{i1_{op}}, c_{i2_{op}}, ..., c_{iM_{op}}\right)$, and peak usage, $\left(c_{i1_p}, c_{i2_p}, ..., c_{iM_p}\right)$, for each consumer (i), as well as the accompanying observable portion of the state on each day (m), $(t_m, C_{t_m-1}, \psi_{t_m})$.

We solve the model for 4,096 types of consumers, corresponding to a fixed grid with eight points of support for each of the four parameters, $(\alpha_h, \kappa_h, \lambda_{h1}, \lambda_{h2})$. The solution to the model yields state-dependent integrated policy functions for peak and off-peak usage on each type's optimal plan, $c^*_{hk^*_h t_{op}}$ and $c^*_{hk^*_h t_p}$, respectively. In addition, we calculate the expectation of the square of optimal usage during off-peak and peak hours, $c^{2*}_{hk^*_h t_{op}}$ and $c^{2*}_{hk^*_h t_p}$, respectively.

The goal of the estimation algorithm is to identify which of the $H = 4,096$ types' behavior from the model best match the behavior of each individual, $i = 1, ..., I$, over the panel of data. We use a least-squares criteria to compare fit, such that the type h that best matches to consumer i is given by

[9] We drop the small fraction of subscribers, less than 2%, for which we do not observe a complete time series.

$$\widehat{h_i} = \min_{\{h=1....H\}} \left[\sum_{m=1}^{M} \tilde{\mathbf{z}}_{ih}' \tilde{\mathbf{z}}_{ih} \right],$$

where

$$\tilde{\mathbf{z}}_{ih} = \begin{pmatrix} c_{im_{op}} - c^{*}_{hk^{*}_{h}t_{m(op)}} \\ c^{2}_{im_{op}} - c^{*2}_{hk^{*}_{h}t_{m(op)}} \\ c_{im_{p}} - c^{*}_{hk^{*}_{h}t_{m(p)}} \\ c^{2}_{im_{p}} - c^{*2}_{hk^{*}_{h}t_{m(p)}} \end{pmatrix}.$$

This process is repeated for each i.[10] Aggregating across the chosen types for each consumer, i, the population weights for each type of consumer, h, is then

$$\widehat{\theta_h} = \frac{1}{I} \sum_{i=1}^{I} \mathbf{1} \left[\widehat{h_i} = h \right].$$

Standard errors on the distribution of types in the population are calculated through a block-resampling procedure. Specifically, a block of dependent data in our application is characterized by a complete billing cycle for each consumer. We repeatedly sample, with replacement, blocks of data for each consumer to create, $r = 1, ..., R$, time series of the same length and dependency structure for each of the I consumers. We then calculate each individual's type for each of the R time series, which yields standard errors for each type's population weight $\widehat{\theta_h}$.

There are a some advantages of the FGFE approach in panel-data applications like ours. First, even compared to the constrained convex optimization problem in the FGRE approach of Fox et al (2015), there can be computational advantages. In that FGRE setting, practical numerical collinearity issues can arise for even moderately dense type grids and parameters that are less strongly identified. The pooling of individuals data to calculate moments creates a mixture problem that the optimization must undo. In instances where the model predicts two types to behave similar at states observed in the data, numerical instability due to collinearity can result. Second, and related to

[10] This approach can easily be adapted to a likelihood criteria for type choice. The assignment of a type to a consumer can also be probabilistic according to the relative likelihood contributions, rather than the deterministic approach taken here. This essentially serves to refine the uniform prior over types that the optimal plan-choice yields into a posterior as in a Bayesian framework. The details of this approach is discussed in Nevo et al. (2016b).

the first advantage, the FGFE approach does not require that the moments used in estimation be linear in the type-specific weights. In the FGRE approaches, this linearity is necessary or the problem becomes a constrained nonlinear optimization problem that is intractable with even a moderate number of types. This limits the moments available to use in estimation, which can lead to difficulty in differentiating between types due to constraints on the types of behaviors predicted by the model that one can exploit during estimation.

Another advantage of the FGFE approach is that it fully characterizes the discrete distribution of types, but in contrast to FGRE demand models like Fox et al. (2015), Nevo et al. (2016a), and other random-coefficient models (e.g. Berry et al. (1995)), the mapping between an individual and a type is preserved ($\widehat{h_i}$). The panel data eliminates the need to aggregate across consumers to form moments, and this permits an individual's type to be inferred, rather than just the distribution of types in the population. In many applications in Industrial Organization, inferring an individual's type rather than only the distribution of types is useful. From the firms' perspective, this may permit different forms of discrimination (third-degree rather than second degree) to be implemented through targeted offerings. Knowledge of each individual's type can also permit a decomposition of the parameters via the minimum-distance procedure of Chamberlain (1982) when observable characteristics of the individual are available, as is done in Nevo (2001). For example, one can regress the parameters describing an individual's type, $(\alpha_{\widehat{h_i}}, \kappa_{\widehat{h_i}}, \lambda_{1\widehat{h_i}}, \lambda_{2\widehat{h_i}})$, on a vector of the individual's observed characteristic to decompose the parameter into observable and unobservable determinants of preferences. The limitation is that the individual's type is persistent or fixed. This is not particularly limiting, because parameters can characterize distributions to explain temporal variation in behaviors like λ_1 and λ_2 in our application.

4.2 Identification

Identification of our model closely follows the discussion in Nevo et al. (2016a). There are a few important differences, each simplifying and improving identification. First, we eliminate the κ_{1h} parameter that led to a satiation point for usage even when speed was unbounded. While we observe much higher speeds than Nevo et al. (2016a), this dimension to the type space is not needed because the limiting case of infinite speed is clearly not in our data, and given the additional computational complexity of our model, eliminating it permits us to consider a denser grid over the other parameters. Second, usage and plan choices are strong sources of identification. Plan choice can be thought of as assigning each type to a plan and putting a uniform prior over the types on each plan, while the usage moments can then distinguish between the types choosing a plan. The flexibility of the FGFE approach is also important here, as we are able to consider richer moments of usage, because we are not restricted to moments that preserve linearity in the type weights. This is how we're able to consider the first and second conditional moments of usage, in contrast to Nevo et al. (2015), which only uses the unconditional first moment.

Perhaps most importantly, we also have an additional source of variation in the price of usage that can help identify a type. Like Nevo et al. (2015), usage-based pricing is particularly helpful, as we observe a large number of marginal decisions by each consumer, weighing the benefit of consuming more content against the increase in the probability of overages (i.e., the shadow price of usage). This variation is helpful for pinning down the primary determinant of an individual's elasticity of demand, $\alpha_{\widehat{h_i}}$. While the fraction of individuals incurring overages in any billing cycle is lower in our data, the proportion of individuals overall incurring overages at any point in the panel is higher so we observe a greater fraction of individuals truly making marginal decisions on usage. However, in addition to this price variation introduced by the nonlinear pricing, we also have extensive variation in the network state, which shifts the cost of consuming content. This is helpful in pinning down an individual's preference for speed, $\kappa_{\widehat{h_i}}$, which would otherwise largely be identified by plan choice alone. It also assists in identifying α_h, which is the elasticity with respect to the complete cost of usage.

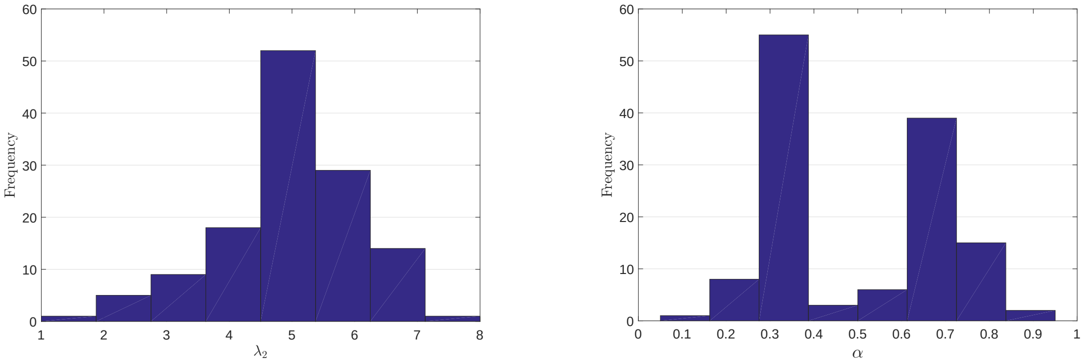
5 Results

We present our estimation results in three parts. First, we summarize our estimates of the types 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 Visualization of Type Distribution Estimates

We estimate a weight greater than 0.01% for 129 types. That is, 129 different types (h) were chosen from the 4,096 candidate types considered. This concentrated mass among a few types is similar to the result from Nevo et al. (2016), which had a slightly smaller proportion of the candidate types receive positive weight. On each plan, the mass is quite concentrated with the most common representing over 30% of the plan's total mass, and the top two types over 40%. Most of the types with positive weight, 85, come from the most-expensive plan, despite it only representing about 2.5% of the sample.

Figure 12: *Marginal Distribution of* λ_2 *and* α



(a) Marginal of λ_2

(b) Marginal of α

Note: The figures are the estimated marginal distributions for λ and α, respectively. Within each panel, the sum of all five bars in each figure total 129, the total number of types.

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 12, we present the marginal distributions for two of the parameters to demonstrate the flexibility of the econometric approach. Figure 12 (a) and (b) give the marginal distribution for λ_2 and α, respectively. The mass for both distributions are well within the chosen support for the parameters, which is necessary so the discrete support of candidate types doesn't constrain optimization over the type weights. The λ_2 parameter that determines the intra-day allocation of usage, along with the network state, has an intuitive shape and distribution. Most consumers, and the average user, have peak usage that averages approximately five times off-peak usage. The mean of λ_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 $\lambda_2 = 1$ (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. The marginal distribution for all of the parameters is summarized in 6.

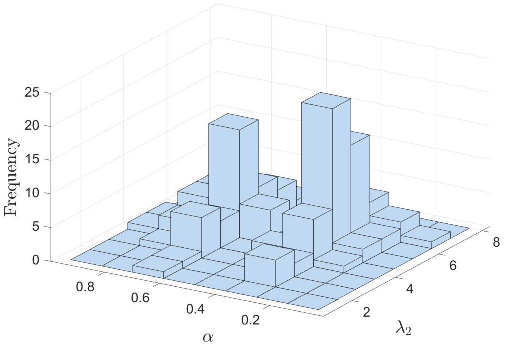
As would be expected given the quite different marginal distributions, the joint distributions are quite irregular. Figure 13 gives the joint distribution of λ and α. The multi-peaked nature of the α distribution is still clearly visible, but there are non-negligible correlations between the two parameters. This demonstrates the importance of the flexibility of our estimation approach, which allows for free correlations between each pair of

Table 6: *Descriptive Statistics for Types*

	Mean	*Median*	*Mode*
α_h	0.62	0.60	0.70
λ_{1h}	8.23	7	7
λ_{2h}	5.71	5	5
κ_h	4.32	4	4

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

Figure 13: *Joint distribution of λ and α*



Note: Joint histogram of λ and α given by frequency counts.

parameters rather than the normality and lack of covariance often assumed in structural econometric applications. This flexibility is reflected in the fit of the model. For all plans, the correlation between the empirical moments and the fitted moments is above 90%. The model also fits patterns in the data not explicitly used in estimation, similar to those reported in Nevo et al. (2016).

5.2 Policy Implications of Preference Estimates

Directly from our estimates, it is straightforward to calculate the distribution of willingness to pay for the central features of broadband service: connectivity speed and usage allowances. Understanding preferences for these attributes have important implications for regulatory guidance in a number of areas.

For connectivity speed, it is straightforward from the model to calculate the value associated with increasing the speed of each type's optimal plan by 1 Mb/s. We simply resolve the model for each type 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, for each type, $\frac{\partial V_1(C_1=0)}{\partial s_{h_{k^*}}}$. We find that a 1 Mb/s increase in speed on each type's optimal plan is valued at and average of \$0.79. This value drops off extremely fast, as a 10 Mb/s increase is valued at less than \$4.

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.[11] Our results suggest the subsidies to cover the costs of access might be quite valuable, as access at even modest speeds is highly valuable, but the larger expenditure to dramatically increase overall speeds would likely not do much to penetration and have quite low value. For example, an HD movie from most OTTV services requires approximately 5 Mb/s, and recent encoding 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, 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 each type's optimal plan. This value equals $\frac{\partial V_1(C_1=0)}{\partial \overline{C}_{h_{k^*}}}$, which is the change in expected utility at the beginning of the billing cycle. We find the median consumer values an additional GB at only \$0.04, while the average is \$0.09 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.[12]

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

[11] This decision can be obtained by going to 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".

[12] 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.

about \$0.08. 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.

5.3 Economic and Technological Solutions to Congestion

Much has been made of the existence of congestion and its impact on customers and the providers of applications and content whose product is degraded during congestion. Our model estimates are helpful for both quantifying the surplus lost from congestion, not necessarily welfare because we have no cost estimates, and also evaluating returns to different solutions to congestion. In our model, the transition process of the network state, ψ, dictates the uncertainty over, and level of, congestion realized by consumers. Thus, the most straightforward way to measure congestion's impact is to simply change the ψ process to always be in the best state where realized and provisioned speeds are equal. We find that completely eliminating congestion results in an average increase of \$6 per month for subscribers. This is perhaps a relatively modest increase in surplus, but entirely consistent with the recently fading preference for increased connectivity speeds. Thus, substantial and costly upgrades to broadband networks well beyond current levels may not be the least-cost way to improve consumer welfare.

An alternative to large-scale investment would be to explore the viability of less-costly economic and technological solutions to maximize existing networks. We explore three potential solutions: peak-use pricing, throttling of connectivity speeds, and local-cache technologies. We compare improvement in consumer surplus under each of these scenarios to the baseline results discussed immediately above where we've eliminated congestion completely.

Throttling Speed

The most-straightforward congestion solution to implement using our model is to compare usage with throttling of connectivity speeds to a baseline case with no congestion. To simulate the effect of such a policy, we present consumers with an option to have their speed slowed after exceeding their usage allowance rather than incur overage

Table 7: *Counterfactual: Throttling*

	Throttling	
Usage and Surplus	*Baseline*	*7 Mb/s*
Daily Usage (GB)	2.5	2.9
Peak Usage (GB)	1.8	2.0
Off-Peak Usage (GB)	0.7	0.9
Consumer Surplus ($)	70.22	78.67
Revenue ($)	57.42	57.09

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 **??**. The effect is a bit counterintuitive, as total usage during peak and off-peak hours increases. However, the majority of consumers opt in, and without the possibility of overages, usage is heavier at all hours. 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 **??**. The first column provides a baseline where congestion is absent from the network, while the second and third columns consider a 30% 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, the bottom row of Table **??** provides a particularly useful insight into the usefulness of peak-use pricing. As the allowance is decreased, the perceived cost of peak usage for the consumer substantially increases. Even though the low elasticity makes the usage response modest, this 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

Table 8: *Counterfactual: Peak-Use Pricing*

	Allowance Reduction		
Usage and Surplus	*Baseline*	*30% Reduction*	*50% Reduction*
Daily Usage (GB)	2.5	2.6	2.4
Peak Usage (GB)	1.8	1.7	1.5
Off-Peak Usage (GB)	0.7	0.9	0.9
Consumer Surplus ($)	70.22	72.33	69.01
Revenue ($)	57.42	57.21	58.54
$\frac{\partial V_1(C_1=0)}{\partial \overline{C}_{h_{k^*}}}$ ($)	0.09	0.12	0.45

Table 9: *Counterfactual: Local-Caching Technology*

	Local-Caching Technology	
Usage and Surplus	*Baseline*	λ_2 *50% Reduction*
Daily Usage (GB)	2.5 GB	3.1 GB
Peak Usage	1.8 GB	1.7 GB
Off-Peak Usage	0.7 GB	1.4 GB
Consumer Surplus ($)	70.22	85.44
Revenue ($)	57.42	58.22

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 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 **??**, 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 a 10-month panel of hourly subscriber usage and network conditions. The key feature of our model is the incorporation of network congestion and allowing it to affect a subscriber's daily consumption decision.[13] There are three sources of variation we exploit in our data. First, we use (shadow) price variation that results from the structure of usage-based pricing's three-part tariff. Second, we use cross-sectional variation in packet loss, our measure of network congestion, across subscribers. Third, our ISP invested in the core network several times throughout 2015, creating times series variation in the overall quality of the network.

Our demand estimates are used to measure the value to subscribers from eliminating network congestion. We find the improved network conditions encourage some subscribers to downgrade, but any loss in revenue is entirely offset by an increase in consumer surplus. Subscribers' realized speeds increased by roughly 18% with each additional Mbps of speed being valued at roughly \$2.87.

There are several extensions to the basic model in this paper that can be explored in future versions. First, we could allow consumers to switch plans. In our sample, around 12% of subscribers make a plan change by either moving to a more expensive or cheaper plan; these subscribers are omitted from our original estimation. Under a usage-based pricing regime, we expect some consumers may upgrade to a plan with a larger usage allowance to account for their growing demand, while others may downgrade to better align their usage with a lower allowance or due to cost concerns.

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.

[13] This differs from previous research such as Nevo et al. (2016).

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.