

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

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

June 6, 2017

Abstract

The growth of the Internet has constrained broadband networks, forcing service providers to search for solutions. We develop a dynamic model of daily usage during peak and non-peak periods, and estimate consumers price and congestion sensitivity using high frequency usage data. Using the model estimates, we calculate usage 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-use 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, peak-use pricing

JEL Codes: L11, L13, L96.

*We thank Shane Greenstein and Eugenio Miravete for helpful discussions, and participants in seminars and conferences for useful feedback. James Myatt provided excellent research assistance. We are grateful to the North American Internet Service Providers that provided the data used in this paper, and 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 video, continues to grow rapidly. Internet Service Providers (ISPs) constantly make investments in an attempt to keep capacity in line with this growth. From 1996 to 2013, these private investments totaled around \$1.3 trillion, or about \$75 billion per year.¹ Historically, broadband investment has been financed by private firms, but the importance of Internet access is leading to public funding and use of additional public policy tools to support and encourage further 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: usage is significantly higher during peak-use periods. Building enough capacity to meet demand during these periods leaves the capacity idle during off peak periods. Both these suggest that an effective approach to dealing with congestion should combine investments with economic solutions to reduce demand during peak periods. In this paper, we explore the effectiveness of several such solutions.

We start by estimating demand for online content using data on hourly Internet usage and network conditions made available 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 variation in the (shadow) price of usage and 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 these estimates to explore the effect of throttling connection speeds, peak-use pricing, and local-cache technology on overall demand and congestion during peak-demand periods.

Before estimating the model, we use our rich data to demonstrate several patterns that inform some of the questions of interest and motivate modeling assumptions. First, there is a very consistent within-day usage pattern: usage is greatest during evening hours

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

²For example, broadband investment to modernize and expand the network was a condition for approval of the Charter/Time Warner Cable merger. (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.

and significantly lower the remainder of the day. This pattern is consistent across days and across consumers with different total monthly usage. Second, measures of network performance are correlated with network utilization, with performance decreasing during peak-use periods. This suggests that any policy that shifts demand from peak to non-peak times will improve network performance. Third, most usage is due to over-the-top video (OTTV) which, at least in principle, can be downloaded during non-peak times with the right incentives and technology in place. Finally, there is a large amount of heterogeneity in usage patterns 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 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 online gaming 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 multiple 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, packet loss, our measure of congestion, drops by 27%, average daily usage increases by 7% and the share of off-peak usage, out of total usage, decreases by 8%. This suggests there is an elasticity of total usage with respect to congestion, as well as substitution between peak and off-peak usage.

To quantify consumers’ willingness to pay for a less congested network, we develop and estimate a model of Internet usage that 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 alter plan choice and usage. Second, we let consumers make both peak and off-peak usage decisions (rather than a single 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 Akerberg (2009), 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, we use the solution to these dynamic problems to estimate the distribution of types by computing, for each household

³To measure congestion we primarily rely on 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 was 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 flexibility of our estimation approach.

Our counterfactual exercise relies on the ability to identify the key parameters within sample and then use the model to simulate the counterfactuals. Price sensitivity, one of the key parameters, is identified from the variation in the shadow price throughout the billing cycle. Overage fees are only assessed if a subscriber exceeds the usage allowance, but a forward-looking subscriber understands today's consumption marginally increases the likelihood of exceeding the usage allowance before the end of the billing cycle. Another key input into the counterfactual exercises, the elasticity of off-peak usage in response to changes in congestion, is identified by the covariation of usage and the (observed) network state. For example, in the data we observe that consumers on congested nodes have a higher fraction of usage off-peak. Similarly, when a node is split consumers immediately experience a less congested network and we see them increasing peak usage and decreasing off-peak usage. Both suggest usage is responsive to congestion.

We find most consumers have a low willingness to pay for faster advertised speeds as well as for increased allowances. On the other hand, consumer surplus is 20% greater per month, on average, if congestion is eliminated during peak-demand periods. From our policy counterfactuals, we find throttling speeds has little effect on usage and actually increases it slightly. The reason usage increases is that some consumers view a slower connection as less of a deterrent than the possibility of incurring overages, which leads to higher usage. Next, we explore peak-use pricing and find it also 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.

Our results, both directly from the model estimates and from the counterfactual simulations, have implications for multiple ongoing policy discussions. First, the low willingness to pay for greater network speeds, but substantial willingness to pay for abating congestion, suggests that when public policy tools are leveraged to encourage investment the focus should be on consistency of service rather than pushing network capabilities.

Second, the low value subscribers place on larger usage allowances demonstrates that usage-based pricing, as currently implemented by our ISP, does not impose a substantial

constraint on the average household. Thus, the zero-rating policies of ISPs⁴ that have received substantial regulatory attention are currently having little impact on OTTV services, consistent with their rapid adoption.

Third, the welfare improvements we find from peak-use pricing and local-cache technologies are likely to be difficult to realize in practice because of the complex incentives facing ISPs. In particular, it is not clear that ISPs have the incentive to implement peak-use pricing in a manner similar to our counterfactual simulation. Permitting caching of OTTV content during off-peak periods may diminish the relative attractiveness of the ISP's own video services. On the other hand, the ISP might have an incentive to make attractive as possible an activity that currently accounts for two-thirds of traffic and presumably provides substantial value. Exploring this question in greater depth requires a model of demand for bundles, pay-TV, and Internet services, and also more detailed data on the subscribers' online activities. We view our model as a useful first step towards a richer exploration of these issues. See Greenstein et al. (2016) for a detailed discussion of related issues.

Our work is related to several strands of papers in the literature. First, this paper, like Nevo et al. (2016), Malone et al. (2016), and Malone et al. (2014), uses high-frequency data to study subscriber behavior on residential broadband networks. However, this paper differs in several important ways from the earlier work. Malone et al. (2016), and Malone et al. (2014) are purely descriptive. Nevo et al. (2016) also estimate demand, and examine the effectiveness of usage-based pricing 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 in several dimensions. Most importantly, our previous work only looked at total daily usage and not the split between peak and off peak. This is driven by the second major difference, a different set of questions focusing on the implications of economic solutions to congestion, which we could not address using our previous model. Finally, in this paper we estimate the model by explicitly using the individual choices over an eleven month period, while in our previous work the estimation relied on aggregate moments generated from individual decisions over a single month (and from a different ISP).

The paper also relates to a broader literature studying various issues in the telecommunications industry. Varian (2002) and Edell and Varaiya (2002) run experiments where consumers face different prices, along with varying allowances and speeds. Goolsbee and

⁴Zero rating policies are policies where content from certain providers does not count towards the allowance. Some have raised concerns with these policies because they can be used to create a price differential between third party OTTV services and those of the ISP. Potentially leading to foreclosure and, according to some, violating the principles on net-neutrality.

Klenow (2006) estimate the benefit of residential broadband from time-use surveys and Hitte and Tambe (2007) show Internet usage increases by roughly 22 hours per month when broadband is introduced. Lambrecht et al. (2007) study how uncertainty with respect to usage influences choices over plans with three-part tariffs. Other related papers are Dutz et al. (2009), Rosston et al. (2013), and Greenstein and McDevitt (2011). More recently, Goetz (2016) studies congestion at network interconnections and Wilson (2017) examines whether public investment in broadband networks crowds out private investment.

The question of intra-day variation in demand and the role of peak-use pricing also comes up in other contexts, real-time pricing of utilities being a prime example. Recent work in the this area includes Wolak (2007), Wolak (2010), Wolak (2016), Strapp et al. (2007), Ito (2014) and Anderson et al. (2017). Gowrisankaran et al. (2016) study the related issue of intermittency in the supply of renewable energy. For a more complete survey of this literature, see Newsham and Bowker (2010) and Fauqui and Sergici (2010).

The modeling in this paper is related to several literatures that study behavior in dynamic settings. First is the literature studying whether consumers are forward-looking and respond to dynamic incentives (Aron-Dine et al. (2015), Chevalier and Goolsbee (2009), Grubb (2015), Grubb and Osborne (2015), and Hendel and Nevo (2006b)). Second, is the literature on estimation of demand in dynamic settings using high-frequency data on consumer consumption behavior (e.g. Crawford and Shum (2005), Hendel and Nevo (2006a), Gowrisankaran and Rysman (2012), Yao et al. (2012), and others). Finally, dynamic incentives like those generated by nonlinear pricing in our application also arise in other settings. In labor applications, compensation schedules that create an association between performance over a fixed period of time and salary make effort decisions by workers dynamic (e.g. Copeland and Monnet (2009), Chung et al. (2010), Misra and Nair (2011)). In health care applications, deductibles and limitations on out-of-pocket expenditures create similar tradeoffs (Einav et al. (2015b) and Einav et al. (2015a)).

2 Preliminary Analysis

In this section, we describe our data sources and sample, and then provide preliminary analysis that motivates the modeling that follows. We first present descriptive statistics on plan selection and monthly usage. Second, we show that there are clear intra-day usage patterns with marked peaks in the evenings, and inter-day usage patterns that demonstrate consumers respond to (shadow) price variation arising due to the plan's three-part tariffs. Next we show network performance is correlated with overall usage and noticeably declines during peak hours, and then provide direct evidence of the effect

of congestion on usage. Finally, we discuss the composition of usage and the external validity of our sample.

2.1 Data Sources

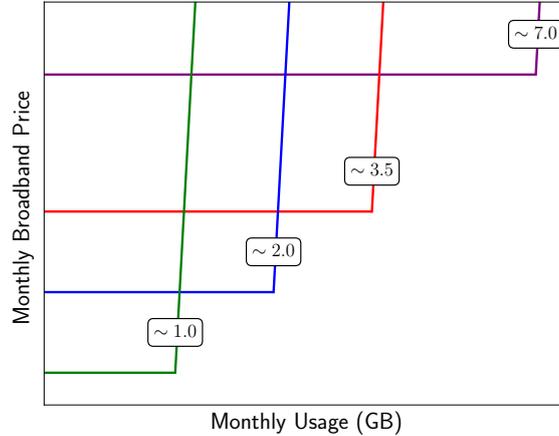
The main dataset we use comes from a North American ISP. The provider offers several plans with features that include a maximum download speed, an access fee, usage allowance, and overage price per GB for data in excess of the usage allowance.⁵ Usage is recorded in GBs for both uploads and downloads, 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 Internet Protocol Detail Records (IPDR), a common measurement platform used by ISPs for usage-based billing. 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 data on hourly utilization by node and 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 also present 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 provides insight into the types of traffic (OTTV, gaming, web browsing, etc) generated by each user. This 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 on the composition of traffic are helpful in explaining the findings from our model, and demonstrate the external validity of the data used in the analysis.

⁵Subscribers are not on long-term contracts, only incurring a disconnection fee if service is canceled.

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. Due to the non-disclosure agreement with the ISP that provided the data we cannot populate exact numbers in this graph and can only display the general relationship. 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.

2.2 Plan Selection and Monthly Usage

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 1.⁶ The steep slope of the line after the allowance is met indicates the fairly high per GB overage fee this ISP charges. 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 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

⁶Due to the non-disclosure agreement with the ISP that provided the data we cannot populate exact numbers in this graph.

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

| | <i>Tier 1</i> | <i>Tier 2</i> | <i>Tier 3</i> | <i>Tier 4</i> | <i>All</i> |
|------------------------|---------------|---------------|---------------|---------------|------------|
| 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 |
| <i>N</i> | 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.

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).

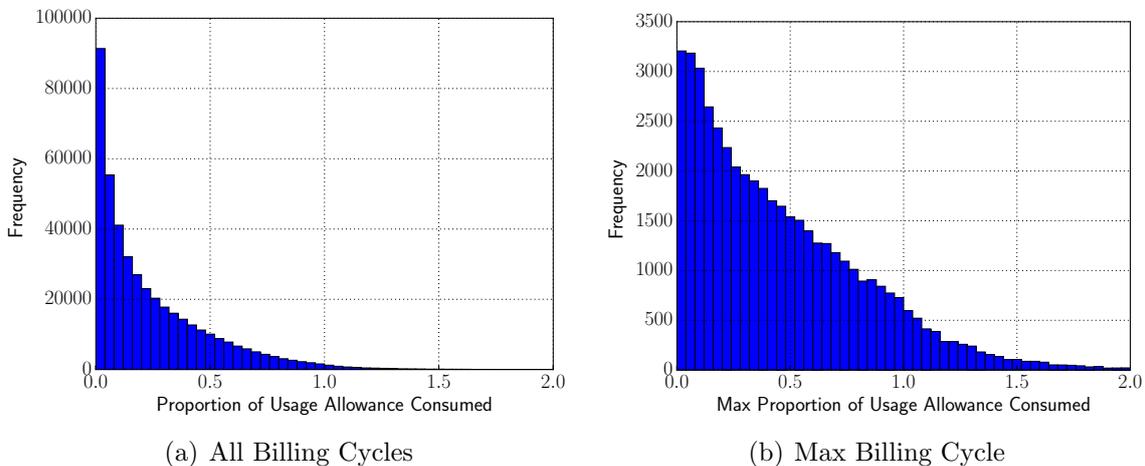
2.3 Temporal Patterns in Usage

In this subsection, we display patterns in usage across days and within days. First, we study the distribution of usage across 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 consumers. We show that a significant fraction of consumers get near the allowance at some point during the 11-month sample. We then provide 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 observe a very regular pattern of peak usage during the evening and much lower usage at other times.

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

Figure 2: *Distribution of Proportion of Allowance Used*



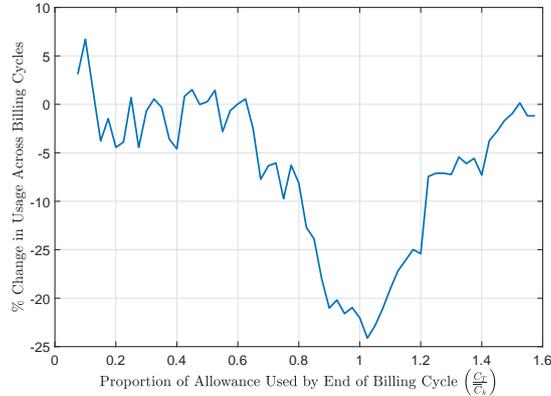
Note: This figure presents two graphs related to the distribution of the proportion of the allowance used by consumer's each month. In panel (a), we present the distribution of this proportion for all customer billing cycles resulting in 11 observations for each customer. In panel (b), we present the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

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%. This probability, together with the high price that the ISP charges for usage over the allowance, is enough to induce significant dynamic behavior.

A natural question to ask is whether consumers respond to the shadow price variation implied by the dynamic problem we pose below. Nevo et al. (2016) explore this question using both within-month and cross-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. Finally, a consumer well above the allowance, who likely also expects to be over the allowance next month, sees no price change.

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

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 billing cycle varies with the proportion of the allowance consumed by a subscriber at the end of the billing cycle.

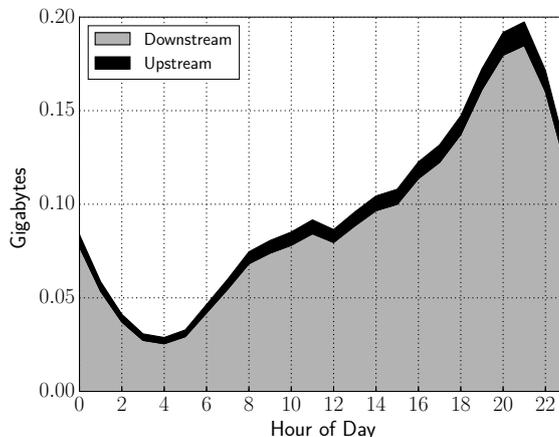
calculate the mean percentage change for groups of subscribers that used various fractions of the allowance by the end of the billing cycle. Like Nevo et al. (2016), we find subscribers facing a price increase at the beginning of the next month consume relatively more at the end of the current month, 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.

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 for both upstream (e.g. uploading a file to Dropbox) and downstream (e.g. streaming movie from Netflix) usage. The proportion of downstream traffic is approximately 90% at every hour of the day. This directional disparity is 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. As we show in the Appendix, these average

Figure 4: *Average Temporal Usage Patterns*



Note: This figure presents statistics on how usage is distributed throughout the day, on average, and the breakdown of usage by direction.

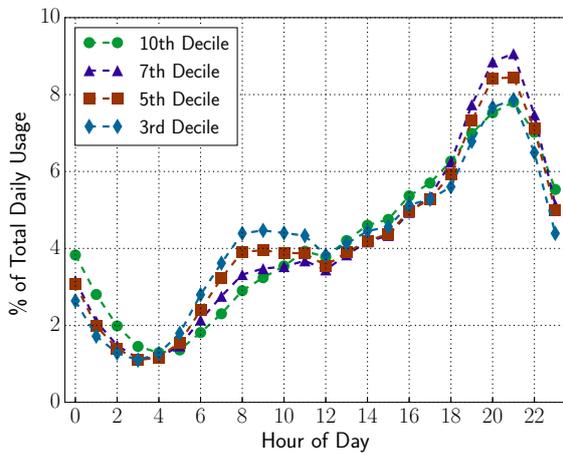
temporal patterns do not differ substantially by the day of the week.

We find 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 consumers, where consumers are assigned to deciles based on their total usage. The heaviest-usage consumers (10th decile) have only a slightly flatter profile throughout the day, revealing a 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.

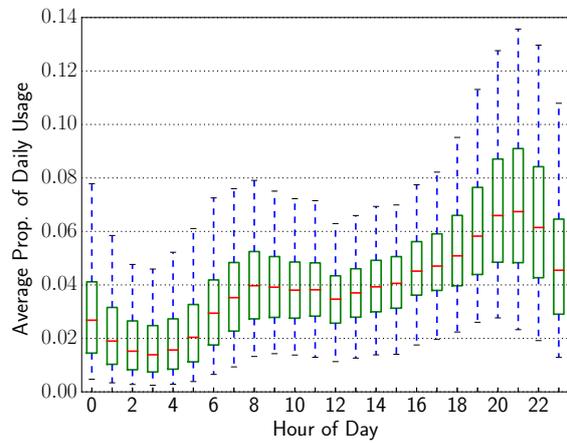
For each hour of the day, Figure 5(b) presents the distribution across consumers of the proportion of usage during that hour. For example, during the 9PM hour 50% (95%) of people have average usage that is less than 6.5% (13.7%) of their average daily usage over our panel (line within the box represents median). The box and whiskers capture the interquartile range (25th and 75th) and the 5th and 95th quantiles, respectively. The dispersion at every hour is indicative of substantially different temporal usage patterns across consumers, albeit not correlated with the consumer’s overall 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 either that during peak-demand times the network is

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 variation in temporal patterns in usage across consumers. Panel (a) reports hourly percentages for deciles 3, 5, 7, and 10, where the deciles are calculated using each consumer’s total usage during the panel. For example, the 10th decile includes consumers in the top 10% of all consumers in terms of average monthly usage. In panel (b), we report variation in the temporal profile across consumers. Specifically, for each consumer we calculate the proportion of their overall traffic during each hour of the day, and then plot the distribution of these proportions at each hour.

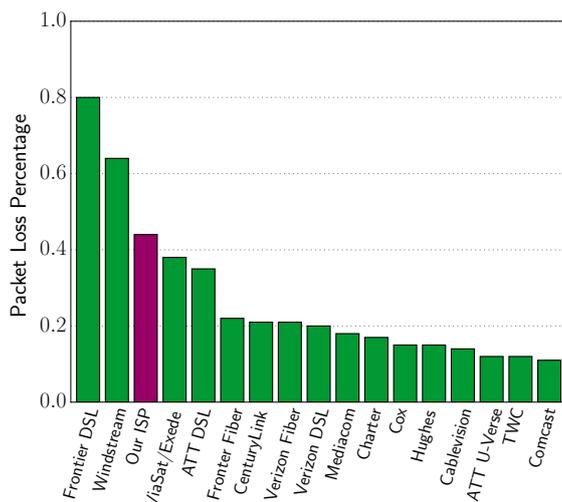
extremely congested, or there is a large amount of excess capacity that sits idle for the remainder of the day.

2.4 Network Congestion

In this subsection we demonstrate the extent of congestion in the network and how congestion impacts usage. We first show the link between congestion and packet loss, a measure of network performance. We also present the sources of variation in congestion in the ISP’s network. Second, we provide direct evidence of the effect of congestion on usage, relying on variation in congestion due to network upgrades.

Congestion and Packet Loss

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

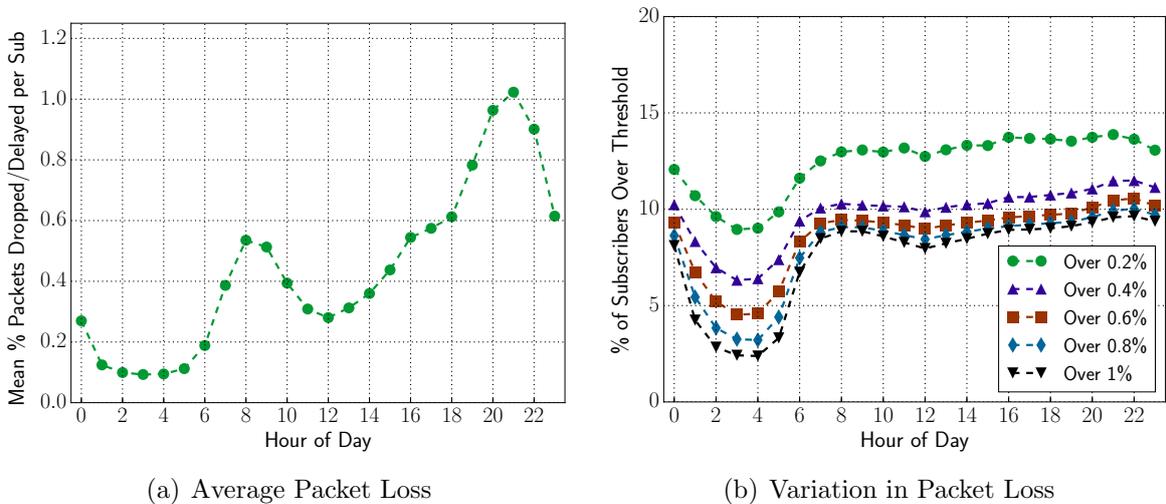


Note: This figure is a reproduction from the FCC’s 2015 Measuring Broadband America Fixed Report, We calculate the the average of hourly packet loss percentages in our sample, which we believe is the most comparable to the FCC statistics, and include it in the figure.

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 repeatedly. 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 is the the hourly proportion of packets dropped/delayed, which we, and others, refer to as packet loss. We use hourly packet loss to measure congestion because it is an individual measure of quality, instead of an aggregate one, and thus better reflects the performance each subscriber faces.

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.⁷ 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).

Figure 7: *Average Hourly Subscriber Packet Loss*

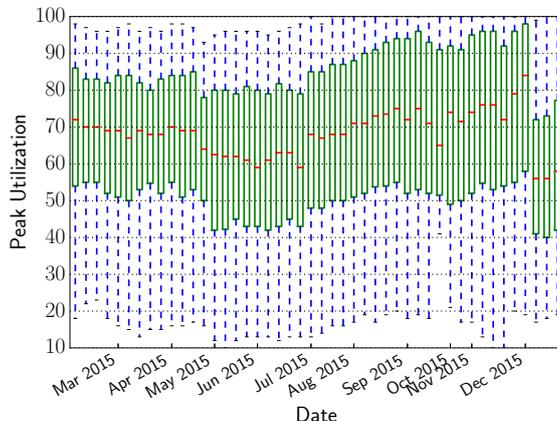


Note: These figures report statistics of average packet loss by hour. For each subscriber in the sample, we calculate hourly packet loss across the panel to generate these figures. In panel (a), we report average hourly packet loss for all subscribers. Panel (b) reports the percentage of packet loss 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.

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 by reporting the percentage of subscribers 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.

⁷ From the FCC report, it is not clear exactly how their statistics are calculated. We believe that the statistics are the average of hourly packet loss percentages from a specific test conducted by the modem.

Figure 8: *Weekly Node Utilization Statistics*



Note: This figure presents the weekly variation in peak utilization of network nodes. The box represents the inter-quantile range in each week, the dashed line is the median, and the vertical dashed lines extend to the 5th and 95th percentiles.

Therefore, the majority of people experience little packet loss over the day, but in some cases, packet loss is 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 viewing. However, if someone is browsing a website, dropping a single particular packet could be the difference in the website loading 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.

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 network 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.⁸ When such a change is made,

⁸This 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*

| | <i>Before</i> | <i>After</i> | <i>Diff</i> | <i>% Change</i> |
|-------------------------|---------------|--------------|-------------|-----------------|
| 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% |
| % Usage During Off-Peak | 29.4% | 27.1% | -2.3 | -7.8% |
| 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. Seven days of data are taken from before and after the node split date to calculate means. These averages are at the node level and are weighted by the number of people on the node.

the network state for the affected subscribers should be improved because there are half as many subscribers sharing the same capacity. If subscriber behavior is responsive to such changes in network quality, we would expect an increase in usage overall, particularly at peak hours.

There are several distinct node splits of this nature 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 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%. Together these imply a 7.8% decrease in the percent of usage that is off-peak. This suggests that there is some degree of unmet demand prior to the node split that is now able to be realized, and provides some evidence of intra-day substitution to avoid congestion during peak hours.⁹

The co-variation of congestion and the share of consumption during off peak is what ultimately identifies the intra-day elasticity of substitution. The analysis of node splits suggests this elasticity is non trivial. A similar conclusion is reached when we regress the ratio of off-peak to total usage on measures of node congestion. We find that more congested nodes have a higher share of off-peak usage and the magnitude in the cross-

⁹Note that the increase in usage could come from a change in subscriber behavior, or bandwidth-adaptive applications becoming more responsive. In further decomposing the usage response, we find that nearly 90% of the 7-day response occurs after 48 hours when consumers are aware of the upgrade, as where an adaptive response would be immediate. Additionally, the moderate drop in usage during off-peak hours is inconsistent with an adaptive response.

Table 3: *Percent Usage by Application*

| Groups | <i>Description (Examples)</i> | <i>% of All Usage</i> |
|----------------|---|-----------------------|
| 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 |

Note: This table reports the composition of aggregate traffic for different types of applications.

section is comparable to what we see from the node splits.

2.5 Composition of Usage and External Validity

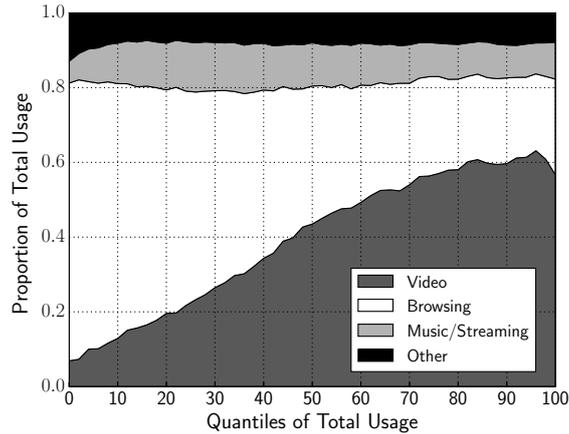
Our main data source, discussed above and used for our analysis, does not provide information on the types of online activities subscribers engage in, only the volume and timing. We do, however, have data on the composition of usage from another ISP’s network. The data are from over 500,000 customers, national in scope, and from the same period, February – December 2015. We do not use these data to estimate the model, 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 the model estimates.

Table 3 presents the percentage of usage for each of several categories. We find that video and music 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 file-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 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).

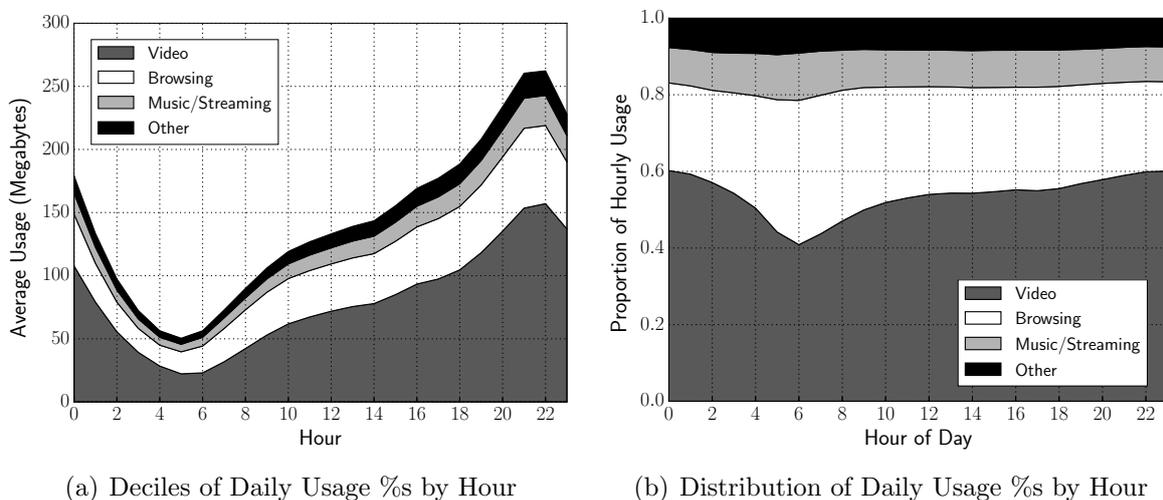
Figure 10(a) and 10(b) present average usage by hour and traffic type, and the proportion that each traffic type accounts for at each hour, respectively. The overall temporal pattern in average usage is nearly identical to the pattern from the data above that is used for our analysis. Video is the most 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 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 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. In contrast, we explicitly incorporate network congestion into the model, and we disaggregate the daily

Figure 10: *Data: Hourly Usage by Group*



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 panel (b), we report the distribution across all consumers of these proportions at each hour.

usage decision into 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, $t = 1, \dots, T$, a subscriber chooses the amount of content to consume during off-peak hours and then peak hours on their chosen service plan, $k = 1, \dots, K$. Plans are characterized by a provisioned speed at which content is delivered in the absence of congestion (s_k), a usage allowance (\bar{C}_k), a fixed fee that pays for all usage up to the allowance (F_k), and an overage price per GB of usage in excess of the allowance (p_k). The menu of plans, and the characteristics of each, are fixed.¹⁰ 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, dG_ψ , which we will estimate from the data and present in the Appendix.

Utility from content is additively separable over all days in the billing cycle, and across

¹⁰Plans 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 ended.

billing cycles.¹¹ Let consumption of content during peak and off-peak hours be denoted by c^p and c^{op} , respectively. Consumption of the numeraire good is given by y . The utility for a subscriber of type h on plan k is given by

$$u_{hk}(c^p, c^{op}, \psi, v) = v_1 \left(\frac{(c^{op} + c^p)^{1-\alpha_h}}{1-\alpha_h} \right) - (c^{op})^2 \left(\frac{v_2 \kappa_h}{\ln(s_k)} \right) - (c^p)^2 \left(\frac{\kappa_h}{\ln(\psi s_k)} \right) + y.$$

The first term captures the subscriber's utility from consuming the content. Marginal utility is declining, as we expect the first unit of any activity (email, web browsing, video, etc.) to bring higher marginal utility than subsequent usage. The convexity of the utility function is quite flexible, nesting everything between log ($\alpha_h \rightarrow 1$) 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, v_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 consumption. 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 $c^{op} \left(\frac{2v_2 \kappa_h}{\ln(s_k)} \right)$. 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 v_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 ex-

¹¹Nevo et al. (2016) provide evidence that content is likely not substitutable across days and billing cycles.

ample, 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 v_1 and v_2 are independent normal random variables with means equal to μ_{1h} and μ_{2h} , respectively, and a common coefficient of variation, ρ_h . The vector of parameters, $(\alpha_h, \kappa_h, \mu_{1h}, \mu_{2h}, \rho_h)$, describes the preferences of a subscriber of type h .

3.2 Optimal Usage

A consumer solves a finite-horizon dynamic programming problem within each billing cycle.¹² 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 v_{1t} and v_{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, conditional on choosing plan k , this subscriber's problem is

$$\begin{aligned} \max_{\{c_t^p, c_t^{op}\}_{t=1, \dots, T}} \quad & \sum_{t=1}^T E_{(\psi, v)} [u_{hk}(c_t^p, c_t^{op}, \psi, v)] \\ \text{s.t.} \quad & F_k + p_k \text{Max}\{C_T - \bar{C}_k, 0\} + Y_T \leq I, \quad C_T = \sum_{j=1}^T (c_j^{op} + c_j^p), \quad Y_T = \sum_{j=1}^T y_j. \end{aligned}$$

We do not discount future utility since we model daily decisions, over a finite and short horizon. Uncertainty involves the realizations of v and ψ . 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 (v_T). Depending on the values of c_T^{op} , ψ_T , and v_{1T} , the consumer will either consume a satiation level of utility such that $\frac{\partial u_{hk}(c_T^p, c_T^{op}, \psi_T, v_T)}{\partial c^p} = 0$, the remaining portion of their allowance such that

¹²The 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.

$\bar{C}_k = C_T$, or incur overages such that $\frac{\partial u_{hk}(c_T^{op}, c^p, \psi_T, v_T)}{\partial c^p} = p_k$. Denote this optimal level of consumption, or the policy function on day T during peak hours, as $c_{hkT}^{p*}(c_T^{op}, C_{T-1}, \psi_T, v_T)$ or more compactly, c_{hkT}^{p*} .

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

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

where $\mathcal{O}_{tk}(c) = \operatorname{Max}\{c - \bar{C}_{kt}, 0\}$, i.e., usage in excess of the remaining allowance ($\bar{C}_{kt} = \operatorname{Max}\{\bar{C}_k - C_{t-1}, 0\}$). The expectation is only over ψ (which impacts the optimal peak-usage policy) because v_{1T} and v_{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

$$E_{(\psi, v)} [V_{hkT}(C_{T-1}, \psi_{T-1})] = \int_{\psi} \left[\int_v V_{hkT}(C_{T-1}, \psi_{T-1}, \psi, v) dG_v^h(v) \right] dG_{\psi}(\psi | \psi_{T-1}),$$

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

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

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

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

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

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 the subscriber selects one of the offered plans ($k \in \{1, \dots, K\}$) or no plan ($k = 0$) such that

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

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. Previous work has studied whether consumers make what look like suboptimal choices ex post in a variety of settings.¹³ In our data, it is not obvious that subscribers systemically make mistakes. 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.

¹³In particular, some research has highlighted what seem like suboptimal choices made by consumers facing non-linear pricing, similar to ours, in cell phone usage (Grubb and Osborne (2015)) and health care (Abaluck and Gruber (2012); Handel (2013)). In contrast, several papers (e.g., Miravete (2003); Economides et al. (2008); Goettler and Clay (2011); Ketcham et al. (2012)) highlight circumstances where individuals make choices that are rational ex post.

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, $c_{hk^*t}^{op*}$ and $c_{hk^*t}^{p*}$, 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 each household belongs to exactly one type and that with enough data this type will be revealed. Then the distribution of types is calculated by aggregating households. 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 unlimited data, the distribution of types for each household will not be degenerate. Since ultimately the aggregate distribution of types is of interest, we do not try to distinguish between these two views.

Our estimation approach is a panel-data modification of Akerberg (2009), Fox et al. (2011), 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 type is defined by a vector of the parameters. We then estimate the distribution of types by matching patterns from the data to those predicted by the behavior of the types.

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, (v_1, v_2) . The relative likelihood values over the H_k candidate types for each consumer are 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 Likelihood Function

For each individual on plan k , $i = 1 \dots I_k$, we have a series, $m = 1 \dots M$, which captures usage at a daily frequency on the chosen plan. This includes both off-peak usage, $(c_{i1}^{op}, \dots, c_{iM}^{op})$, and peak usage, $(c_{i1}^p, \dots, c_{iM}^p)$. 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}).

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, $(\alpha_h, \kappa_h, \mu_{1h}, \mu_{2h}, \rho_h)$. The solution to the model yields state-dependent policy functions for peak and off-peak usage on each type's optimal plan, $c_{hkt}^{op*}(C_{t-1}, \psi_{t-1}, v_t)$ and $c_{hkt}^{p*}(c_t^{op}, C_{t-1}, \psi_t, v_t)$, respectively. Additionally, the parametric structure of the model (distribution of v), yields probabilities associated with any realization of the policy functions, $\mathbb{P}[c_{hkt}^{op*}(C_{t-1}, \psi_{t-1}, v_t) = c]$ and $\mathbb{P}[c_{hkt}^{p*}(c_t^{op}, C_{t-1}, \psi_t, v_t) = c]$, conditional on the observed states. These probabilities, together with the Markov process for ψ (dG_ψ) that is recovered directly from the data (and presented in the Appendix), 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 , $h = 1, \dots, H_k$. Further, consider one consumer's sequence of usage on plan k , $(c_{i1}^{op}, \dots, c_{iM}^{op})$, and $(c_{i1}^p, \dots, c_{iM}^p)$, and the values of the associated observed states $\{t_m, \psi_m, C_{t_m}\}_{m=1, \dots, M}$. The likelihood that candidate type h generated this sequence of data, conditional on the observed states, equals

$$\mathcal{L}_{ih} = \prod_{m=1}^M \mathbb{P}\left[c_{hkt}^{op*}(C_{m-1}, \psi_{m-1}, v_m) = c_{ikm}^{op}\right] \times \mathbb{P}\left[c_{hkt}^{p*}(c_m^{op}, C_{m-1}, \psi_m, v_m) = c_{ikm}^p\right] \times dG_\psi(\psi_m | \psi_{m-1}).$$

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

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,

$$\mathbb{P}_{ih} = \frac{\mathcal{L}_{ih}}{\sum_{j=1}^{H_k} \mathcal{L}_{ij}}.$$

To form estimates of the posterior probabilities, $\widehat{\mathbb{P}}_{ih}$, we calculate probabilities from the solution to the model via simulation over v , i.e., $\widehat{\mathbb{P}}\left[c_{hkt}^{op*}(C_{m-1}, \psi_{m-1}, v_m) = c_{ikm}^{op}\right]$ and $\widehat{\mathbb{P}}\left[c_{hkt}^{p*}(c_m^{op}, C_{m-1}, \psi_m, v_m) = c_{ikm}^p\right]$, respectively, and use the same estimate of the Markov process for the network state, \widehat{dG}_ψ , 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 \mathbb{S}_k is the proportion of individuals selecting plan k . The summation is only over $i = 1, \dots, I_k$ 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 ($r = 1, \dots, R$) perform the estimation using samples created through sampling with replacement, calculating the standard errors directly from the distribution of estimates.

4.2 Identification

Our estimation approach uses two choices: the 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, and not just aggregate moments.

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 mass 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, as 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.

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 decisions are revealing about the individual's preferences. Fixing the shadow price of usage at zero, the responsiveness of c_t^p to variation in the the network state identifies the slope of 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 \bar{C}_k , it will also increase its slope (like congestion). For a fixed network state (ψ), the mean and variance of c^p and c^{op} identify the parameters of the v_1 distribution. Finally, the distribution of v_2 that determines the relative marginal cost of usage during off-peak and peak periods is identified by the co-variation between the the share of usage which is off-peak and congestion.

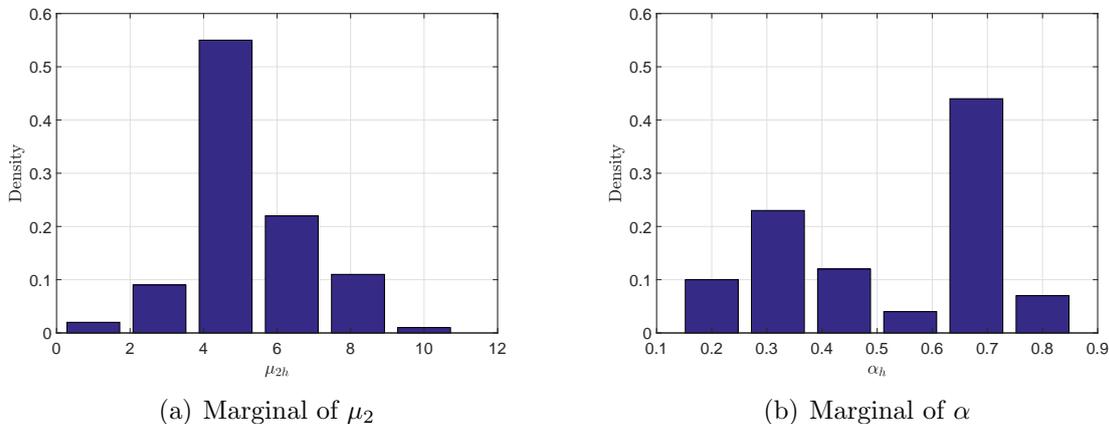
5 Results

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. 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 type 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.

Figure 11: *Marginal Distribution of μ_2 and α*



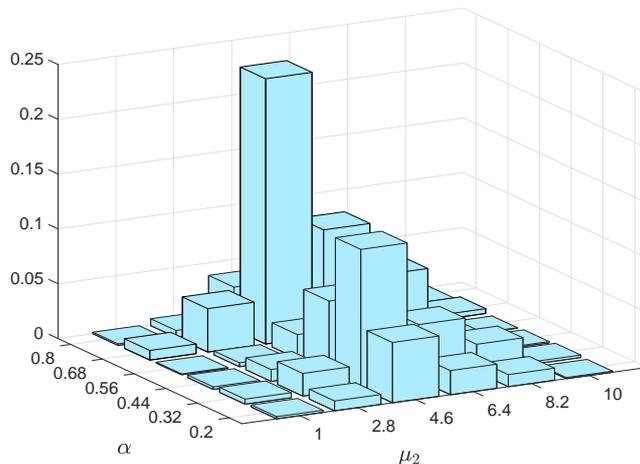
Note: The figures are the estimated marginal distributions for μ_2 and α , respectively.

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 in virtually no redistribution of mass outside the chosen support.¹⁴ 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 similar patterns in how usage is distributed between peak and off-peak periods, but there is still substantial heterogeneity around that mode. Consistent with this observation from our data, we observe a pronounced modal value for μ_2 , while the heterogeneity in temporal usage patterns is also captured with some types with $\mu_2 \approx 1$ (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 distri-

¹⁴For 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.

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



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

butions are quite irregular. Figure 12 gives the joint distribution of μ_2 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 parameters rather than the normality and lack of covariance often assumed 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, both of which have direct policy implications.

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),

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

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 an 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.¹⁵

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 encoding and compression technologies continue to reduce requirements, which suggests that a home with 100 Mb/s 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.

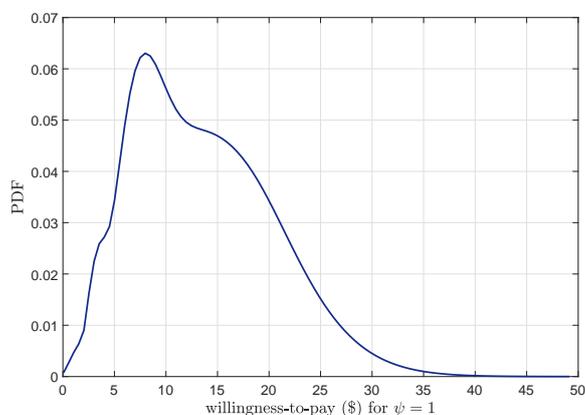
Similarly to preferences for speed, our model yields an estimate of the value of one more GB added to the usage allowance on each type’s optimal plan. This value equals

$$\frac{\partial E_{(\psi,v)} [V_{hk1}(C_1 = 0, \psi = 1)]}{\partial \overline{C}_{h_k^*}},$$

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),

¹⁵This 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”.

Figure 13: *PDF of willingness-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.

suggests that current usage allowances are well ahead of most consumers' 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.¹⁶

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 \$0.14. This is quite small, because even if you substantially decrease the usage allowances, the marginal value does not 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-

¹⁶The 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.

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.¹⁷ 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 by approximately 40% annually.¹⁸ 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.

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 level of congestion and resolve the consumer problem, which would imply a new congestion, and so forth until usage levels and congestion were consistent with each other. This involves iterating between a dynamic problem for a large number of consumer types, and an equilibrium calculation of congestion, which computationally is not feasible. For the calculations below we focus on the equilibrium aspect and simplify the dynamics when solving the counterfactuals.¹⁹ Specifically, we

¹⁷As a sanity check on this number, we can compare it to the cost of providing a less congested network. If we take 200 consumers and put them on a node, they will be congestion free in the short run. Our estimates suggest that in aggregate these consumers will be willing to pay roughly \$3,000 a month for this upgrade. The cost of installing a new node varies widely, but \$50,000 is a reasonable “average” number. This cost is equal to roughly a year and a half of what the consumers are willing to pay. Given current usage growth rates this seems in the ballpark of the time it would take the new node to get congested.

¹⁸The 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.

¹⁹In an earlier version of this paper we went to the other extreme, where we resolved the new dynamic

provide a baseline simulation from the model where we simplify the consumer’s problem to the decision of how to allocate usage within a single day with a proportional reduction of the monthly usage allowance. 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.

If the daily usage allowance is not binding, each of the subscribers’ peak and off-peak usage satisfies

$$v_1 (c^{op} + c^p)^{-\alpha} = \frac{2c^{op}v_2\kappa_h}{\ln(s_k)}$$

$$v_1 (c^{op} + c^p)^{-\alpha} = \frac{2c^p\kappa_h \ln(\psi s_k) - (c^p)^2\kappa_h \frac{1}{\psi} \frac{\partial \psi}{\partial c^p}}{\ln^2(\psi s_k)},$$

where v is a realization of the preference shocks. If the usage allowance is binding, the overage price is added to the right-hand side of both equations. 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 v 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 the differential tradeoffs introduced by each of the alternatives.

Throttling Speed

One way to deal with congestion is to lower realized speeds during peak hours of consumers who have exceeded their allowance. This is the way that overages are implemented on some cellular networks. To simulate the effect of such a policy, we slow consumers down during peak-use periods after exceeding their usage allowance instead of 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 prefer the penalty of a low speed than having to pay the per-GB charge. Their additional usage improves consumer welfare, while there is only a slight decrease in ISP revenue due to the absence of overage charges. The results when speed is throttled to 3 Mb/s are similar. Overall, throttling does not seem to achieve the goal of shifting demand to off-peak hours for residential broadband networks.

problems but held congestion fixed rather than solving for the equilibrium value. The results were qualitatively similar.

Table 4: *Counterfactual: Throttling*

| | <i>Baseline</i> | <i>7 Mb/s (s_k)</i> | <i>3 Mb/s (s_k)</i> |
|---------------------|-----------------|----------------------------------|----------------------------------|
| Daily Usage (GB) | 2.6 (0.03) | 2.9 (0.03) | 2.8 (0.03) |
| Peak Usage (GB) | 1.9 (0.02) | 2.0 (0.02) | 1.9 (0.02) |
| Off-Peak Usage (GB) | 0.7 (0.01) | 0.9 (0.01) | 0.9 (0.01) |
| Consumer Surplus | 68.31 (0.07) | 72.40 (0.07) | 73.02 (0.07) |
| Revenue | 59.75 (0.06) | 57.09 (0.06) | 58.18 (0.06) |

Note: This table reports how subscriber behavior changes under throttling of connection speeds relative to a baseline simulation with the current usage-based pricing schedules. Standard errors are calculated using the block-resampling procedure described in Section 4.

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. For this to have any effect, we need to also decrease the allowance. The results are in Table 5. The first column provides the baseline using the current usage-based pricing schedules. The second and third columns consider 25% and 50% reduction in the baseline allowance when only peak usage is counted against that allowance. Consistent with the results from the analysis of node splits, we find that peak usage responds to a higher price whether it be in the form of overages or congestion, and off-peak usage decreases relatively less. 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. However, we find that the willingness to pay for an extra GB during peak periods 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 transferable to off-peak hours. One way to accomplish that is the introduction of local-caching technologies.

Local-Cache Technology

OTTV services account 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

Table 5: *Counterfactual: Peak-Use Pricing*

| | <i>Baseline</i> | <i>25% Reduction (\bar{C}_k)</i> | <i>50% Reduction (\bar{C}_k)</i> |
|---------------------|-----------------|---|---|
| Daily Usage (GB) | 2.6 (0.03) | 2.7 (0.03) | 2.4 (0.03) |
| Peak Usage (GB) | 1.9 (0.02) | 1.8 (0.02) | 1.7 (0.02) |
| Off-Peak Usage (GB) | 0.7 (0.01) | 0.9 (0.01) | 0.7 (0.01) |
| Consumer Surplus | 68.31 (0.07) | 69.06 (0.07) | 65.24 (0.07) |
| Revenue | 59.75 (0.06) | 58.11 (0.06) | 60.22 (0.06) |

Note: This table reports how subscriber behavior changes under peak-use pricing relative to a baseline simulation with the current usage-based pricing schedules. Standard errors are calculated using the block-resampling procedure described in Section 4.

Table 6: *Counterfactual: Local-Caching Technology*

| | <i>Baseline</i> | <i>25% Reduction (μ_2)</i> | <i>50% Reduction (μ_2)</i> |
|-----------------------|-----------------|---|---|
| Daily Usage (GB) | 2.6 (0.03) | 2.8 (0.03) | 3.2 (0.03) |
| Peak Usage (GB) | 1.9 (0.02) | 1.8 (0.02) | 1.7 (0.02) |
| Off-Peak Usage (GB) | 0.7 (0.01) | 1.0 (0.01) | 1.3 (0.02) |
| Consumer Surplus (\$) | 68.31 (0.07) | 75.88 (0.08) | 81.70 (0.09) |
| Revenue (\$) | 59.75 (0.06) | 60.97 (0.06) | 63.01 (0.06) |

Note: This table reports how subscriber behavior changes with the introduction of local-caching technologies relative to a baseline simulation with the current usage-based pricing schedules. Standard errors are calculated using the block-resampling procedure described in Section 4.

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 (e.g., Netflix’s phone app 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 (μ_2). In Table 6, we provide estimates of the effect if local-cache technologies were to reduce the cost of off-peak usage by 25% and 50% for all consumer types, respectively. This of course may substantially understate its effect, as the heaviest of users consume much more OTTV services and would benefit disproportionately 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 or avoid by complementing the technology with peak-use pricing.

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 Akerberg (2009), 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 strategies 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. 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 may 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 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

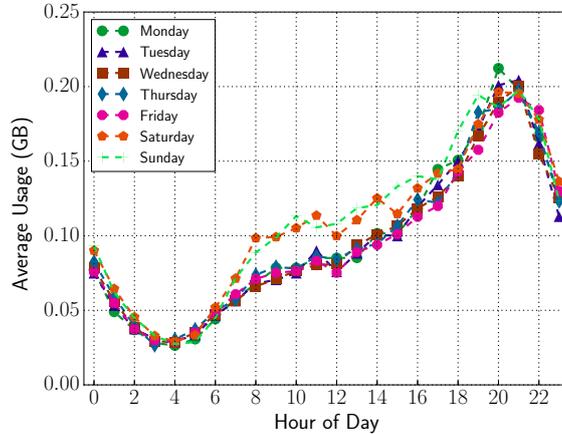
- Abaluck, Jason and Jonathan Gruber (2012). “Choice Inconsistencies Among the Elderly: Evidence from Plan Choice in the Medicare Part D Program.” *American Economic Review*, 101(4): 1180–1210.
- Ackerberg, Daniel A (2009). “A New Use of Importance Sampling to Reduce Computational Burden in Simulation Estimation.” *Quantitative Marketing and Economics*, 7(4): 343–376.
- Anderson, Laura M, Lars Gårn Hansen, Carsten Lynge Jensen and Frank A Wolak (2017). “Using Real-Time Pricing and Information Provision to Shift Intra-Day Electricity Consumption: Evidence from Denmark.” *Department of Economics, Stanford University*.
- Aron-Dine, Aviva, Liran Einav, Amy Finkelstein and Mark Cullen (2015). “Moral Hazard in Health Insurance: How Important Is Forward-Looking Behavior?” *Review of Economics and Statistics*, 97(4): 725–741.
- Chevalier, Judith and Austan Goolsbee (2009). “Are Durable Goods Consumers Forward-Looking? Evidence from College Textbooks.” *Quarterly Journal of Economics*, 124(4): 1853–1884.
- Chung, Doug, Thomas Steenburgh and K Sudhir (2010). “Do Bonuses Enhance Sales Productivity? A Dynamic Structural Analysis of Bonus-Based Compensation.” *HBS Working Paper 11-041*.
- Copeland, Adam and Cyril Monnet (2009). “The Welfare Effects of Incentive Schemes.” *Review of Economic Studies*, 76(1): 93–113.
- Crawford, Gregory and Matthew Shum (2005). “Uncertainty and Learning in Pharmaceutical Demand.” *Econometrica*, 73(4): 1137–1173.
- Dutz, Mark, Jonathan Orszag and Robert Willig (2009). “The Substantial Consumer Benefits of Broadband Connectivity for US Households.” *Internet Intervention Alliance Working Paper*.
- Economides, Nicholas, Katja Seim and Brian Viard (2008). “Quantifying the benefits of entry into local phone service.” *The RAND Journal of Economics*, 39(3): 699.

- 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.
- Einav, Liran, Amy Finkelstein and Paul Schrimpf (2015a). “Bunching at the Kink: Implications for Spending Responses to Health Insurance Contracts.” *Working Paper*.
- Einav, Liran, Amy Finkelstein and Paul Schrimpf (2015b). “The Response of Drug Expenditure to Non-Linear Contract Design: Evidence from Medicare Part D.” *Quarterly Journal of Economics*, 130(2): 841–899.
- Fauqui, Ahmad and Sanem Sergici (2010). “Household Response to Dynamic Pricing of Electricity: A Survey of 15 Experiments.” *Journal of Regulatory Economics*, 38(2): 193–225.
- Fox, Jeremy T, Kyoo il Kim and Chenyu Yang (2016). “A Simple Nonparametric Approach to Estimating the Distribution of Random Coefficients in Structural Models.” *Journal of Econometrics*, 195(2): 236–254.
- Fox, Jeremy T, Stephen P Ryan and Patrick Bajari (2011). “A Simple Estimator for the Distribution of Random Coefficients.” *Quantitative Economics*, 2(3): 381–418.
- Goettler, Ronald and Karen Clay (2011). “Tariff Choice with Consumer Learning and Switching Costs.” *Journal of Marketing Research*, 48(4): 633–652.
- Goetz, Daniel (2016). “Competition and Dynamic Bargaining in the Broadband Industry.” *Princeton University Working Paper*.
- 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.
- Gowrisankaran, Gautam, Stanley S Reynolds and Mario Samano (2016). “Intermittency and the Value of Renewable Energy.” *Journal of Political Economy*, 125: 1187–1234.
- Gowrisankaran, Gautam and Marc Rysman (2012). “Dynamics of Consumer Demand for Durable Goods.” *Journal of Political Economy*, 120(6): 1173–1219.
- Greenstein, Shane and Ryan McDevitt (2011). “The Broadband Bonus: Estimating Broadband Internet’s Economic Value.” *Telecommunications Policy*, 35(7): 617–632.

- Greenstein, Shane, Martin Peitz and Tommaso Valletti (2016). “Net Neutrality: A Fast Lane to Understanding the Tradeoffs.” *Journal of Economic Perspectives*, 30(2): 127–150.
- Grubb, Michael (2015). “Consumer Inattention and Bill-Shock Regulation.” *Review of Economic Studies*, 82(1): 219–257.
- Grubb, Michael and Matthew Osborne (2015). “Cellular Service Demand: Biased Beliefs, Learning, and Bill Shock.” *American Economic Review*, 105(1): 234–271.
- Handel, Benjamin (2013). “Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts.” *American Economic Review*, 103(7): 2643–2682.
- Hendel, Igal and Aviv Nevo (2006a). “Measuring the Implications of Sales and Consumer Inventory Behavior.” *Econometrica*, 74(6): 1637–1673.
- Hendel, Igal and Aviv Nevo (2006b). “Sales and Consumer Inventory.” *The RAND Journal of Economics*, 37(3): 543–561.
- Hitte, Loran and Prasanna Tambe (2007). “Broadband Adoption and Content Consumption.” *Information Economics and Policy*, 74(6): 1637–1673.
- Ito, Koichiro (2014). “Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing.” *American Economic Review*, 104(2): 537–563.
- Ketcham, Jonathan D, Claudio Lucarelli, Eugenio J Miravete and M Christopher Roebuck (2012). “Sinking, Swimming, or Learning to Swim in Medicare Part D.” *American Economic Review*, 102(6): 2639–2673.
- 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.
- Miravete, Eugenio (2003). “Choosing the Wrong Calling Plan? Ignorance and Learning.” *American Economic Review*, 93(1): 297–310.

- Misra, Sanjog and Harikesh Nair (2011). “A Structural Model of Sales-Force Compensation Dynamics: Estimation and Field Implementation.” *Quantitative Marketing and Economics*, 9(3): 211–257.
- Nevo, Aviv, John Turner and Jonathan Williams (2016). “Usage-Based Pricing and Demand for Residential Broadband.” *Econometrica*, 84(2): 411–443.
- Newsham, Guy R and Brent G Bowker (2010). “The Effect of Utility Time-Varying Pricing and Load Control Strategies on Residential Summer Peak Electricity Use: A Review.” *Energy Policy*, 38(7): 3289–3296.
- 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.
- Strapp, James, Chris King and Sharon Talbott (2007). “Ontario Energy Board Smart Price Pilot Final Report.” *IBM Global Business Services and eMeter Strategic Consulting for the Ontario Energy Board*.
- Varian, Hal (2002). *The Demand for Bandwidth: Evidence from the INDEX Experiment*, volume Broadband: Should We Regulate High-Speed Internet Access? Brookings Institution.
- Wilson, Kyle (2017). “Does Public Competition Crowd Out Private Investment? Evidence from Municipal Provision of Internet Access.” *Working Paper*.
- Wolak, Frank A (2007). “Residential Customer Response to Real-Time Pricing: The Anaheim Critical Peak Pricing Experiment.” *Center for the Study of Energy Markets*.
- Wolak, Frank A (2010). “An Experimental Comparison of Critical Peak and Hourly Pricing: The PowerCentsDC Program.” *Department of Economics, Stanford University*.
- Wolak, Frank A (2016). “Designing Nonlinear Price Schedules for Urban Water Utilities to Balance Revenue and Conservation Goals.” *National Bureau of Economic Research*.
- Yao, Song, Carl Mela, Jeogwen Chiang and Yuxin Chen (2012). “Determining Consumers’ Discount Rates with Field Studies.” *Journal of Marketing Research*, 49(6): 822–841.

Figure 14: *Day-of-Week Variation in Temporal Usage*



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

Appendix

In this Appendix we present some additional descriptive statistics and results.

Intra-Day Usage by Day of the Week

We find that average temporal patterns do not differ substantially by the day of the week. Figure 14 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.

Markov Process Governing the Network State

In Table 7 we present the Markov process we use for the network state. Packet loss is split into seven bins that are used to study how persistent packet loss is day-to-day. 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. This matrix is used to solve the model. For the estimation procedure, all we need are day-hour observations of daily

Table 7: *Transition Matrix of Peak Packet Loss*

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

| | <i>Mean</i> | <i>Median</i> | <i>Mode</i> |
|------------|-------------|---------------|-------------|
| 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.

consumption and the observed peak packet loss state for each account in the sample.

Statistics of the Distribution of Types

The marginal distribution for each of the parameters, calculated from the 732 candidate types receiving positive density, is summarized in Table 8. With the exception of α , each of the parameters has a mode equal to the median. However, for each of the parameters other than α , the mean is also 10% of more different than the median, capturing skew in the distributions.