When the Rich get Richer: A Case Study of High Speed Tier Usage Behavior

Sarthak Grover, Roya Ensafi and Nick Feamster
Princeton University

Paper #0, 7 Pages

Abstract

We present an analysis of usage patterns of Comcast subscribers whose service was increased from one high service tier to another. This study focuses on changes in usage behavior of two groups of users in the same city, a control group (105 Mbps tier link) and a treatment group (250 Mbps tier link). Members of the treatment group were randomly selected from the control group to receive 250 Mbps without their knowledge.

Previous work has shown that users who are already maximizing their usage on a given access link will continue to do so when they are migrated to a higher service tier. We study how users who are already on service plans with high downstream throughput respond when they are moved to a higher service tier without their knowledge. We find that subscribers who are already using most of their available capacity do not use significantly more capacity when they are moved to a higher service tier; in contrast, 50% of the subscribers who have low traffic demand increase their 95 percentile peak usage by 10% on upgrading the service plan. We also show that the median daily usage increases by 20% during off-peak hours.

1 Introduction

PARA 1: SEE OTHER USAGE CASE STUDIES ABOUT IMPORTANCE
(importance of usage. important for fcc. important for isps capacity planning. bottlenecks to usage can be the user themselves, the transit, or the service. how the fcc is considering usage based benchmarks so its important to understand who the heavy users are and when do they consume the most Sarthak: should this fcc be in the last intro para or the discussion section?)

PARA 2: lower tier has been studied (dasu, africa, etc). but higher tier what happens? what happens when people are already satisfied?
(previous study shows usage is tough to determine. choices are random too many variables [36]. dasu guys study usage with many factors by monitoring transfer bytes throughout the world. designed as a natural exp they show that apart from capacity, price and performance (latency pack loss) also play an imp role in determining usage. the law of diminishing returns when accounting for other factors.)

PARA 3: Our study
(we study the change in usage of subscribers who do not need any more bb are offered more without their knowledge by controlled exp. these 2219 dev are selected randomly from the same locations and same tier and ISP (keeping price and performance latency and plots similar). Compared with 18355 devices sharing the same demographics but getting what they paid for, hopefully their demands and choices are actually similar due to such control. Allows an ISP to plan for capacity based on trends in usage demand - if demand don't increase users are satisfied with high tier.)

PARA 4: Interesting results and contributions
(Our contributions show the complicated nature of individual usage and capacity, and motivate the need of demand based tiers and benchmarks in the future. We see off peak usage increases for users not utilizing the link anyway. We also seee that in our high speed tier dataset, prime time hours become late (8-12). Also peak usage increases for low util users daily. In the dataset over three months only 11 users out of 1500 actually went over 105 Mbps.) Sarthak: maybe just interesting results and not contributions

PARA 5: Higher perspective of interesting results and a line on policy makers
(The analysis of usage behavior to discuss different perspectives of bb utilization. We see that greedy rich users do not increase their usage so isps may conclude that they are satisfying their customers fully. But we also see that frugal rich users increase usage, especially in off peak, so conclude that as a consumer or policy maker capacity is still impacting perf. Both parties may have opposing views but look at the same observation. Motivates further study of the relationship between bb and capacity, esp. with similar control exp, and find out why this increase happened? Also look into usage demands as a metric used to determining broadband benchmarks.)

PARA 6: Roadmap of the paper
(section 2 background on previous usage work and industrial trends. section 3 data characterization and sanitization. section 4 empirical analysis looking at usage, peak usage, prime time, asymmetry, persistence and prevalence Sarthak: talk again a bit about usage - most interesting is how many

did NOT change behavior . section 5 is discussion on the differing perspectives of utilization (fcc vs isp) Sarthak: and different types of users taxonomy. and concluding on how fcc considering usage definitely needs more input from our community)

2 Background and Related Work

1. previous academic studies of usage - study low tier and natural experiments. Our approach is highly controlled experiments

2. industrial trends like sandvine and the FCC studies of usage and reports [4]

3. Last para: no one Broadband analysis has recently attracted much attention from the research community and the general public given its important business and policy implications. A number of efforts have focused on characterizing the availability and performance of broadband services around the world [1, 2, 5, 12, 20, 28, 31, 33]. The focus of our work is on exploring broadband services in their broader context, evaluating the complex interplay between broadband service characteristics, their market features and user demand. Different aspects of the complex interplay between user behavior, network services and operation has been explore in previous work. Some recent studies have examined the relationship between user behavior, network services and the providers. In Dobrian et al. [13] the authors show that poor connection quality can have a negative impact on a users quality of experience. Blackburn et al. [3] study how user behavior affects the economics of cellular operators. Chetty et al. [7] perform a user study to understand the effects of usage caps on broadband use. Other efforts have explored additional factors that may influence service demand, including the weather [6], service capacity [36] and the type of region [8].

(instead of this say we did it) The difficulty or outright impossibility of conducting controlled, randomized experiments of user behavior at Internet scale has been pointed out before. In his SIGCOMM 2011 Award presentation, Vern Paxson pointed to this issue and suggested the use of natural experiments to explore potential causal relationships with observational data. In a recent paper, Krishnan and Sitaraman [21] explore the use of related quasi-experimental design (QED) to evaluate the impact of video stream quality on viewer behavior and Oktay et al. [24] relies on it for causal analysis of user behavior in social media. We opted for natural experiments, rather than QED, as we consider the control and treatment groups to be sufficiently similar to random assignment.

3 Data Source and Characterization

Introduce control and treatment terminology, sanitization, data characteristics and first view of per subscriber bw must ask comcast how many days it has been since the users tier upgrade - is it short term or long term (compared with the africa)

Our dataset consists of network usage byte counters reported by Comcast gateways every 15 minutes from October 1, 2014 to December 29, 2014. There are two sets of broadband tiers that were used to collect this data: *control*set, consisting of households with a 105 Mbps access link, and the *treatment*set, consisting of households that were paying for a 105 Mbps access link, yet were receiving 250 Mbps instead. Users in the test set were selected randomly and were not told that their access bandwidth has been increased. There were more than 15000 gateway devices in the control set, with varying usage over the three months, and about 2200 gateway devices in the test set. **confirm - these were reported by Comcast gateways right?**

3.1 Data Description

The raw data sets provided by Comcast consisted of the *treatment*set, and 8 separate *control*sets consisting of more that 15k unique households, over different date ranges within the three months. Each dataset contains the following relevant fields: Device ID, sample period time, service class, service direction, IP address, and the bytes transferred in the 15 minute sample slot, as described in table 1.

Field	Description
Device_number	Arbitrarily assigned CM device identifier
end_time	Fifteen minute sample period end time
cmts_inet	Cmts identifier (derived from ip address)
service_direction	1-downstream, 2-upstream
octets_passed	Byte count

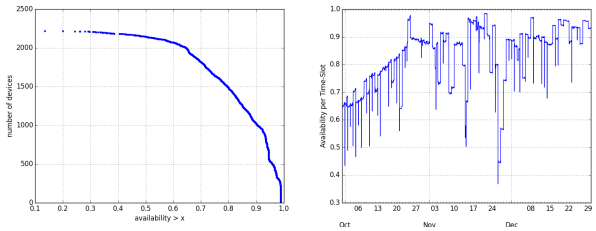
Table 1: *Field Descriptions for Comcast Dataset by Comcast*

3.2 Data Sanitization

Our initial analysis of data transferred per time slot showed that certain gateway devices were responsive only for brief periods. We also noticed that certain time slots had a very low response rate throughout the dataset. **Why? asked comcast - waiting for response** .

We evaluate the fraction of responsiveness of a gateway throughout the dataset, as well as the fraction of responsiveness per time slot, and call this the **heartbeat**. Figure 1 shows how the number of devices decreases for a higher heartbeat requirement. Based on the common trend of this plot throughout the *treatment*and *control*datasets, we decided to only choose gateway devices with an heartbeat of at least 0.8.

We sliced the sanitized *treatment*set based on the date range of each individual *control*set for comparison. We compared each of these tests individually to ensure that there are no outliers. We refer to the *treatment*and *control*sets in this case simply as datasets $set_1 - set_8$, where *set* is *treatment*or *control*. We also sliced and combined the sanitized data to give us *control*and *treatment*data for each month, referred to as set_{oct}, set_{nov}, set_{dec}. Finally, we combine all *control*sets to form a large concatenated dataset over the same date range as the complete *treatment*dataset, and we refer to this simply as set_{full}.



(a) *Heartbeat by device* **(b)** *Heartbeat by date*

Figure 1: *Heartbeat, based on gateway device responsiveness. (Make common eps plot of heartbeat – 8 control sets (half filtered) + test sets vs availability.)*

In the following analysis, we only present results for set_{full}, unless the behavior of an individual dataset varies significantly from the overall behavior and requires mention.

3.3 Relevance of the Data

REDO: not as limitations

In this section we describe how the Comcast database collected is both granular as well as unbiased. This database enables us to study usage behavior in a controlled setting. Beside, because of following properties, it is legitimized our use of it to compare and validate the *FCC*policy.

Study Byte Counters: The purpose of this work is to study the usage characteristics, irrespective of the application responsible for such usage. Limiting ourselves to just byte counters makes our analysis easily extendible to any ISP, and the FCC, interested in doing a similar study at a larger scale, without the risk of leaking PII. A study of applications has already been performed extensively by Sandvine [], as well as other researchers.

Granularity of 15 minutes: Broadband usage evaluated by commercial groups [], or governmental survey bodies, usually employed by the FCC, tends to focus on aggregated usage statistics over months, long term trends, and applications. In our work we specifically focus on data transferred in 15 minutes, to avoid short term bursts that max out the capacity, but account for long term heavy flows (such as real time entertainment and voip calls) that will continuously max out the access link. This gives us a granularity fine grained enough to study major changes in usage characteristics (such as peak trends) while ignoring short term bursts of traffic (such as browsing)

Note that byte counter readings collected every 15 minutes from multiple households were synchronized for consistency in measurements.

High Tier Measurements: We limit ourselves to analyzing usage patterns in the high capacity access link tier only. The *treatment*dataset was collected by increasing the capacity from 105 Mbps to to 250 Mbps for 2200 randomly selected users, without their knowledge. This served a two-fold purpose in avoiding biases that studies on usage and capacity suffer from: (a) *Avoid behavioral change bias:* offering users with high capacity a further increase without their knowledge avoids the risk of behavioral changes that may occur when one purposefully buys a higher bandwidth connection; and (b) *Avoid frustrated user bias:* users already have a high capacity that gets upgraded, instead of opting for an upgrade because their previous capacity was insufficient for their usage. Studying datasets with these biases will always show a positive correlation between usage and capacity, and by examining a single high capacity tier, we avoid this.

Single ISP, Same Location: No bias between service plans, pricing model, and traffic treatment. Controlled setting. Paths + performance should be similar and unbiased by the ISP as data is from one city. Also avoids local behavioral biases (if any). This gives us a highly controlled setting to study usage behaviors in an unbiased manner across a very large set of users (15k *control*and 1500 *treatment*households). Thus we believe that are conclusions will be representative of broadband behavior in a general American urban city. We expect the baseline behavior of all users to be similar, and in fact, interpret any differences between the *control*and *treatment*set behavior as aggregate changes that occurred due to the an increase in access link capacity.

4 Empirical Analysis

Sarthak: merge in a better manner: behavior, prime time, utilization, peak ratio, asymmetry, prevalence

We evaluate the usage behavior of the $treatment_{full}$[1] and the $control_{full}$[2] sets based on the criteria in table 2. We interpret the behavior of the datasets both separately as well as comparatively: (1) general inferences drawn from analyzing the dataset, and (2) comparative inferences drawn from observing changes in user behavior due to the upgrade in access link bandwidth (as explained in section 3.3)

First, we examine usage behavior and prime-time ratio as aggregates seen at the ISP. We use the total data usage per subscriber parameter to study these quantities. Next, we evaluate the utilization and peak time for each household in our dataset. We present the ISP perspective of utilization, and compare it to a users' perspective (FCC by proxy). We discuss a taxonomy of users based on their usage behavior and requirements.

Parameter	Definition	Agency
Prime $\text{Time}_{original}$	7:00 PM - 11:00 PM	FCC
Prime Time	8:00 PM - 12:00 AM	Authors
Prime Time Ratio	$\frac{\text{avg usage in peak (prime-time) hour}}{\text{avg usage in off-peak hour}}$	Sandvine
Peak Period	Time of network 95% of max	Sandvine
Peak Ratio	$\frac{\text{90\%-ile of max daily usage}}{\text{median of daily usage}}$	Authors
Usage per Subs.	$\frac{\text{aggregate data usage in time slot}}{\text{number of contributing subscribers}}$	Authors

Table 2: *Evaluation Criteria*

[1] household byte counters from (unknown) users with a 250 Mbps access link

[2] household byte counters from users with a 105 Mbps access link

4.1 Usage Behavior

In the next section, we analyze spacio-temporal network usage behavior: **Time Series Behavior (TS):** aggregating the usage (and utilization) per household over time (daily or weekly, per time slot). **Distribution Across Devices (CDF):** aggregating over time slots per day to measure utilization per device. We use the prime-time ratio and peak usage as criteria to evaluate usage behavior and interpret utilization.

To characterize diurnal user behavior as observed at the ISP, we first calculate usage per subscriber (table 2), and then plot the median and 90%-ile of total usage over a week for both *treatment*and *control*sets (figure 2).

We observe that the rise to the peak prime time hour usage on weekdays is not plateaued like the pattern observed on weekends (and holidays). A generic (median) weekday aggregate usage consists of a rise in usage that starts early in the morning that builds up to the prime-time period, peaks, and then falls sharply. We do not observe a trough in mid afternoon (between 2:00 PM – 6:00 PM), as is usually the case for overall usage observed at US Fixed access providers [**?**].

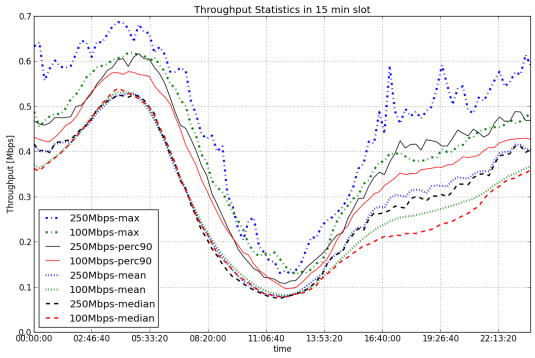


Figure 2: *agg (days) over means (devices): aggregate has no trough, peaks in the evening hours*

Comparing the *treatment*and *control*sets, we observe that the median prime time and late night behavior is very similar (7:00 PM – 7:00 AM), but during off peak daytime (work) hours, the *treatment*set has a higher median than the *control*set. There was no change in prime time behavior in the evening, and an increased usage in off-peak daytime hours.

Concluding remark.. Conclusions for the FCC. Conclusions for the ISP

see todo.txt, todo.tex for analysis comments, explain, check

4.2 Prime Time Ratio

Prime Time: The daily diurnal nature of usage patterns across many households naturally requires the provider to design networks capable of handling load at the peak times in a day. Such peak times are usually observed during evening hours, and the data transferred at this time is called peak usage. The FCC defines *Prime Time* as the local time from 7:00 PM to 11:00 PM [1], when many households heavily consume real-time entertainment traffic (video), seen as primarily responsible for high usage during these hours. Latency and performance are adversely affected during prime-time, causing bottlenecks at home, the last mile, in transit, or at the content server. For example, the Sandvine Global Internet Phenomena Report [3] showed that devices in the same household selected Netflix's own CDN (OpenConnect) during off-peak hours, and third party CDNs (with differing performance) during prime-time. This may happen because Netflix OpenConnect is over-utilized during prime time [6].

Prime Time Ratio: To measure the concentration of network usage during prime time, Sandvine defined the *Prime-Time ratio* as the "absolute levels of network traffic during an average peak period hour with an average off-peak hour". Based on the FCC definition of prime-time hours (7p-11p), we measure the daily prime-time ratio of set_{full} in section 4.2.

To characterize the prime-time ratio, as defined in **??**, we calculate the aggregate data transferred at the ISP in an average prime-time hour, and divide it by the off-peak average.

Prompted by the monotonically increasing trend of usage behavior during daytime hours on weekdays (figure 2) we calculated the prime-time ratio for each four hour period throughout the day to find the evening hours with the largest ratio. In our dataset, the prime time ratio peaks at 8:00 PM – 12:00 AM, rather than FCC's definition of 7:00 PM – 11:00 PM. This discrepancy could be limited only to the high tier households in our dataset, but we deem that unlikely. Another reason could be that prime time is delayed globally with the rise in real time entertainment's contribution to traffic.

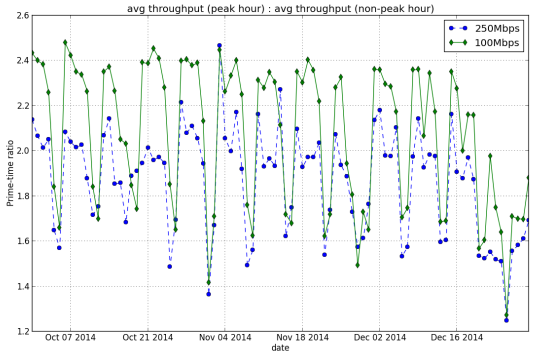


Figure 3: *Prime Time ratio showing weekly pattern + differences during holiday periods (Thanksgiving, Christmas)*

We use our updated definition of Prime Time (table 2) to calculate and plot the Prime Time ratio per day for the *treatment*and *control*sets in figure 3. A comparison shows that the *treatment*set's prime time ratio is 10% lesser, supporting

[3]The Sandvine Reports [6, 7]are released bi-annually and contain a detailed analysis of aggregate Internet usage. They are also referred to in the FCC reports [1–3]

the observation from section 4.1 that showed that the usage during prime time is similar across both sets, but the usage in off peak hours is higher in the *treatment*set.

see todo.txt for remaining analysis and concluding remarks

4.3 Interpreting Utilization

Utilization can be characterized in many ways, and each definition can answer a particular question. Furthermore, as stated in the previous section, multiple parties may interpret utilization differently, based on their aims and requirements.

Utilization as the average data rate for each device in each time slot portrays the overall spacio-temporal aspect of our dataset. We plot a distribution of the rate per device-time in figure 4. This shows that median utilization of household users is the same in both sets (50 kbps). A comparison between the *treatment*and *control*set shows that there is no affect on the overall adoption.

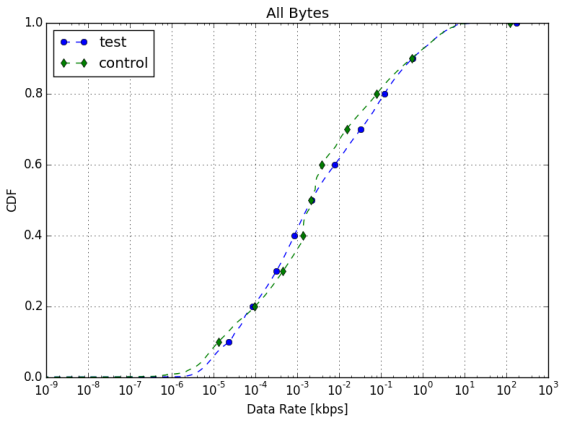


Figure 4: *CDF of data rate per time slot for all devices (agg view of data): Overall not much change due to capacity increase. Median data rate 2bps for 3 months x thousands of devices!*

Based on the measurement methodology, we study the highest utilization seen by a household both in its lifetime, and on each day. Our aim is to examine the peak usage per household, and study if the behavior changes due to an upgrade.

Figure 5a provides a distribution of the highest average data rate a household achieves. To avoid outliers, we also plot the 90%-ile of the max data rate achieved by households in both *treatment*and *control*sets. We see that a median household is expected to achieve the highest data rate of between 1 – 10 Mbps over its lifetime. This is much lower than the access link capacity, indicating that the median device has a utilization ratio (avg data rate:capacity) under 0.1 in our dataset. The number of households that increased their peak utilization beyond the *control*set's 105 Mbps capacity were negligible.

Surprisingly, we see that 30% of the households from the *treatment*set have a low peak utilization (under 0.1 Mbps), while 40% of the *control*set households are under 0.1 Mbps. Thus, the absolute peak utilization does not increase when compared to the access link capacity, but there is certainly an increase in peak utilization of devices that had a low requirement, due to the change in capacity.

To investigate this further, we also study the *peak utilization per device on a daily basis*. Figure 5b shows that for 30% of the devices, the maximum data rate in the *treatment*set is consistently higher than the *control*set, albeit no where near the actual access link capacity.

This is similar to the behavior observed in figure 2, showing that the peak usage during prime-time is unaffected, but lower utilization throughout the day is higher for the test set. We speculate that there could be two possible reasons for this increase in utilization: (1) short term downloads and/or web browsing achieves a slightly better data rate on a small time scale, or (2) real-time video quality is slightly higher, but not enough to completely saturate the access link capacity. Unfortunately, we miss these short lived, or consistent, events due to a 15 minute time slot granularity and only looking at byte counters.

Different Perspectives of Utilization: Sarthak: this should be only in the discussion?? We take this opportunity to reflect on the interpretation of the disparity of peak utilization per device, as shown in figure 5a. The ISP may interpret this as no change in peak usage, as the prime-time usage remained the same based on aggregated usage, even in prime-time. Thus, we believe that given the opportunity, the provider will not invest to offer a higher access link unless it in a region showing such low demand unless it is guaranteed profit, or is forced into deployment by an external agency.

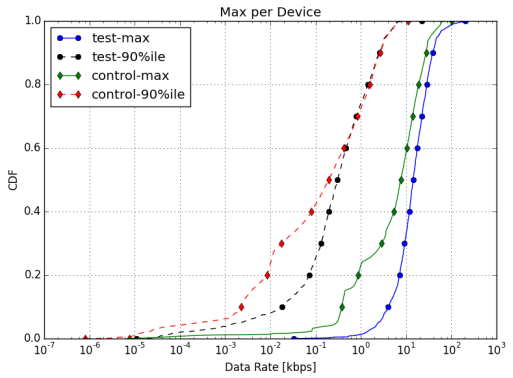
In contrast, the consumer (and therefore the FCC) might be convinced that individually the usage behavior of a household is affected by the increase in access link capacity, especially for households with a lower utilization. We believe that this is the perspective the FCC takes when considering deployment and adoption of broadband services.

needs work on conclusion Although our unbiased experiment still shows a certain correlation between utilization and capacity, it also contradicts the law of diminishing returns [**?**]. **sanity check**

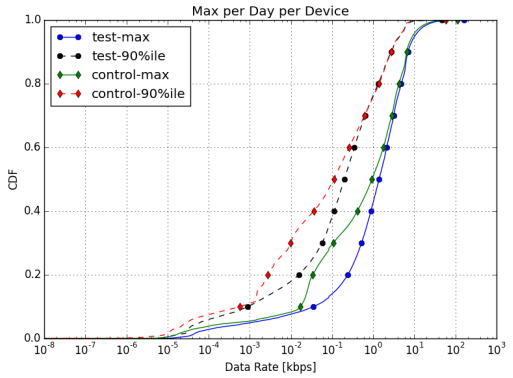
4.4 Peak Usage

Internet usage throughout a day follows diurnal sleep-patterns, and researchers have shown that such patterns are in fact correlated with GDP, Internet allocations, as well as electrical consumption of a region [5]. This makes the study of usage behavior extremely relevant to the governmental bodies responsible for development, such as the FCC, when considering policy decisions.

Peak Ratio: The Sandvine Reports show that although the mean usage has remained stable for the past few years, usage during peak-times has increased drastically [6]. To measure this growth, they introduce the concept of peak period, measured when the network is within 95% of its highest point. Although, these reports present a good view into aggregate usage patterns over a month, they neglect to analyze usage



(a) *CDF of max per device: test set has higher (max) average data rate below 10 kbps. 30% of devices in the control set have a max data rate of 2 kbps while 30% of test set has a max data rate of 10 kbps. (sanity check numbers, redo plot)*



(b) *CDF of max per device daily*

Figure 5: *Peak Utilization: The maximum data rate varies for test and control set for low data rates, and this variation is present daily.*

characteristics individually. Inspired by their definition, we measure the disparity between the 90 percentile of the peak and median usage of each household within a day, and call this the *Peak-Ratio*. In section 4.4 we show that the peak ratio can be used to divide users in the same tier based on their usage behavior.

To further characterize and compare the deviation of data rate for the *control*and *treatment*set, we examine *peak-ratio* as defined above. Figure 6 shows that the median peak-ratio for each device in the *treatment*set is much larger than that of the *control*set. **replace much larger with the exact number or percentage** . Sarthak: Taken together with our observations of a lower prime-time ratio of the *treatment*set (section 4.2) this implies that there are households in the *treatment*set that achieve a peak-ratio > 1, but not during the prime-time hour. We believe that these households might actually be small businesses or work-at-home users that peak during daytime hours instead of evening hours.

The median peak-ratio per device itself shows a large range, from 1 to 10e6 (figure 6), and the maximum peak-ratio per device was an order higher. Clearly there are some households

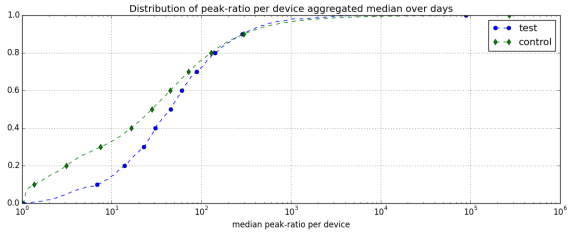


Figure 6: *Median peak ratio per device showing that test set has higher daily ratio (50 times by median). Thus ISPs should condition their networks to 50 times the median usage for each user added in the worst case scenario.*

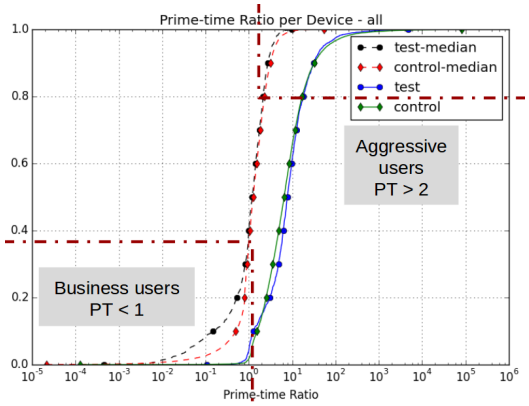


Figure 7: *(old) Prime Time ratio + usage can be used to divide users into four sets: aggressive all time + non aggressive all time, aggressive peak time, aggressive non-peak time (business hours). 30% PT < 1: possibly businesses with normal work-hours . 20% PT > 2: aggressive prime-time streamers*

that have a very even usage throughout the day (low peak ratio), and others that are extremely aggressive only at certain times (high peak ratio). We plot this segregation in figure 7.

TO PLOT? :

- peak ratio cdf vs no of devices
- peak ratio cdf vs time of day where peak occurred
- no of devices cdf vs time of day where peak occurred

Based on differing usage profiles within the same high tier bandwidth, we suggest that the FCC adopt multiple benchmarks based on usage characteristics to better characterize broadband availability, deployment, and adoption in the US. Such multiple benchmarks can be the minimum broadband speed required per user based on the kind of traffic expected during a day. ISPs can also offer these users better plans based on hour-of-the-day or usage caps to encourage more off-peak usage. These users probably don't cause latency spikes in PT.

4.5 Traffic Asymmetry

just some stats on asymmetry: maybe merge this with usage??

- **maybe this doesn't need a separate section, just comment on asymmetry in each of the above**
- Cisco vs Alcatel: upload is increasing vs there's still too much download
- Claim that its reducing due to uploads, but content is mostly download. Is the ratio still 10:1 (FCC thinks its 25:3)
- Compare with Sandvine asymmetry stats
- Talk about 3 Mbps comparison with observed uplink in control and test set

4.6 Prevalence

Sarthak: add plot on the 11 users that changed behavior - but need to explain the stats without these aggressive users

Sarthak: it'll be awesome to add persistence to this plot after segregating rich greedy and rich frugals and reporting stats - easily done in pandas

5 Discussion

5.1 Different Perspectives of Utilization

The FCC has the responsibility to increase the availability and deployment of broadband throughout the US (with the broadband threshold benchmark defined as 25 Mbps in downlink and 3 Mbps in uplink). Their progress report states that: given the option, users will adopt a higher tier bandwidth [], thereby meriting the high investment. However, a survey conducted by NCTA showed that the largest deterrent to broadband adoption is that users do not *need* broadband (the second largest is the cost). The conflicting view of the ISP is that the cost of deployment in an unchartered area is too high, unless a significant number of households *need* it. Thus, both parties are asking the same question: do people *need* a higher capacity, i.e., what is their *utilization* as compared to the capacity?

Previous research shows that the utilization will increase as the capacity of the access link increases [], and also that utilization[4] and capacity follow a law of diminishing returns [**?**]. However, such studies have been biased by studying users that actually required a higher capacity for their usage. It is inevitable that such a correlation would exist for such households, whose utilization is bottlenecked by the ISP.

In this work, we ask a more fundamental question: *how much does the user behavior change with increasing capacity*. Specifically, when the capacity is already very high and the user has not opted for an increase, does their utilization still vary with capacity? Both the FCC and the ISPs have a different perspective of the utilization:

The FCC perspective: *Utilization as adoption* of a higher capacity link when available (but not under the constraints of a much higher cost). The answer to this question is important to the FCC to encourage further deployment of high tier links throughout the US. Essentially, if *any* change is observed in link utilization due to the upgrade in our dataset, the FCC may interpret that as *adoption* to the higher available tier.

The ISP perspective: *Utilization as a capacity bottleneck*, i.e., if the ISP can show that the *maximum utilization* of a household does not vary with increasing capacity, it will prove there is not enough demand to offer a higher tier. The ISP needs the answer to this question for future capacity planning, and the cost-analysis for the investment of new technology in any area. For example, Google Fiber is now expanding to Salt Lake City, from where we received our dataset. The analysis of change in user behavior with capacity will estimate the number of users that actually *need* the higher capacity service offered by Google.

Thus there is a need to measure *utilization* at times when users *need* the Internet capacity the most: the peak usage.

EVERYTHING BELOW THIS IS TODO

5.2 User Taxonomy

The Sandvine reports present a taxonomy of users based on their contribution to real-time entertainment traffic. We incorporate a similar definition based on contribution to data traffic, along with our observations of utilization, to present a taxonomy of the users in our dataset. One category of users is the non-utilizers, i.e., non-aggressive low bandwidth users, that contribute less than SOME THRESHOLD PERCENTILE to the daily data transferred Sarthak: these the ISP can ignore, also they probably don't need this tier as their utilization from the previous section must be super low . The second category is of users contributing most aggressively to the data at the ISP Sarthak: these users will probably gobble up a higher capacity link if given a chance - they're the ones who effect all our graphs.. Need to check this claim . We further subdivide this high utilizing subcategory based on differing prime-time ratio and peak-ratios follows... **need to think and analyze this further: technical definition to do the analysis**

- Aggressive All-Time: Users having a low peak-ratio due to a lower variance. Is also expected to have a low prime-time ratio.
- Aggressive Prime-Time: The usual streamer with a high prime-time ratio and a high peak ratio.
- Aggressive Non-Prime-Time: Possibly a business user with a low prime-time ratio but a high peak ratio

6 Conclusion

[4] called demand in their work

References

[1] Federal Communications Commission. Measuring Broadband America - 2014, April 2014. (Cited on page 4.)

[2] Federal Communications Commission. Tenth Broadband Progress Report No 14-113, February 2014. (Cited on page 4.)

[3] Federal Communications Commission. Eleventh Broadband Progress Report No 15-10A1, February 2015. (Cited on page 4.)

[4] Federal Communications Commission. International Broadband Data Report (Fourth), February 2015. (Cited on page 2.)

[5] Lin Quan, and Heidemann, John and Pradkin, Yuri. ANT Evaluation of the Diurnal Internet, October 2014. (Cited on page 5.)

[6] Sandvine. Global Internet Phenomena Report - 1H, April 2014. (Cited on pages 4 and 5.)

[7] Sandvine. Global Internet Phenomena Report - 2H, November 2014. (Cited on page 4.)