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

Sarthak Grover, Roya Ensafi, Nick Feamster
Princeton University

Paper #0, 3 Pages

Abstract

The Federal Communications Commission (FCC) recently declared that "advanced telecommunication capabilities" are not being deployed in the US in a timely fashion [1]. In an attempt to encourage Internet Service Providers (ISPs) to invest in deploying advanced broadband, the Commission has requested comments on issues pertaining to availability, deployment, and adoption of broadband.

In this work, we provide a much required input to the FCC's open questions, supported by our analysis of usage patterns. We present case study examining the relationship between supply (availability) and demand (adoption) in a controlled experiment. Our analysis shows that peak user demand within a single high speed tier is highly diverse, even when accounting for factors such as capacity, cost, performance, and location. This work motivates the need to adopt user demand as a benchmark to further the deployment of "advanced" broadband capabilities in the US.

1 Introduction

2 Background and Related Work

3 Data Source and Characterization

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 test 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. TODO: confirm - these were reported by Comcast gateways right?

3.1 Data Description

The raw data sets provided by Comcast consisted of the test 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.

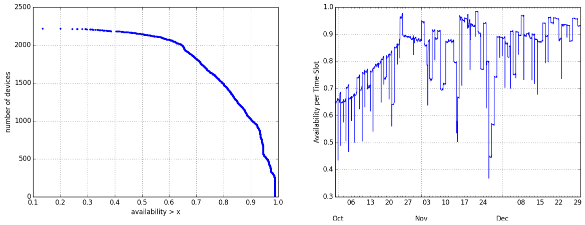
Field	Description
Device_number	Arbitrarily assigned CM device identifier
end_time	Fifteen minute sample period end time
cmts_inet	Cmts identifier (derived from ip address)
service_direction	1-downstream, 2-upstream
octets_passed	Byte count

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. TODO: 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 test and control datasets, we decided to only choose gateway devices with an heartbeat of at least 0.8.



(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.)*

We sliced the sanitized test 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 test and control sets in this case simply as datasets $set_1 - set_8$, where set is test or control . We also sliced and combined the sanitized data to give us control and test 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 test dataset, and we refer to this simply as set_{full}.

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

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 test 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 test 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 test set behavior as aggregate changes that occurred due to the an increase in access link capacity.

4 Empirical Analysis

5 Discussion

6 Conclusion

References

[1] Federal Communications Commission. Eleventh Broadband Progress Report No 15-10A1, February 2015. (Cited on page 1.)