Analysis of the Comcast Device Usage Dataset

Sarthak Grover

Meeting: April 21 2:00 PM ET

Key Findings

- **No Change:** Heavy hitters (almost half the users) do not change their 95% peak usage behavior.
- **Change:** Lighter users (the other half) show a significant increase in 95% peak demand due to the treatment.
- **General Patterns:** Peak hours (prime-time) are later in the day: 8p to 12a.

Note: Presented results (slides 1-15) are for downlink data only. Uplink results follow in the backup slides

Hypothesis and Control Parameters

- **H:** Subscriber demand and usage behavior doesn't change as the access tier is upgraded from 105 Mbps to 250 Mbps
- Controlled Experiment in SLC, Utah
  - Cost, Performance, Location, Demographics assumed to be the same
  - User is not informed, behavior assumed unchanged
  - **Capacity** is the only changing parameter

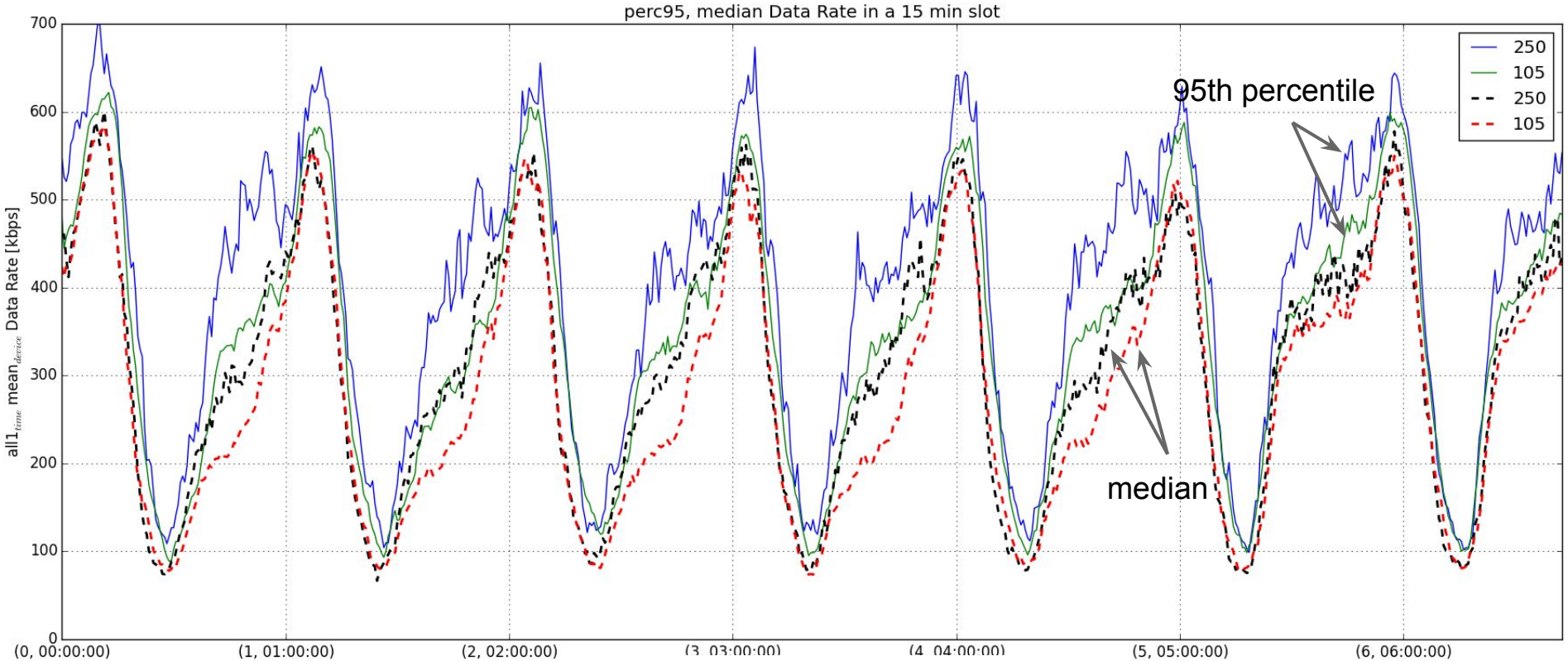
- Capacity and demand are correlated. When does this break down?

Data Description

- Values: Data transferred (bytes) per household (subscriber) every 15-min time slot
- Time Series: Three months: Sept 30 - Dec 29, 2014
- Control group
  - Access tier: 105 Mbps
  - Number of Significant Households*: 4845
- Treatment
  - Access tier: 250 Mbps (unknown to the user)
  - Number of Significant Households*: 1519

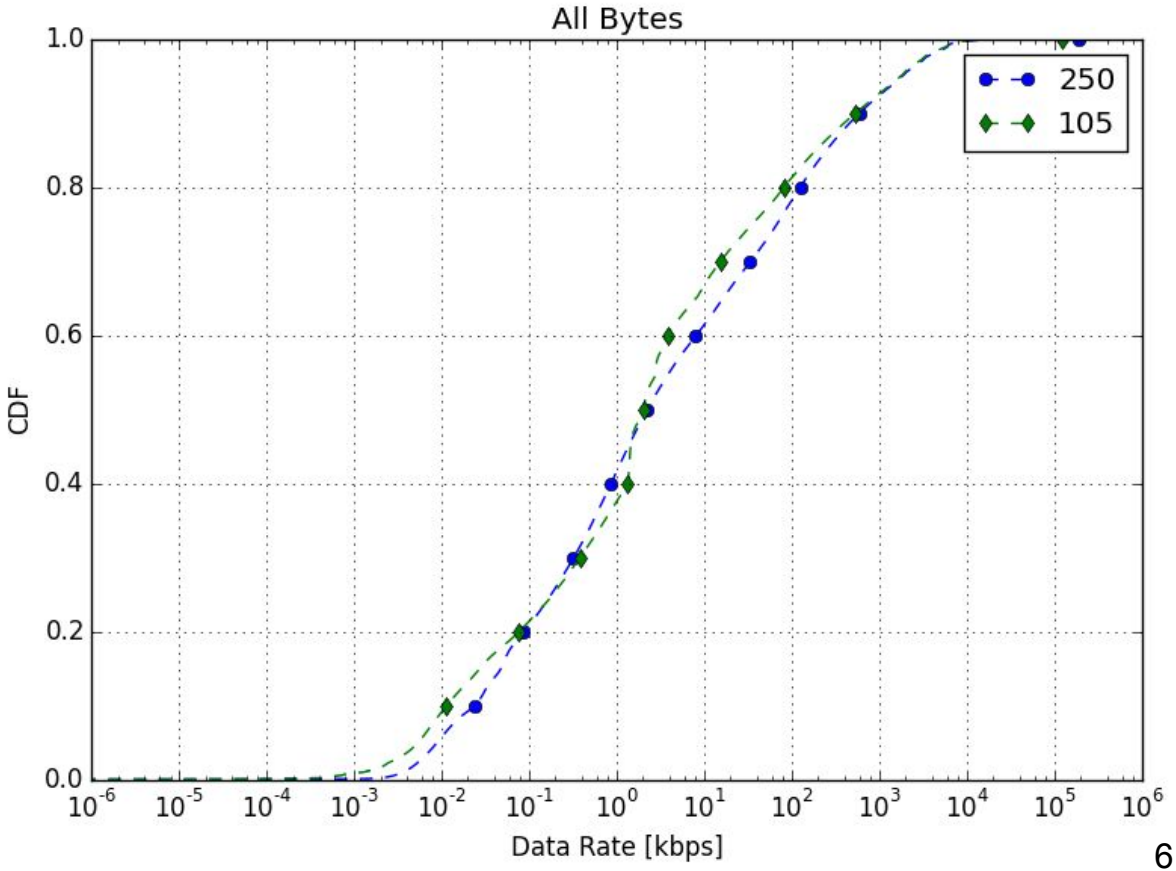
*Significant Households: devices that sent their byte counters at least 80% of the time in 3 months

Average Data Rate per Subscriber in a Week



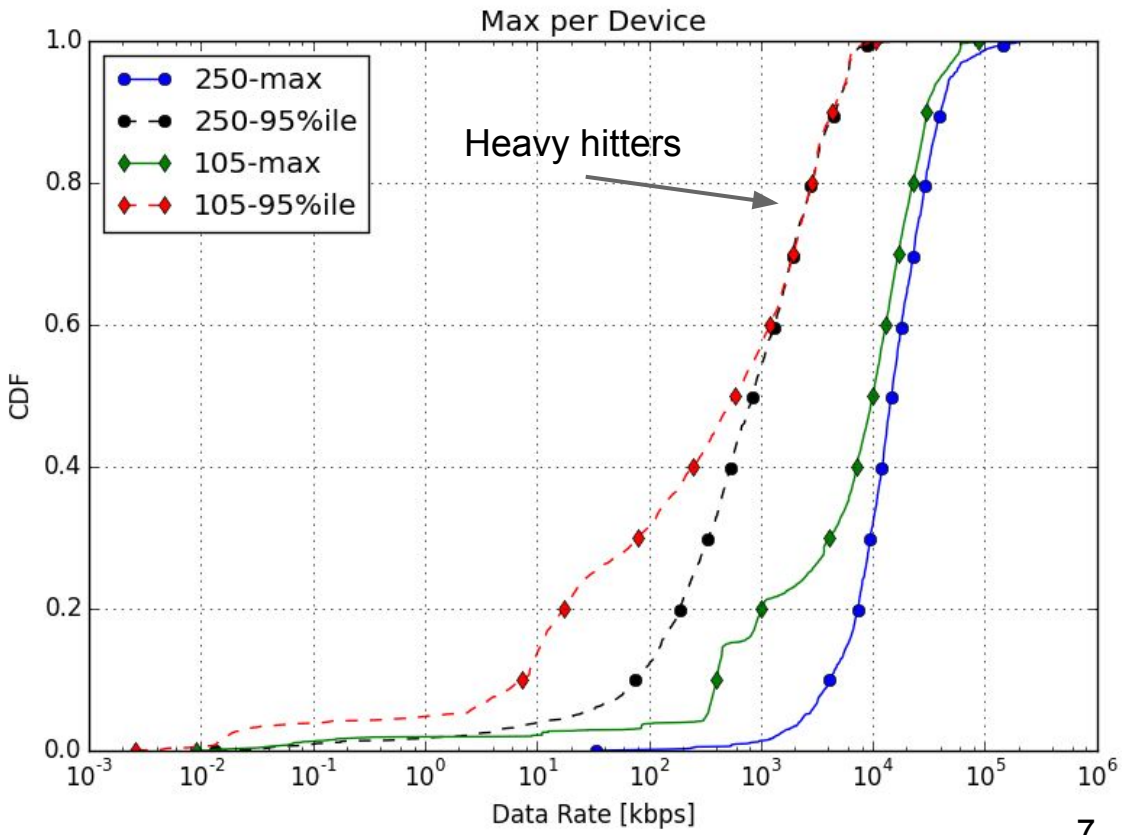
Distribution of Average Data Rate

- Distribution of average data rate (kbps) per 15-min time slot
- **Very similar distributions of bytes transferred**
- Median data rate is approx. 2 kbps for 3 months.



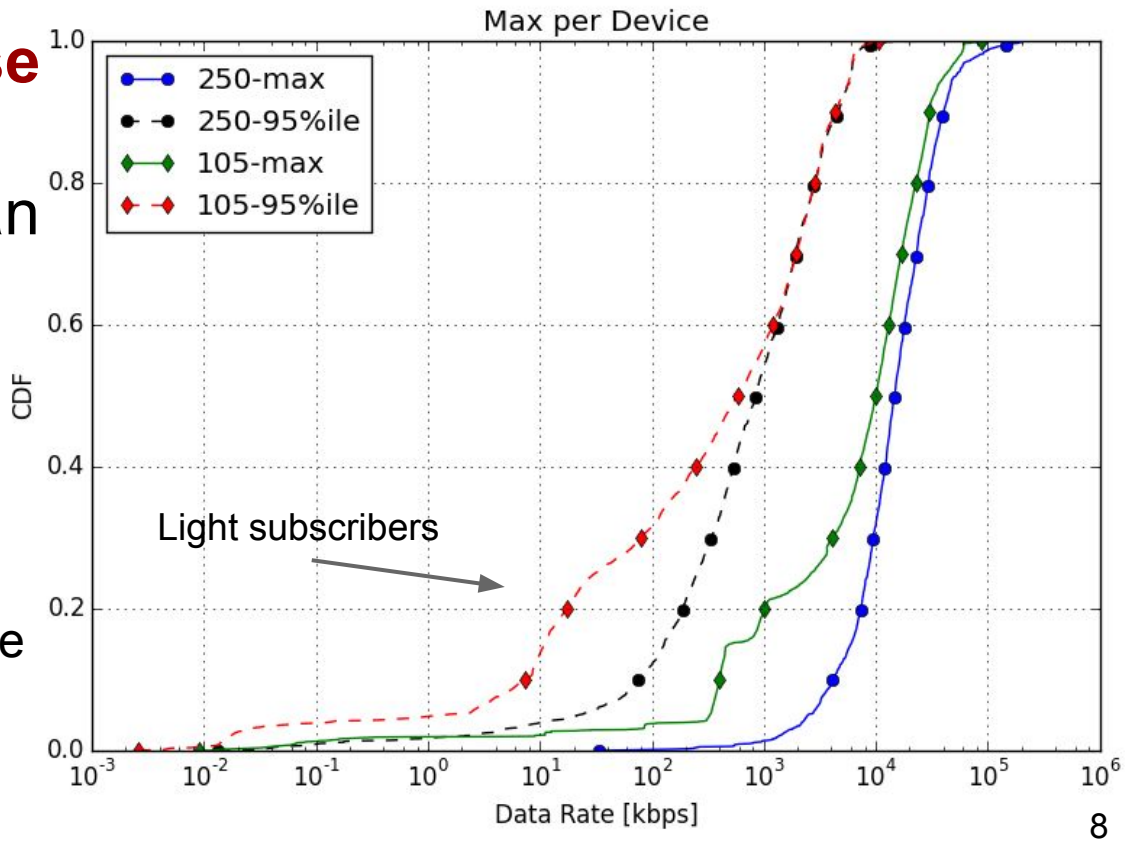
Distribution of Max Data Rate per Subscriber

- Defining utilization as the ratio of max demand (or 95% peak demand) to the link capacity.
- **Heavy hitters**: half the households in control and treatment have a high link utilization (between 10-100%). **They do not change.**



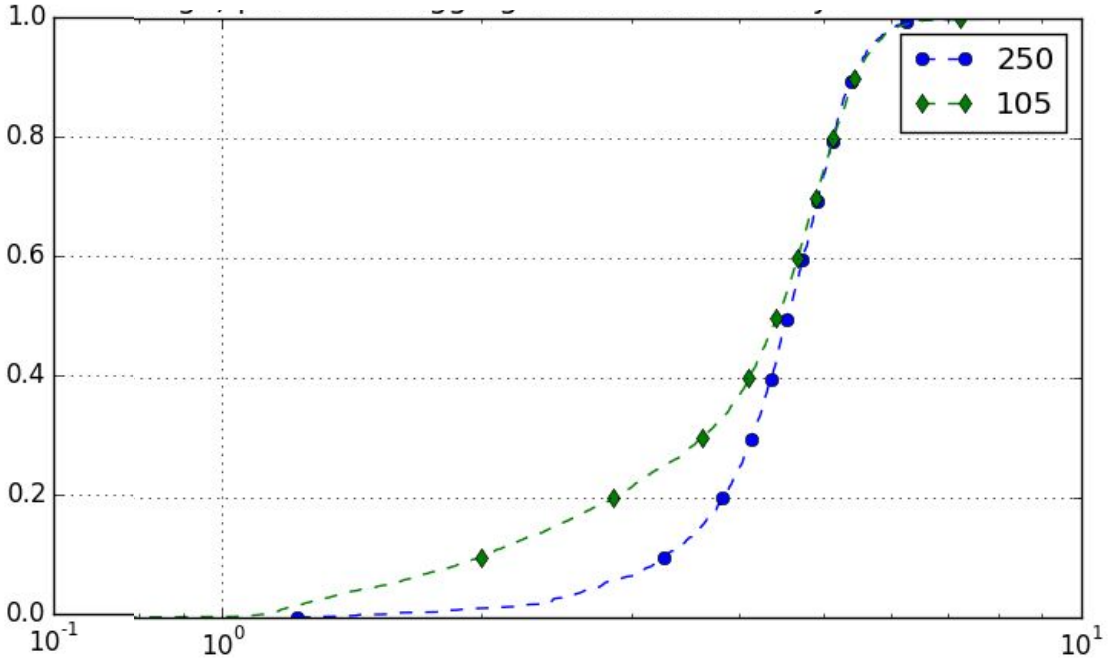
Distribution of Max Data Rate per Subscriber

- **Light subscribers**: half the households **increase their utilization** from less than 1% to less than 10% in the treatment.
- This happens daily!
- Why? Needs examination.
  - Is it short bursty flows?
  - Background time-insensitive apps (unknown to user)?
  - Did user actively change behavior?



Distribution of Daily Trough to Peak Ratio

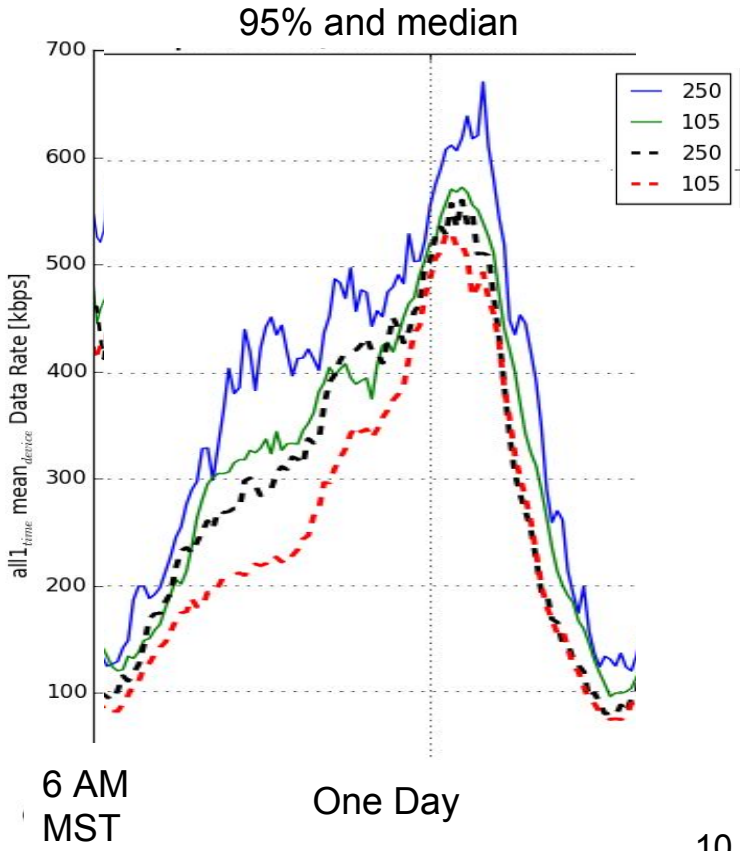
$$\text{peak ratio} = \frac{\text{95-\%ile data rate (per day)}}{\text{mean data rate (per day)}}$$



- Average peak ratio per subscriber increases for 40% of households with a lower ratio (from 1-3 to 2-4).
- Variance per subscriber is higher in treatment

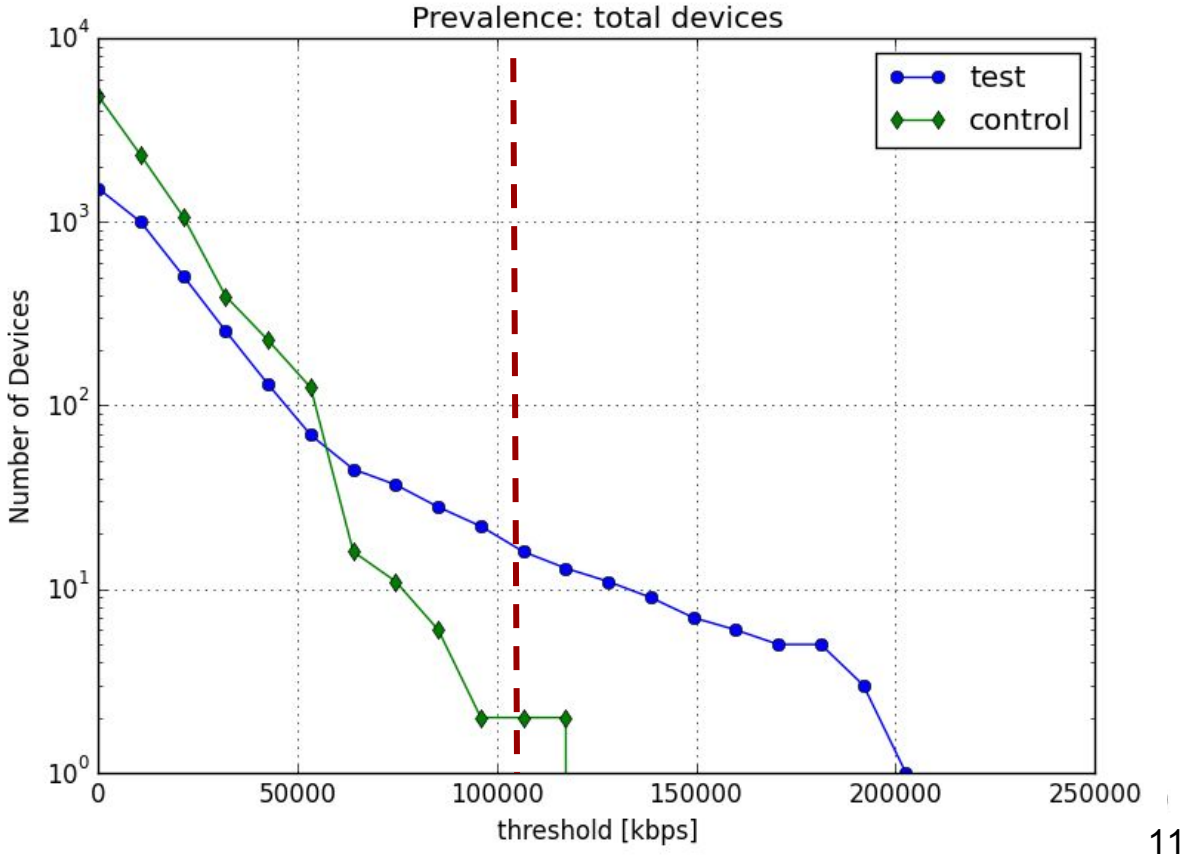
Average Data Rate per Subscriber on a Day

- Prime Time is 8PM to 12 AM
- Demand increases from morning to prime time, without any trough in mid-day.
- 95% of the daily subscriber data rate is much higher for the treatment, so some devices do increase demand.
- **Median daily subscriber data rate is also higher during off-peak hours**



Outliers: Devices with High Demand

- Number of devices reaching a threshold data rate in their lifetime
- Beyond 100,000 kbps only 2 subscribers in the control group, and 11 subscribers in the test group



Conclusions: (1) No Changes

- Usage Patterns:
  - **Overall data rate** and average utilization are similar
  - Heavy users who utilize most of the link capacity **do not increase** their demand
  - Average **peak ratio doesn't rise** for majority of subscribers
  - Prime time seems to be **8p-12a** for both control group and treatment
- This suggests that (1) the series are similar and (2) the ISP is **not the bottleneck for heavy users.**

Conclusions: (2) Changes

- **Peak demand per device**
  - For time slots with **low data rate over 15 min**, peak demand per subscriber is higher in the test set
  - Median data rate per subscriber during off-peak hours also increases
- There are a few outliers that have very high demands (high utilization) that increase in the treatment.

- This suggests that **increasing ISP capacity** (tier) **affects demand** for very high speed tiers, even though the link is **not fully utilized** by the affected subscribers.

Next Steps

The dynamic relationship of user demand, tier capacity, and link utilization needs further study.

Question: How much does the demand change with capacity? Is the change significant?

- We expect a correlation that follows the law of diminishing returns
- Compare with controlled experiment for **lower speed tiers** to study the significance of this change.

Question: Why did demand increase for low utilizing users?

- Look at **smaller time scales**, devices inside the home, and applications used (DPI)

Thanks

[Backup Slides]

Outline of Backup Slides

- Data and Sanitization
  - Questions on Data Collection
- Extended Results
  - Generic Time Series (Downlink)
  - Peak Utilization per Day (Downlink and Uplink)
  - **Prime Time Ratio (PT)**
    - Time series over 3 months (Downlink and Uplink)
    - Distribution over subscribers (Downlink and Uplink)
    - User Taxonomy based on PT and Usage

DATA AND SANITIZATION

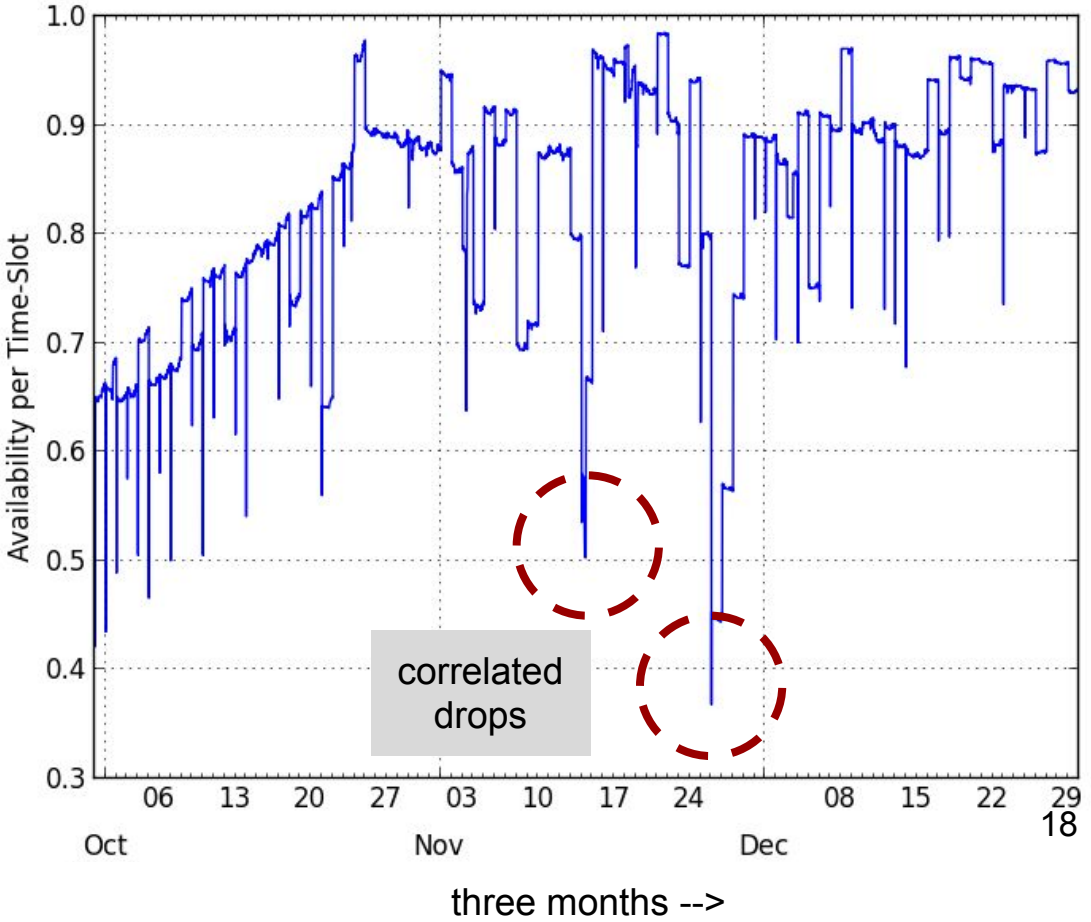
- Data sanitization using subscriber heartbeat
- There are correlated drops in data on certain days.

- Final data metrics: http://sites.noise.gatech.edu/~sarthak/files/comcast/separated/metadata.txt

For results web access code please email: sgrover@cs.princeton.edu

Data Sanitization: Responsive Devices per Time-Slot

- Some devices don't contribute much data
  - Discard
- Correlated non-responsiveness at certain times
- 28-30 Nov, only 35% of the subscribers were ON?



Data Sanitization: Heartbeat > 80%

- Control group: approx 2000/4000 devices had high availability in each set.
  - control1-3: Oct
  - control4: Oct, Nov, Dec
  - control5-7: Nov
  - control8: Dec
  - control: Combine above to one large set over 3 months (more than 4000 devices)
- Test set: approx 1500/2200 devices had high availability
  - 3 months: Oct 1, 2014 - Dec 29, 2014

Results: Treatment (250 Mbps) vs Control (105 Mbps)

Data: Questions for Comcast

- Data Description
  - What is **access link capacity** for control1-8?
  - What is '**service_class_name**', '**device_key**'?
    - We use **total octets** per time slot per device.
  - Are the IP addresses ('**cmts_inet**') anonymized?
    - Confirm: Very few IPs vs Devices, and all in SLC
    - Test set has 2200 unique 'Device_number' but only 25 unique IP addresses in Salt Lake City.
- Data Sanitization
  - **Availability** of devices **drops** suddenly at times.
  - Is this due to outage, collection problems, or slicing?
    - Can we get fixed data if it was a slicing issue

Evaluation Criteria

Calculate average data rate for each device in each slot and compare with access link capacity (250 Mbps or 105 Mbps)

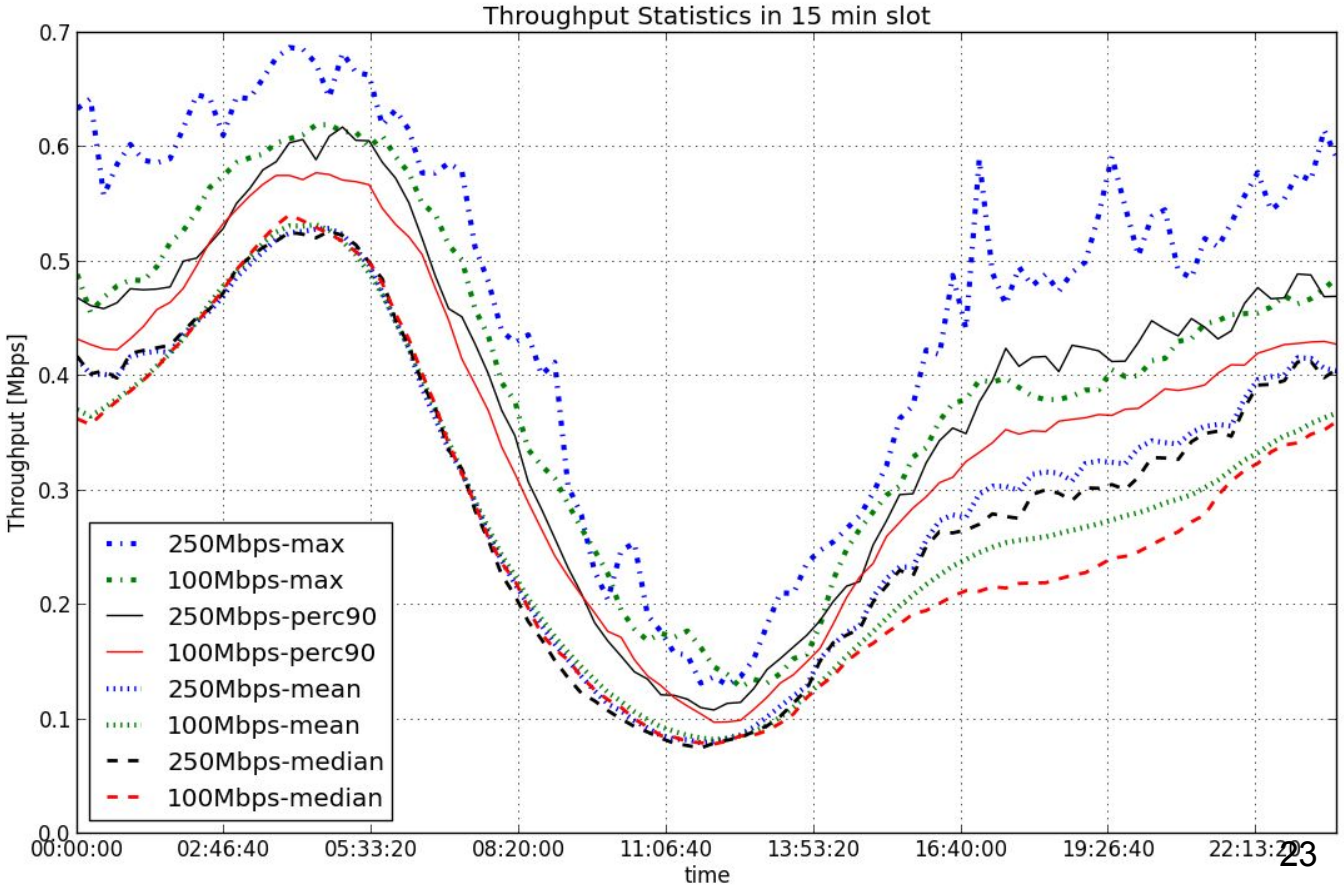
- **Usage pattern**: compare utilization, i.e., average data rate per 15 mins.
- **Peak utilization**: compare max data rate of devices over a day
- **Prime-time ratio**: compare average data rate during prime-time to average data rate outside prime time

Usage Pattern: Description

- **Q1: Average data rate per day, aggregated across devices**
  - aggregated by mean, median, max, 95-percentile across devices
- Calculate average data rate using bytes transferred per 15 min slot
- **Q2: Does the average data rate change when we go from 105 Mbps (control) → 250 Mbps (treatment)**
- Plot a distribution of the average data rate seen in both datasets

Stats on a generic day

- green, red are 100 Mbps.
- blue, black are 250 Mbps.
- max, and 90%ile are consistently higher for test set (not big difference for 90%ile)
- means and medians are indistinguishable during busy hour, but in other hours test set has high mean.
- So ratio throughout the day between 90%ile:median may vary for test, control. During busy hour, test will be larger than control, but during others it may be very different



The importance of measuring peak

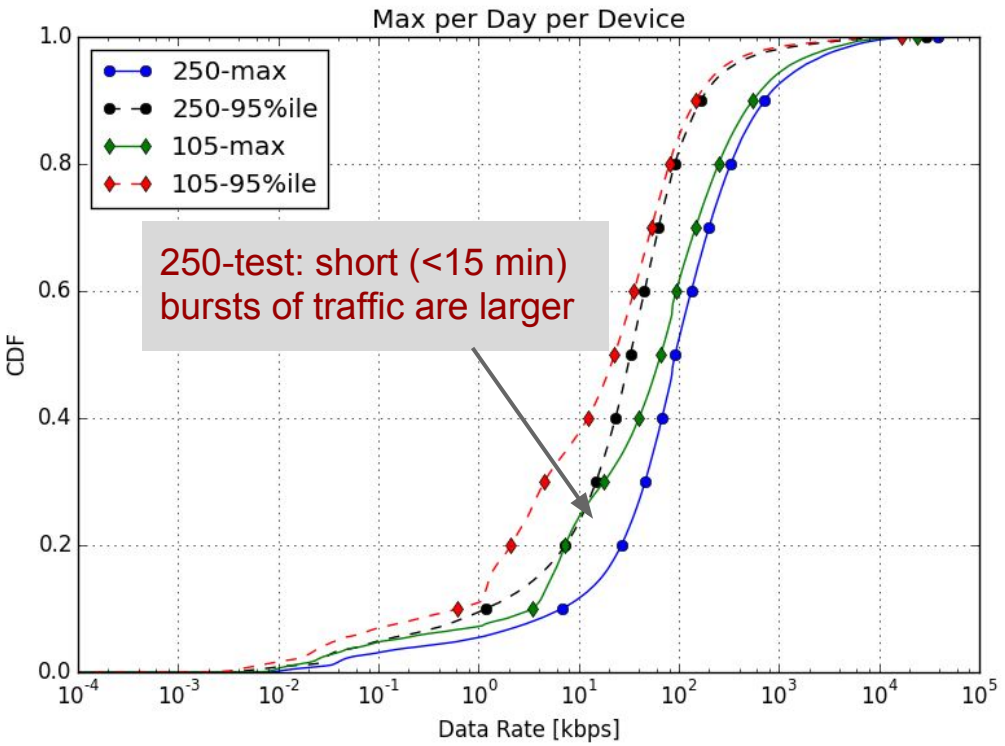
- Prime Time
- Mostly low usage and similar diurnal pattern
- Capacity planning: ISP load at busy hours
- CDNs/Servers: peak latency at prime time
- 63% real time entertainment, 32% netflix, 14% youtube. [Sandvine, BBAmerica]
- Utilization should be measured in prime time not full day

Peak Utilization: Description

- Capacity planning is concerned about “prime-time” and peak behavior, i.e., max data rate per device is more important than the average rate of that device.
- Idle hour behavior is expected to stay the same, regardless of access link
- **Q: Does maximum (or 95-%ile) data rate per device change with access link?**
- Calculate the **maximum data rate for a device** over the three months, and compare max rate (per device) for test set and max rate (per device) for control set

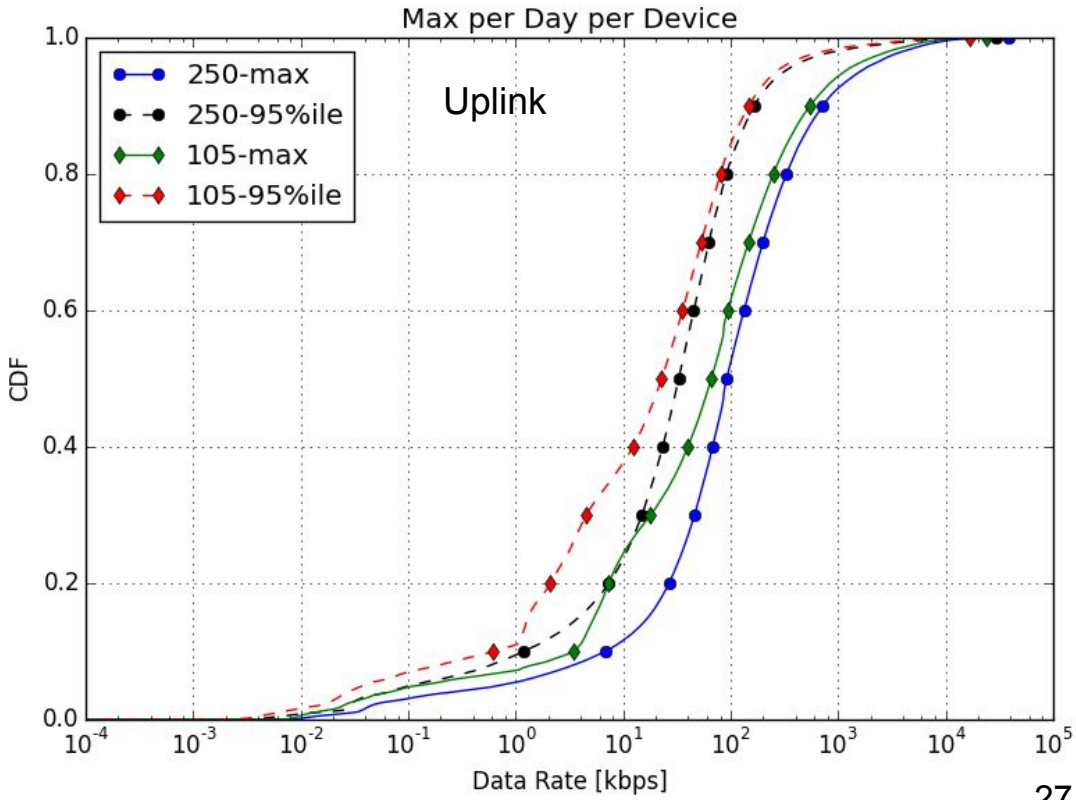
Downlink Peak Utilization: Max Data Rate per Day per Device

- Max seen by a device per day, in 3 months
- Consistent increase for daily max data rate per device
- **15 min granularity misses information**
  - short faster data bursts
  - better video quality
  - baseline is different



Uplink Peak Utilization: Max Data Rate per Day per Device

- Similar for uplink
- Each day demand increases but for users who do not utilize the link (their max per day peak demand is less than 100 kbps)
- Are these background apps that the user doesn't care about? Or did the user actively change behavior?
  - need higher granularity DPI data to conclude



Prime Time ratio

http://riverside.noise.gatech.edu:8083/separated/full/df_best_primetime_hour.html

- Test and control set max at 2-6 AM UTC
- That is 7p-11pm PST, but 8p-12am MDT
- **No dips in the middle of the day**

	nonpeak_c	nonpeak_t	peak_c	peak_t	start_time	stop_time	test_ratio	control_ratio
0	0.252280	0.278608	0.436101	0.446679	00:00:00	04:00:00	1.603255	1.728638
1	0.244622	0.272400	0.475461	0.476701	01:00:00	05:00:00	1.750003	1.943654
2	0.241829	0.269569	0.496987	0.496832	02:00:00	06:00:00	1.843061	2.055114
3	0.246810	0.273880	0.482536	0.485319	03:00:00	07:00:00	1.772015	1.955087

Prime-Time Ratio: Description

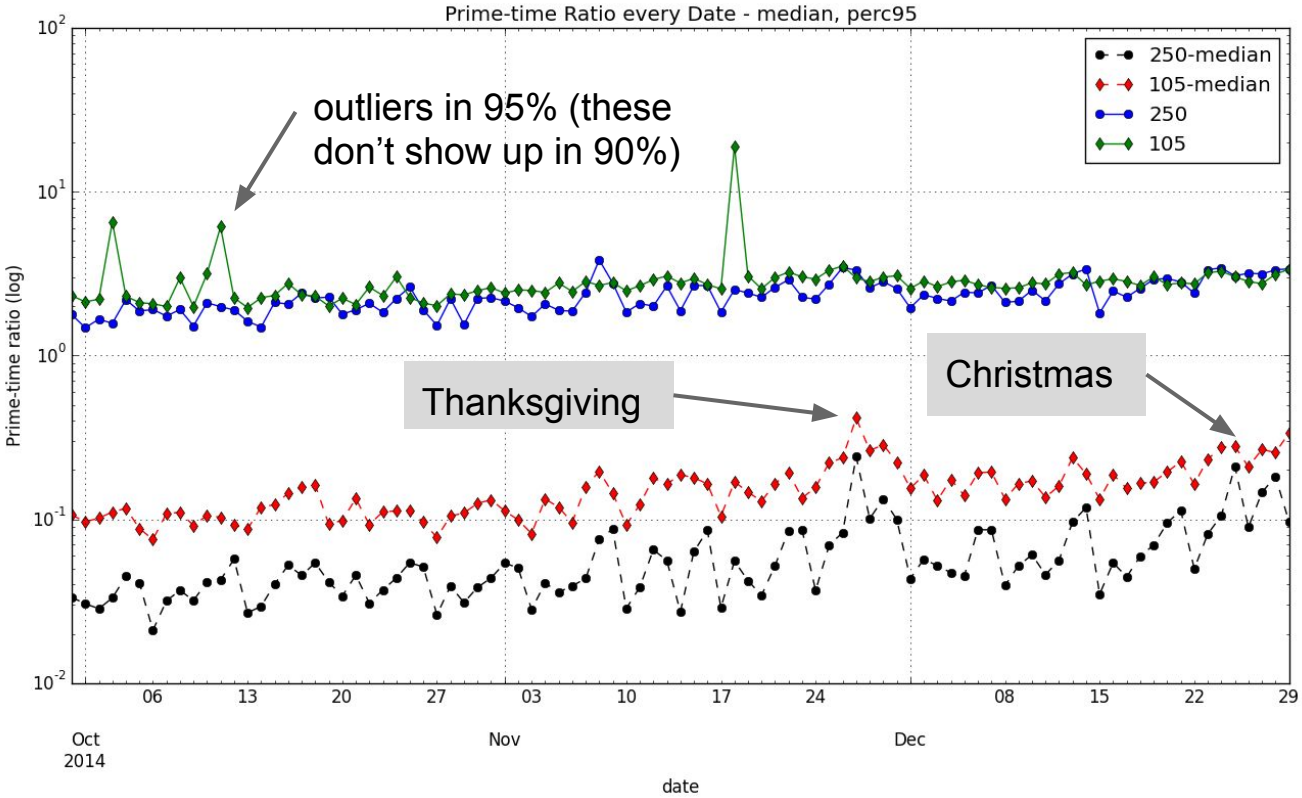
- Sandvine defines Network Prime-Time ratio to measure the concentration of network usage during the prime-time evening hours.
- FCC says prime-time is 7-11 PM.
- Prime-Time ratio = absolute levels of network traffic during an average peak period hour with an average off-peak hour.
- **Q: Does the Prime-Time ratio vary for test and control sets?**
- Measure PT = avg data rate during a peak hour period : off-peak period.

Prime-Time Ratio (PT): Results

- 8p - 12a shows a higher prime-time ratio than 7p - 11p.
- The curve in convex, i.e. there is only one peak time in a day and no troughs.
- Distribution of prime time ratios over control and test set is very similar
- The average prime-time ratio for control sets is consistently higher than test set

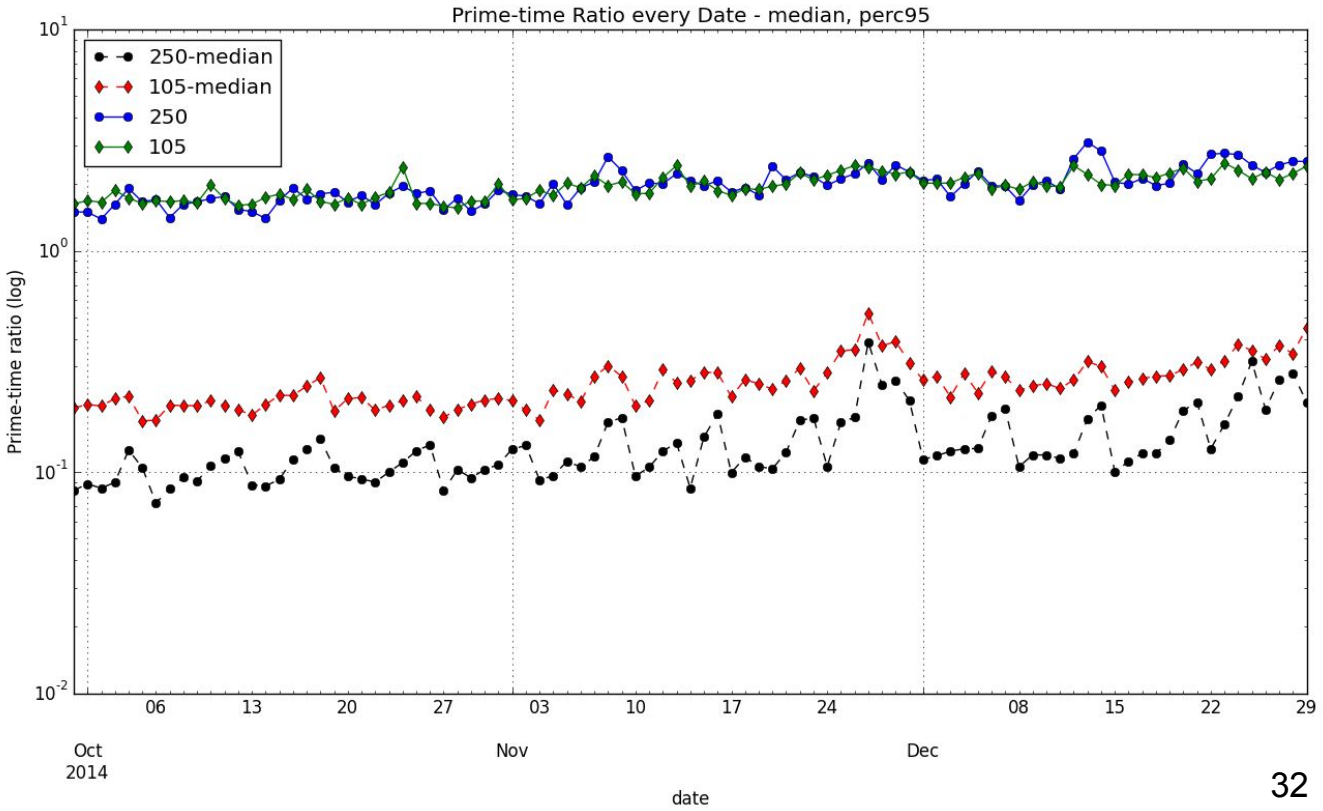
Downlink Prime-Time Ratio: Control Set > Test Set PT

- Diurnal pattern shows higher ratio for 5 weekdays, and low ratio on weekends.
- The median PT ratio over subscribers per day has decreased due to treatment. Why?
- Is there an increasing trend?



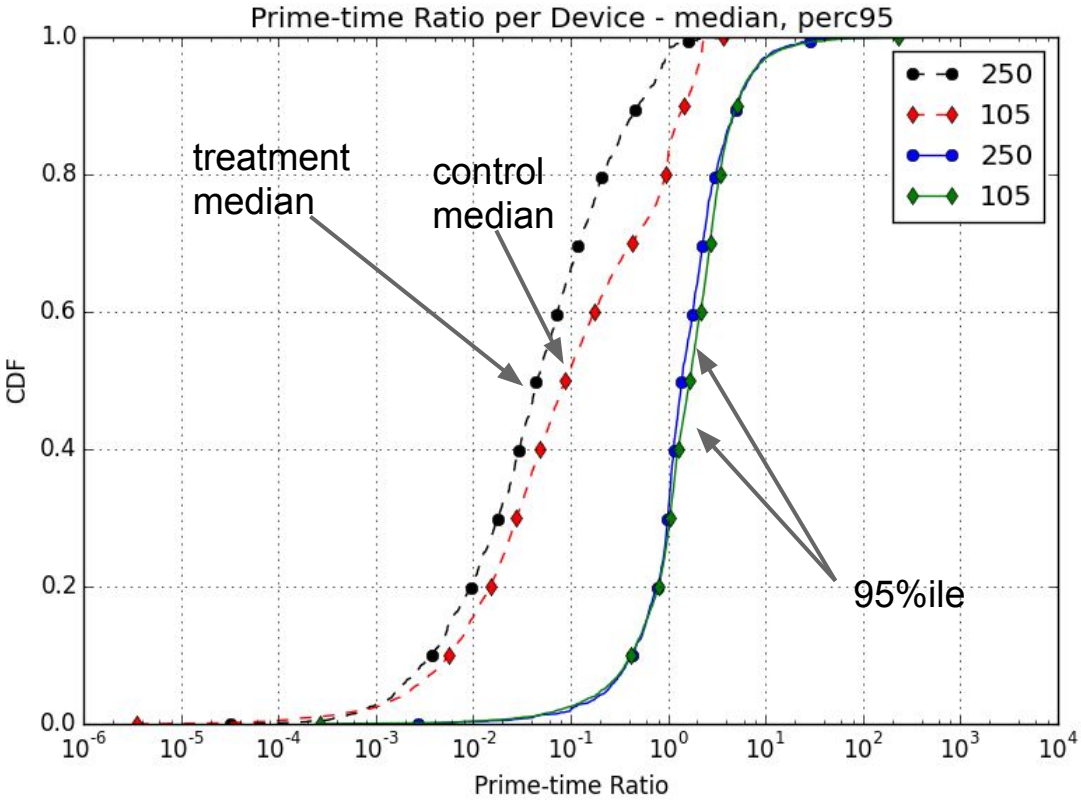
Uplink Prime-Time Ratio: Control Set > Test Set PT

- Similar to downlink, but shows a clearer increasing trend overall.
- Treatment decreases the median PT.
- Is this seasonal, or is internet usage during prime time really increasing this fast?
- This data needs longitudinal study



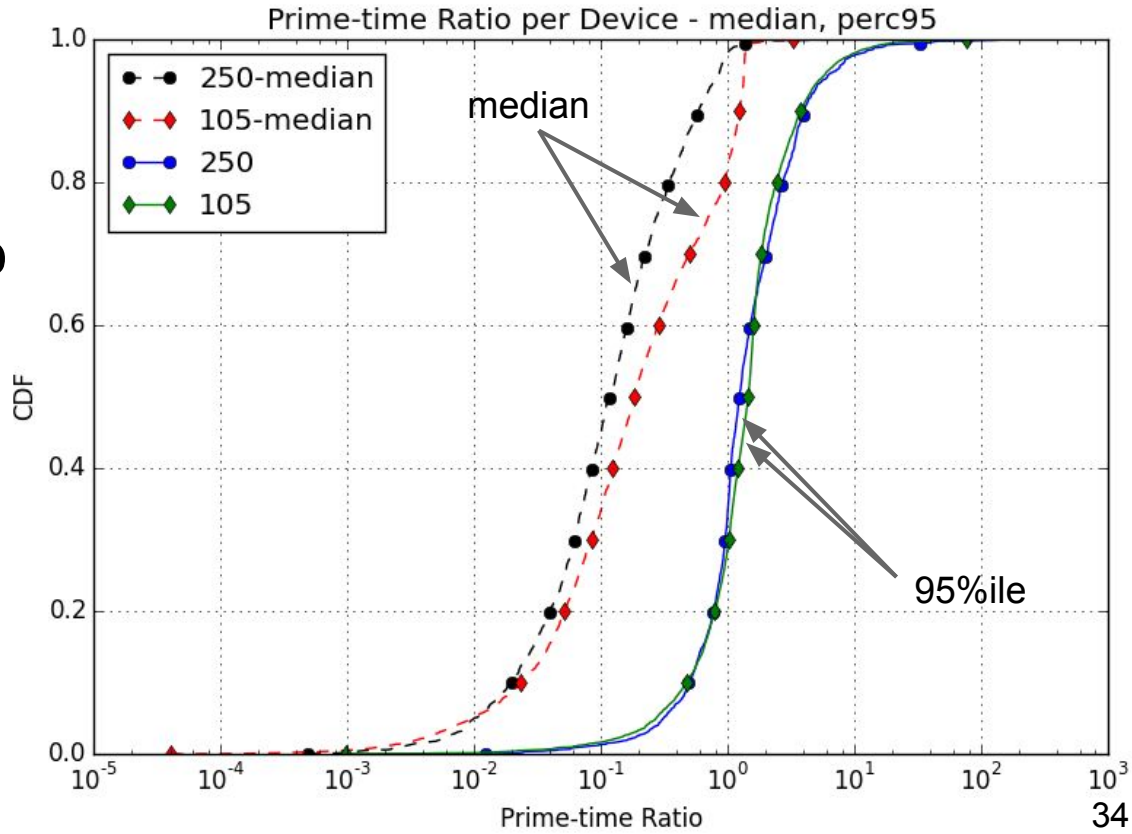
Downlink Prime-Time Ratio: Changes in PT due to treatment

- Prime time ratio per subscriber (median over its lifetime) **decreases** due to treatment. Even though time series of prime time did not indicate this)
- The 95% PT (over subscribers) does not vary significantly



Uplink Prime-Time Ratio: Distribution of PT over Devices

- Similar decrease in median uplink prime-time ratio due to treatment

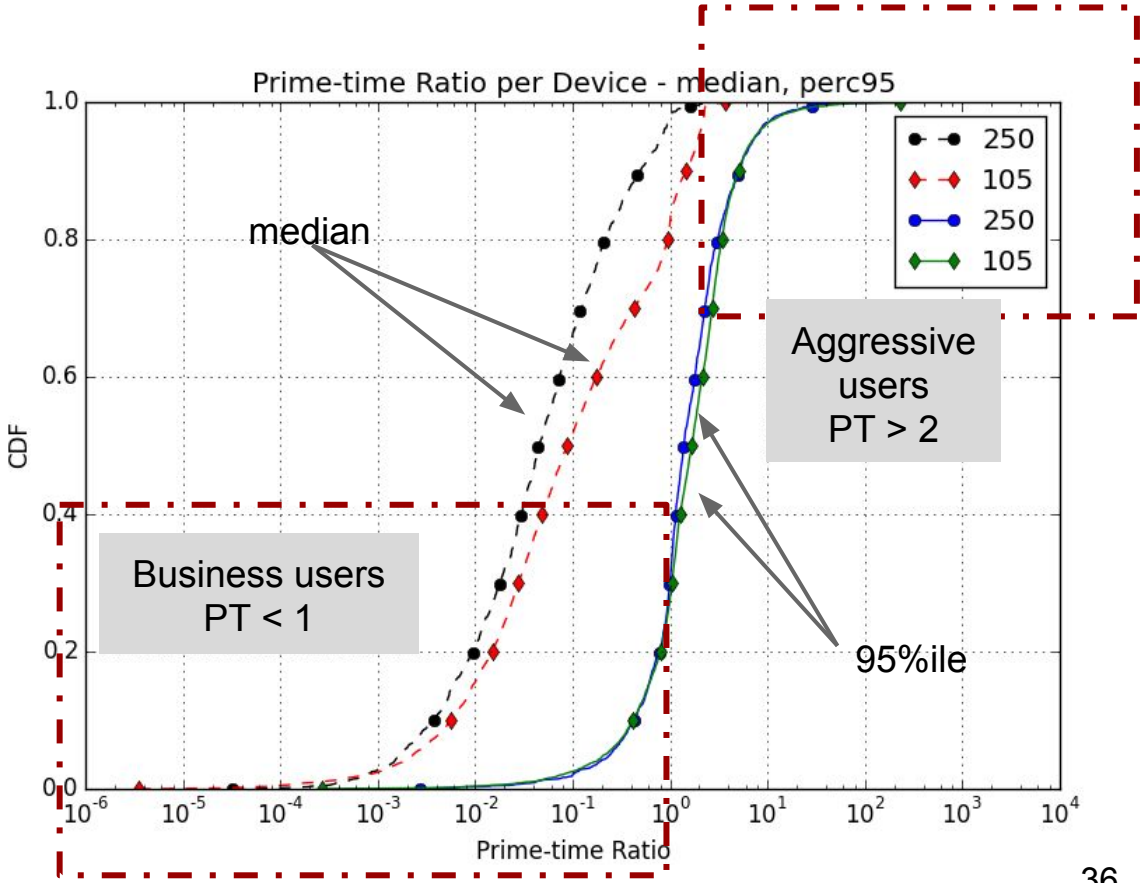


User Taxonomy according to PT

- Some users have a PT < 1
  - They were more aggressive in off-peak hours (day time) than peak hours (evening)
  - As data is from residences, these could be work at home
- Taxonomy based on user behavior (time and quantity)
  - Business users (PT < 1)
  - Streamers (PT > 2)
  - Aggressive all the time (PT~1, usage > avg)
  - Non-aggressive (PT~1, usage < avg)

User Taxonomy based on PT and Usage

- Based on 95 %ile PT ratio of the subscriber
- 30% PT < 1: possibly businesses with morning work-hours
- 20% PT > 2: aggressive prime-time streamers



Fin.

For sanitizing, analysis, and plotting scripts please contact: sgrover@cs.princeton.edu