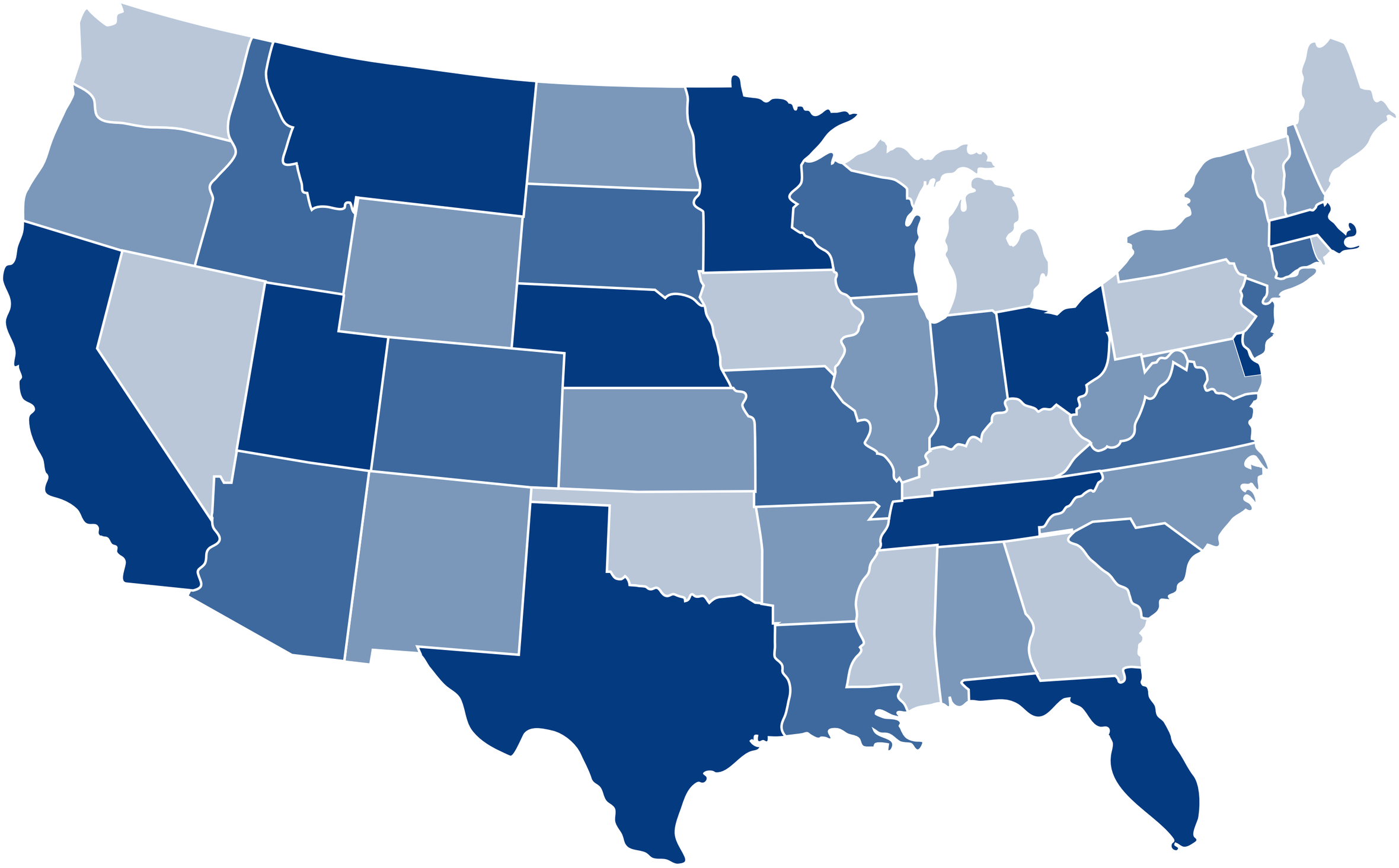

Measuring Broadband America

Validated Data Cleansing, Thirteenth Report

This document is based on September/October 2021 data. It outlines the data cleansing processes used to generate the ‘validated’ dataset from the ‘raw’ dataset. The ‘validated’ September/October 2022 data published by the FCC will have had these operations performed upon it.

Measuring Broadband America

Information

This document is based on September/October 2022 data. It outlines the data cleansing processes used to generate the ‘validated’ dataset from the ‘raw’ dataset. The ‘validated’ September/October 2022 data published by the FCC will have had these operations performed upon it.

Measuring Broadband America 13th Report Validated data can be found at:

https://data.fcc.gov/download/measuring-broadband-america/2023/validated-data-sept2022.tar.gz

The SQL scripts used to conduct these tasks are available in the file:

https://data.fcc.gov/download/measuring-broadband-america/2023/sql-cleanup-scripts-sept2022.tar.gz

Data Processing Flow Validation Operations

Remove all data from September 24th to September 27th 2021 and from October 5th 2021

- The FCC reporting period ran between the following dates: September 12-13, 16-21, October 4-6, 11-13, 15-17, 19-31 in 2022, referred to as the "September-October 2022 reporting period," were analyzed to generate the charts.
- Data outside of the date ranges above was removed from the dataset.
- Only participants who provided a minimum of 5 days of valid measurements during the testing period were included in the September / October 2022 results.
- Data was only charted when results from at least 45 separate Whiteboxes was available. Instances of 44 or fewer Whiteboxes were noted for possible future augmentation.

Handle panelists that changed ISP intra-month

Some panelists changed ISP mid-month. In situations where this occurred we removed the data for the ISP that they spent the shortest period on. For example, if the panelist changed from ISP A to ISP B on September 10th, we would remove data prior to this date because there would be a larger dataset (i.e., September 11th to the 30th) for their performance on ISP B.

A daily log of the panelists' public addresses was used to determine when they changed ISP. This table records on a daily basis the owner of the netblock that the panelist's public facing IP address resides in. This allowed us to identify when people changed ISPs quickly and reliably.

We had two mechanisms for identifying panelists that changed service tiers:

Firstly, ISP-supplied panelist validation information informed us of which service tier the panelist was subscribed to. Some ISPs also provided the date at which they began this service.

However, in cases where the ISP was unable to validate the panelist in question or their validation was delayed, we used the following process:

- Find the difference between the average sustained throughput observed for the first three days in the reporting period from the average sustained throughput observed for the final three days in the reporting period (if the unit wasn't online at the start or end, then we take the first/final three days that they were actually online.
- If this difference is over 50%, we examined the downstream and upstream charts for this unit.
- Where an obvious step change is observed (e.g. from 25Mbit/s to 50Mbit/s), flag the data for the shorter period for removal.

General Data Cleansing

- Only data from specific units was considered for analysis (see unit_to_include.csv).
- Test times were translated from UTC to local time according to each unit's timezone (detailed in unit_timezones.csv).
- Some units had part of the month's data excluded to account for mid-month moves, as described above. See unit_days_to_include.csv for details of which days' data were kept for each unit.
- Only the httpget (downstream throughput), httppost (upstream throughput), udplatency (UDP latency/loss) and webget (web browsing) results were considered for the FCC analysis.
- All results from non-Stackpath targets were removed from the httpget, httppost and udplatency tables.

Speed test (download, upload) cleansing

- Only tests which ran over multiple TCP connections and over IPv4 were considered in the analysis.
- All failed tests were removed. Failed speed tests were not considered in the analysis.

UDP latency/loss cleansing

- All test instances (one per hour, per unit) with less than 50 samples (out of a potential maximum of 600) were removed.
- All test instances where a unit's packet loss exceeded 10% within a single hour were removed. Such a high level of loss would render a connection unusable and is considered an anomalous event.
- All test instances where any round trip time was reported as 0.5ms or lower were removed.
- All test instances where the range of a unit's of individual round trip times exceeded 300ms were removed.
- Only tests which ran over IPv4 were considered in the analysis.

Web Browsing cleansing

- All test instances where the page load time exceeded 30 seconds were removed.
- All test instances where the page load time was lower than 0.5ms were removed. Such low values would indicate that the page was not requested from the internet. For example, some intermediate network device (such as the CPE) could have intercepted the request and replied with its own content (usually an error page when the connection is down). Such cases are not representative of fetching the real webpage.
- All failed tests were removed. Failed webpage load tests were not analyzed in this report.