Final Report

This document summarizes the work done for the BU Spark! project, Spring 2021, MAPC Broadband Digital Equity in MA. This document represents the final report for our work this semester.

Date: 04/19/21

Student Team

This project has two different teams, denoted MAPC team 1 and MAPC team 2. We represent team 2. There are five students, and one project manager for team 2:

- Adam Streich
- Jenny Li
- Nathan Lauer
- Yutong Shen
- Zhixing Zhao

The project manager is Kamran Arif.

Contact

- Ryan Kelly, RKelly@mapc.org Digital Services lead at the MAPC
- Matt Zagaja, mzagaja@mapc.org , Lead civic web developer at the MAPC

Organization

The Metropolitan Area Planning Council - MAPC

Purpose

In this deliverable, we present our work for MAPC regarding the MLAB and Ookla datasets. This is our final report; we present the motivation for the work, the steps we took throughout the semester, results found, and a discussion of these results.

Table of Contents

Abstract

In this work, we construct datasets of measured internet speeds from two different organizations, MLAB and Ookla, in the year 2020, for the state of Massachusetts. MLAB measures speed when someone queries Google along the lines of "how fast is my internet," and measures a simulated network request as if it propagated across a significant portion of the larger internet network. Ookla, on the other hand, measures speed at speedtest.net, and presents a measurement of someone's local ISP server's speed. We further analyze this data on a per municipality basis, and the MLAB data on a per provider basis. We present descriptive statistics of internet speeds during 2020 for the state of Massachusetts, and maps of upload and download speeds for each municipality in the state of Massachusetts. We also present this data as correlated against household income data from the 2014-2018 census. Finally, we discuss a number of key findings in the analysis of these datasets. First, there is a significant difference between measured Ookla speeds and measured MLAB speeds. Second, there exists an upwards correlation between MLAB broadband speeds and median household income; as median household income increases, average broadband speed increases as well. Third, the vast majority of municipalities are significantly under the desired 100/100 download/upload speeds in mega-bits-per-second (Mbps), with many not even reaching 50 Mbps. Fourth, there exists significant disparity in broadband coverage across the state.

Motivation

The Metropolitan Area Planning Council (MAPC) provides planning capacity to municipalities within Massachusetts in a number of different capacities. Recently, they have turned their attention towards broadband, by trying to help municipalities better understand the available internet broadband within their region. Our team joined the MAPC in this endeavor, to collect and analyze broadband data from MLAB and Ookla.

While it may seem trivial, it is actually not so easy to answer questions about internet broadband such as:

- How fast is my internet?
- Is my internet fast enough?
- What providers are available to me, and are there differences in their broadband speeds?

For example, measuring internet speed might amount to a simple measure of the speed of a local ISP server, or be as complex as measuring a real-time observed speed of a fetch request from some geographically distant server. Further, while many ISPs may claim to be able to provide certain speeds, it may not be so clear that the observed speeds match the marketed speeds.

Thus, we are working with the MAPC to try and build a dataset that can answer some of these questions, and provide a basis for answering questions that may inform public policy, such as:

- Is there a correlation between median household income and broadband speed?
- Are there municipalities where the available broadband options do not meet requirements for nominal modern internet usage -- say for example, with large zoom calls being nearly ubiquitous for remote schooling and work -- and how limited are they?
- Can we observe differences in broadband access among different providers in different areas of the state?

To answer these questions, we built datasets from Ookla and MLAB of internet speed measurements throughout the 2020 calendar year. We also pulled household income data from the 2014-2018 census. Aside from the aggregation of data, we also provide analyses of the data, which are sufficient starting points for informing policy decisions.

Ookla Dataset

Ookla provides a service at speedtest.net, where clients can test their internet speed. Broadly speaking, the measurement taken is a measure of the speed of a client's connection with their internet provider or ISP. This differs somewhat from the MLAB data, where Ookla is nominally measuring the broadband provided by an ISP, as opposed to simulating a nominal request across the entire network.

Initial Steps

We started by reading the documentation provided by Ookla, hosted on their public facing GitHub page, at https://github.com/teamookla/ookla-open-data. Here, they provide instructions for accessing their open data, which is the portal we are using to analyze their data in Massachusetts. They provide three levels of access: directly via AWS S3, a few download links, and CLI tool for downloads to the terminal. Additionally, data files are provided in two formats: shapefiles and parquet files.

We started by downloading one of the example shapefile links, but found this to be quite confusing. We were not quite able to find a manner to view the data, or begin to understand it for analysis purposes. We then tried the parquet format, which was more accessible. Parquet is a columnar storage system provided via the Apache Organization, and can be used with any tool in the Hadoop architecture. It is also easy to integrate with python; we managed to find simple "parquet-to-json" and "parquet-to-csv" tools through python Pandas.

AWS and Programmatic Access

For more regular access, we created our own AWS account. We created a root user, and each of us set up our own IAM accounts. For this, we installed the AWS CLI tools, so that we could each access AWS S3 from our local terminals, and for programmatic access.

Date Pre-Processing

In the subsequent data pre-processing procedure, we discovered that the data could be accessed and read directly from the API Endpoints of Ookla with the aids of “GeoPandas” package. The package will read in the shapefiles as a Pandas DataFrame, which makes the data easier to clean. Thus, using the AWS S3 urls provided by Ookla, we used GeoPandas to access the data for each quarter of 2020.

That data, however, was not limited to Massachusetts. In order to downsample the data to just Massachusetts, we also used GeoPandas, on a list of the boundaries for each county in MA obtained from the Census Bureau. Then, since the data is provided by Ookla in shapefile format, we were able to run a joins operation on these two data sets. This yielded just the subset of data that is within a county boundary of some county in Massachusetts. Further, we were then able to label each data point with its associated county.

Labeling by Municipality

After deliverable 1, the Ookla data was labeled with county information, as this information is publicly available and easy to obtain. Unfortunately, labeling this data by county does not correlate well with the previous work done by MAPC; as much of MAPC's work is grouped by municipality -- a finer grain resolution than by county -- it was necessary to further granularize the data points in Ookla, by labeling each data point with a specific municipality.

Fortunately, MAPC already had a dataset with a geographically defined area for each polygon. That data can be found here: https://datacommon.mapc.org/browser/datasets/390. Using this dataset, we were able to label each row in the Ookla data with municipality information.

Data Schema

This section describes the schema of the data, and we provide an example data point for reference.

Columns:

- quadkey: a key that identifies the tile
- avg_d_kbps: the average download speed in kilobits per second within the tile
- avg_u_kbps: the average upload speed in kilobits per second within the tile
- avg_lat_ms: the average latency of the tests in this tile.
- tests: The number of tests that contributed to the other values in this tile.
- devices: The number of unique devices that contributed to the data in this tile.
- geometry: list of latitude/longitude pairs, that collectively form the polygonal shape of this tile.
- index_right: joined column index number
- objectid: identifier for object from joined table
- muni_id: numeric identifier for the municipality.
- municipal: name of the municipality
- shape: list of latitude/longitude pairs, that collectively form the polygonal shape of this municipality.

Example data point:

0302332123102031,240473,108651,9,5,3,"POLYGON ((-71.1749267578125 42.252917783302, -71.16943359375 42.252917783302, -71.16943359375 42.2488517007209, -71.1749267578125 42.2488517007209, -71.1749267578125 42.252917783302))",216,228,73,Dedham,8E080000660100000800100 06A690000B621070001000000A5BAB4CEB2159EC09FB9A411E8F109A0EB0AE896CE019CF7D403C0F702A0BF08FC970488CC09ECD6068CE306ECE90888BF05D0A530ACF90BE4E96B8CA523FCBC4BA4EF1EE8B027A8DA19D0E313D0B024C094AD05F8B0C009FCB97EF4CAEC01F4820AE4F414F0C5

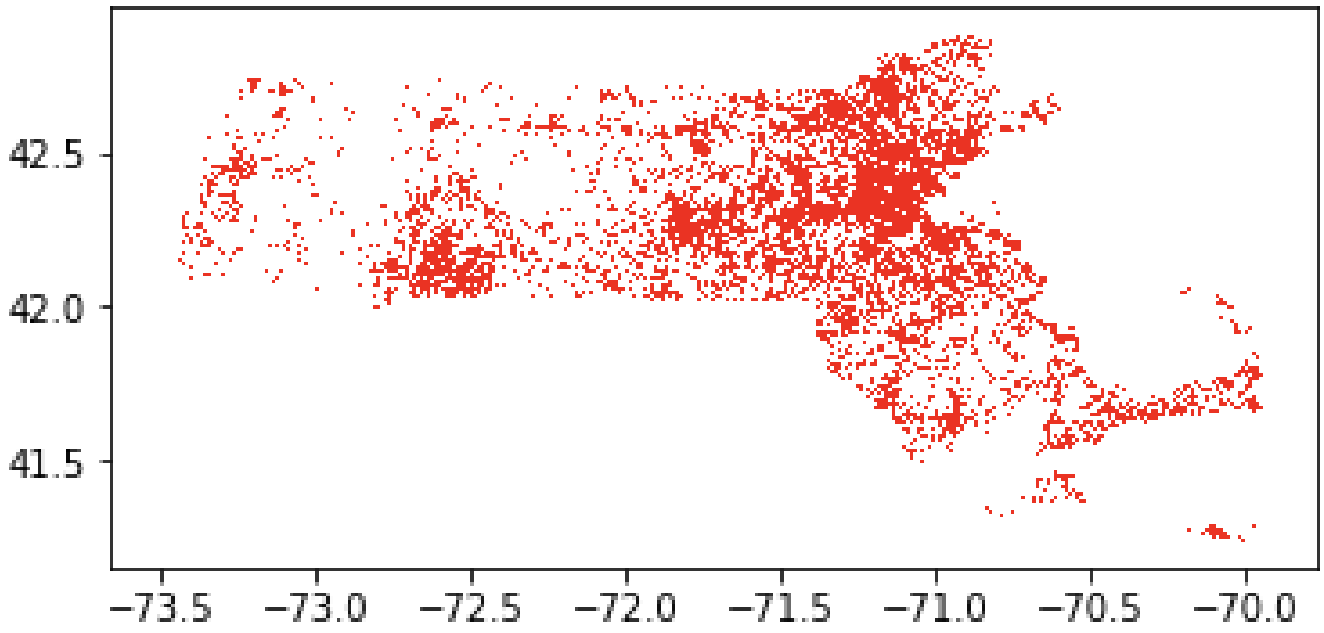
Data Size

Because Ookla provides aggregated information, and we have currently limited our time range to just 2020, the Ookla data is relatively small. We provide the file sizes here

- quarter 1: 588 KB
- quarter 2: 9.2 MB
- quarter 3: 9.5 MB
- Quarter 4: 8 MB

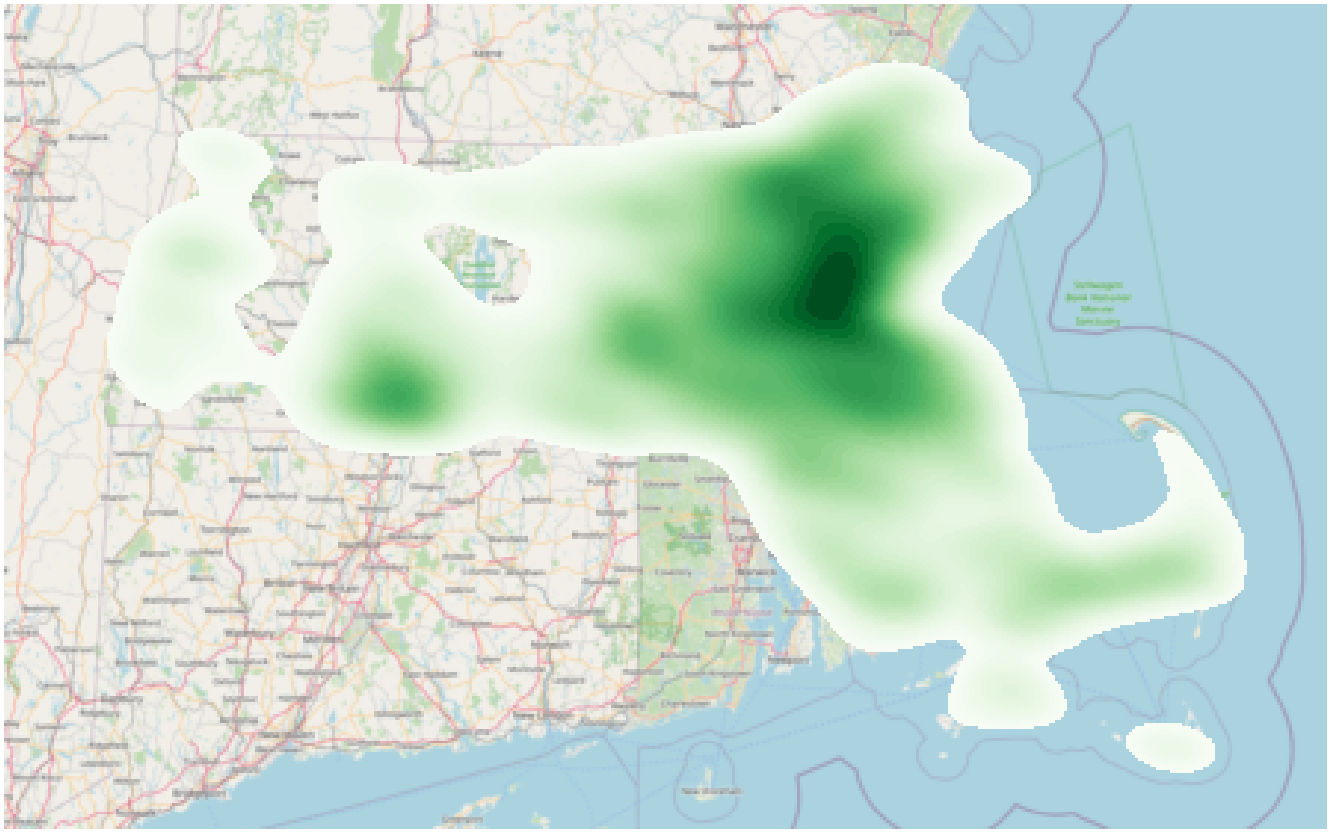
Ookla Geographical Density

Since Ookla data is constructed with tiles labeled with geographical information, we were able to produce density maps showing where in the state the measurements were collected. Here is that map:



The vertical axis is latitude, and the horizontal axis is longitude.

We also present this data as a heatmap:



From the scatter plot of the location of each data point, and the heatmap, we can see that the data around the Boston area and Springfield area are denser than average, and there are only a few data points in rural areas. In fact, there are many areas within the state that either don't have coverage at all, or have only minimal amounts of broadband coverage.

Ookla Basic Statistics

Top 25 Cities with Fastest Average Download Speed

City	Average Download Speed, Kbps
West Bridgewater	205943.28276
Montgomery	205546.3
Bridgewater	199698.68388
Gardner	197069.1943
East Bridgewater	191234.58481
Whitman	187003.88316
Ware	185472.76952
Erving	184774.20513
Everett	184351.24286

City	Average Download Speed, Kbps
Dracut	183783.50973
Palmer	182919.57853
Ashby	181227.69079
Revere	180672.46409
Peabody	179239.06
Haverhill	178558.80971
Lowell	177470.17117
Norton	176841.28748
Chelsea	175945.49351
Fall River	175309.47971
Monson	174890.07237
Milton	174573.90667
Groveland	174141.37981
Westminster	173648.61179
Boston	173572.17082
Middleborough	172813.69739

Cities with the Most Average Number of Devices

City	Average Number of Devices
Somerville	24.791
Cambridge	21.39189
Brookline	18.10905
Boston	16.2856
Malden	15.30633
Everett	15.03333
Chelsea	14.59091
Watertown	13.70169
Medford	11.64208
Arlington	11.12895
Revere	10.8232

City	Average Number of Devices
Quincy	10.56224
Belmont	10.22945
Newton	9.78877
Lawrence	9.05793
Lowell	7.88031
Provincetown	7.71429
Waltham	7.36449
Melrose	7.34375
Winthrop	7.25373
Lynn	7.14741
Winchester	7.04709
Swampscott	6.73604
Salem	6.61962
Worcester	6.52941

MLAB Dataset

The Measurement Lab is an open source project, and aims to advance internet research by providing useful information to anyone about their internet performance. Notably, if you type something along the lines of "how fast is my internet" into Google, MLAB will execute the operations necessary to measure the speed. Unlike Ookla, MLAB attempts to give a more realistic sense of internet speeds, within the context of the larger internet network, and not limited to the speeds provided by a specific ISP.

MLAB is particularly useful, because all of their data is accessible for free, and their tools are entirely open source.

Data Access and DownSampling

The MLAB data is provided primarily through SQL query access, via Google BigQuery. Notably, the size of the data is massive. The client provided us with some example queries; we executed a modified version of these for all of MA in the year 2019, and some ~680 GB of data was returned.

The MLAB data is quite large, and thus it was necessary to downsample the data, so that we could work on it effectively. To do so, we sampled a small, but representative portion of the data: we aggregated all measurements in 2020 that occurred within a time range of 8:00am-8:30am, 12:00pm - 12:30pm, 3:30pm-4:00pm, and 8:00pm-8:30pm. The idea here is that these are normal times when people are accessing the internet, and importantly, during school hours when students may be expected to access the classroom via Zoom, or some other digital streaming interactive service. We argue that this downsampling approach is reasonable, because:

- The sampled times are during critical hours of the day
- The amount of data should be significantly smaller than the entire 2020 data set.
- It seems a fair assumption that these times are representative of normal to heavy broadband use, which is the subset of data that is most important.

MLAB Automated Scripts

Google BigQuery provides a webportal where SQL queries can be run, and they also provide client libraries where queries can be executed against BigQuery from within an external script. It was a bit tricky to set up, but we managed to obtain BigQuery access through python, using the Google Cloud SDK.

With this, we set up the following query:

```
SELECT
  a.TestTime AS TestTime,
  NET.SAFE_IP_FROM_STRING (client.IP) AS IP,
  a.MeanThroughputMbps AS MeanThroughputMbps,
  a.MinRTT AS MinRTT,
  client.Geo.city AS City,
  client.Geo.Latitude AS Latitude,
  client.Geo.Longitude AS Longitude,
  client.Network.ASNumber AS ProviderNumber,
FROM
```

```
  `measurement-lab.ndt.unified_uploads`
WHERE
  client.geo.CountryCode = "US"
  AND client.Geo.region = "MA"
  AND date BETWEEN "{year}-{month_num}-{day_num}"
  AND "{year}-{month_num}-{day_num}"
  AND (
    (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num}
{first_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {first_hour}:30:00.000",
"UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num}
{second_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {second_hour}:30:00.000",
"UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num}
{third_hour}:00:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {third_hour}:30:00.000",
"UTC")
    )
    OR (
     a.TestTime BETWEEN TIMESTAMP("{year}-{month_num}-{day_num}
{fourth_hour}:30:00.000", "UTC")
     AND TIMESTAMP("{year}-{month_num}-{day_num} {fifth_hour}:00:00.000",
"UTC")
    )
  )
```

Note: it was necessary to filter by both date and TestTime, as it appears that BigQuery uses the date field as a method of distributing work across servers. Without this filter, an error was returned.

Data Schema

This section describes the schema of the data, and we provide an example data point for reference.

Unlike the Ookla data, this data is not aggregated, and each data point represents an individual measurement. This likely means that even with the data provided as is, there is some useful cleanup and preprocessing that can be done, such as aggregating measurements from the same device within the timeframe, and other such things.

Columns:

- TestTime: the time at which the test took place, includes both a date and a time.
- IP: the IP address of the device being tested
- MeanThroughputMbps: average broadband throughput in megabits per second. Note that this is different than the download and upload speeds of Ookla, since they are measuring slightly different things.
- MinRTT: Minimum round trip time, the time it takes to send a signal or data packet and receive back the corresponding acknowledgment
- City: the city in which the test occurred
- Latitude: latitude location of the test
- Longitude: longitude location of the test
- ProviderNumber: the autonomous system number of the nearest server system for the test.

Example data point:

2020-10-01 12:01:22.316165 UTC, TBPhGA==, 5.924981493972454, 21.283, Wellfleet, 41.9289, -70.0186, 7922

Data Size

Since we have limited the MLAB data to 2020, and within just 4 time slots (8:00am-8:30am, 12:00pm-12:30pm, 3:30pm-4:00pm, and 8:00pm-8:30pm), then csv file for MLAB is large, but not unreasonably large. It contains 437432 rows of data, and is approximately 49MB in size.

Labeling MLAB Data

Unlike the Ookla data, the MLAB data was already labeled with municipality, and thus the step of labeling each data point by municipality was unnecessary. The MLAB data was also labeled with "ASNumber," which refers to the number assigned to the Autonomous System which controlled the network via which each speed test was conducted. Unfortunately, the data did not come with the name of the organization that operates each Autonomous System, and therefore it became necessary to map each of these numbers to a well defined organization.

To do so, we used a publicly available listing of Autonomous Systems, found here: http://www.bgplookingglass.com/list-of-autonomous-system-numbers

With this information, we added a new column to the MLAB schema, containing the name of the Provider that runs each of the various Autonomous Systems. Unfortunately, this type of information is not obtainable with the

Ookla data, and therefore this labeling was particularly important, in order to be able to run analyses on a per-provider level.

MLAB Descriptive Statistics

With the MLAB data labeled with providers, we computed a number of descriptive statistics over the entirety of the dataset, to get a better understanding of the data contained within. In particular, we produced the following metrics:

Basic Statistics

- average MeanThroughputMbps: 43.7
- median MeanThroughputMbps: 12.6
- mode MeanThroughputMbps: 0 11.8
- Standard Deviation MeanThroughputMbps: 91.66

Top 5 Fastest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
HGE-NET - Holyoke Gas & Electric Department	192.128319
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	134.443643
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	82.131921
ASN-QWEST-US NOVARTIS-DMZ-US	49.514844
ALKERMES - ALKERMES INCORPORATED	23.363405

Botton 5 Slowest Providers on Average, with at least 1,000 Measurements

Provider Name	Average Mbps
SPCS - Sprint Personal Communications Systems	2.376271
CELLCO - Cellco Partnership DBA Verizon Wireless	5.019919
ASN-SHREWS - Shrewsbury Electric and Cable Operations	6.070376
T-MOBILE-AS21928 - T-Mobile USA, Inc.	8.656454
RR-NYSREGION-ASN-01 - Time Warner Cable Internet LLC	9.678461

Counts of Measurements per Provider, with at least 1,000 Measurements

ProviderName	Measurement Count
COMCAST-7922 - Comcast Cable Communications, Inc.	182906
UUNET - MCI Communications Services, Inc. d/b/a Verizon Business	89356
ALKERMES - ALKERMES INCORPORATED	68790

ProviderName	Measurement Count
CHARTER-NET-HKY-NC - Charter Communications	21136
RCN-AS - RCN	15222
ASN-QWEST-US NOVARTIS-DMZ-US	12355
CELLCO - Cellco Partnership DBA Verizon Wireless	6878
T-MOBILE-AS21928 - T-Mobile USA, Inc.	6453
RR-NYSREGION-ASN-01 - Time Warner Cable Interne...	6096
SPCS - Sprint Personal Communications Systems	3502
ASN-SHREWS - Shrewsbury Electric and Cable Oper...	2579
LIGHTOWER Lightower Fiber Networks (LIGHT-141)	1369
HGE-NET - Holyoke Gas & Electric Department	1130

Top 25 Cities by Measurement Count

City	Measurement Count
Needham	34697
Boston	23807
Somerville	15488
Bedford	14515
Ashland	13688
Cambridge	11789
Devens	9150
Worcester	7709
Springfield	7146
Acton	6645
Dorchester	6479
Watertown	6094
Brighton	5325
Arlington	5212
Brookline	5135
Concord	5022
Newton Center	4791

City	Measurement Count
Wellesley Hills	4639
Lexington	4532
Lowell	4415
Quincy	4402
Waltham	4274
Milton	4024
Framingham	3870
Andover	3795

Splitting MLAB Data into Sub-Datasets by Municipality

Since the desired granularity for the data is on the municipal level, we produced a series of csv files each limited to the MLAB data that was collected only within the relevant municipality. This would allow for easier analysis within each of the municipalities, since the overall data set is quite large. Thus, should some analysis focus on a particular municipality, or a short list of municipalities, these files would come in handy.

There are 101 municipalities within MAPC purview, and therefore we produced one output csv file for each of these. They each have the exact same schema as the overall MAPC data, but with data limited to the relevant municipality. For example, there is a single for Acton MA, with a total of 6645 data points contained.

Computing the Average Broadband Speed Per Provider Per Municipality

Here, we produced a csv file with the average broadband speed of each provider, in each municipality. For example, here is an excerpt from that file:

...

Abington,BIGLEAF - Bigleaf Networks LLC,1.9385489829867468,44.9945

Abington,CELLCO - Cellco Partnership DBA Verizon Wireless,14.1583484092067,53.14081818181819

Abington,"COMCAST-7922 - Comcast Cable Communications, Inc.",10.609942074630164,25.27878453038673

Abington,LIGHTOWER Lightower Fiber Networks (LIGHT-141),46.193328643945755,22.193

Abington,"UUNET - MCI Communications Services, Inc. d/b/a Verizon Business",175.27244063910456,14.569966666666653

...

The first column here is the name of the municipality; in this case, we are focused on Abington. The second column is the name of the Provider, or more specifically, the name of the organization that runs the Autonomous System via which a given test was conducted. The third column is the average

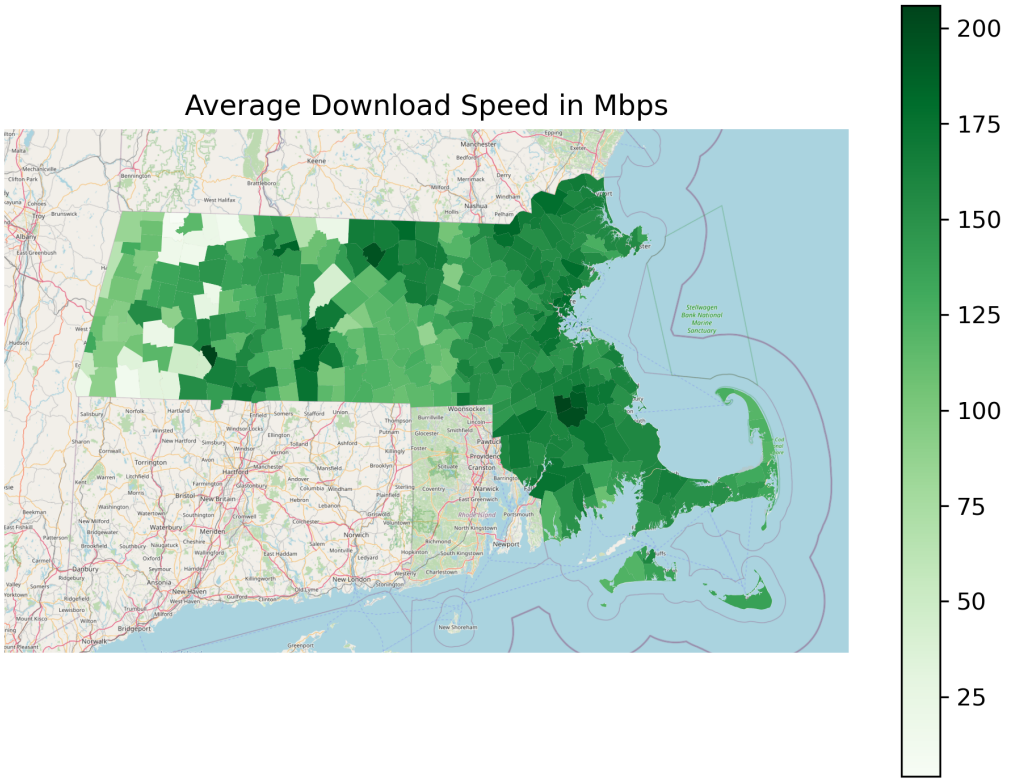
MeanThroughputMbps, and the fourth column is the average MinRTT, which stands for Minimum Round Trip Time.

As can be seen here, Lightower had an average MeanThroughputMbps of 46.2 megabits-per-second, while Verizon was considerably faster, at 175.27 megabits-per-second.

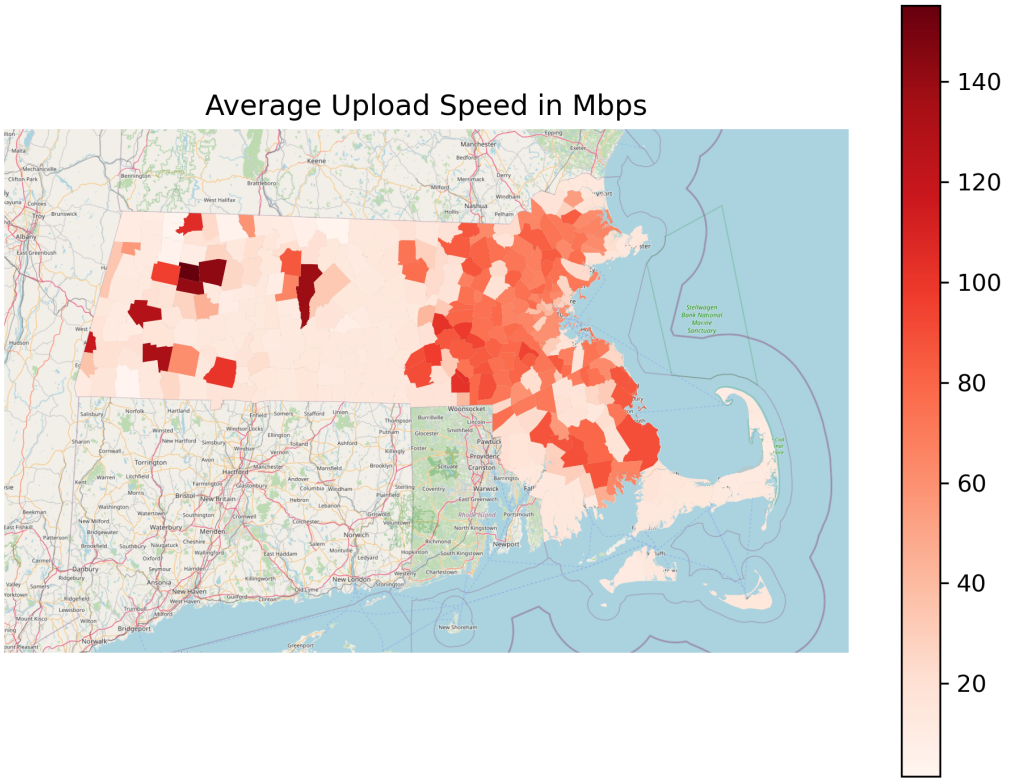
Results

There are a number of results which we present here. First, for the Ookla dataset, we show maps of the download and upload speeds across the state of MA, delineated per municipality. Second, we show screenshots of a final interactive html file which shows tooltip information about Ookla upload and download speed for each municipality. Third, we present scatter plots of measured broadband speeds against median household income from the 2014-2018 census. Finally, for the MLAB data, we present similar scatter plots for a number of the primary providers.

Ookla Upload and Download Speeds



Here, we chart the download speeds in mega-bits-per-second (Mbps), per municipality. Each city is colored by the average download speed across 2020, where darker green means faster speeds (up to 200 Mbps), and lighter green/white means slower download speeds. As can be seen, there is a significant difference in observed speeds for the various municipalities.



Here, we see the upload speeds, per municipality in MA. There are a number of items to note:

- There are large discrepancies between download and upload speed, in many cities.
- In general, upload speeds are much slower than download speeds. The max download speed was near 200 Mbps, while the max upload speed is closer to 140 Mbps, nearly 60Mpbs slower!
- There is an entire section in the middle of the state where there is effectively no upload broadband available. Within this region, many of the municipalities have respectable download speeds, despite the lack of upload abilities.

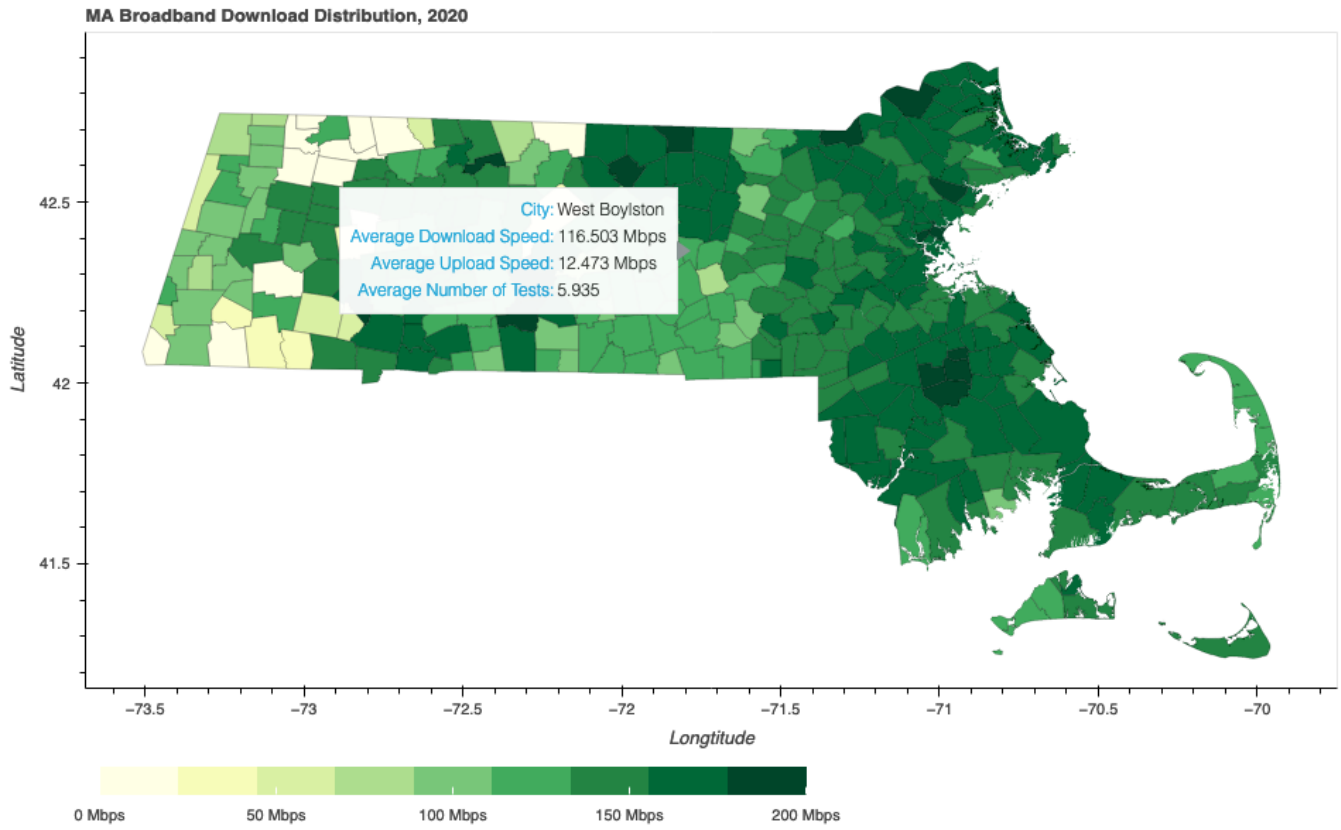
The MAPC was concerned that there has existed an imbalance between download and upload speeds for some time, because people often care more about download speed than upload speed. Unfortunately, this problem is particularly important during the present time, since due to the pandemic, there is an extensive amount of video communication ongoing for both business and education.

Ookla Interactive HTML Map

We generated an html file, that allows a user to interact with this data in a visual sense. The page presents a map of the state of Massachusetts, with each municipality colored according to measured download broadband speeds. Then, as a user mouses over each of the municipalities, a tooltip appears with four data points:

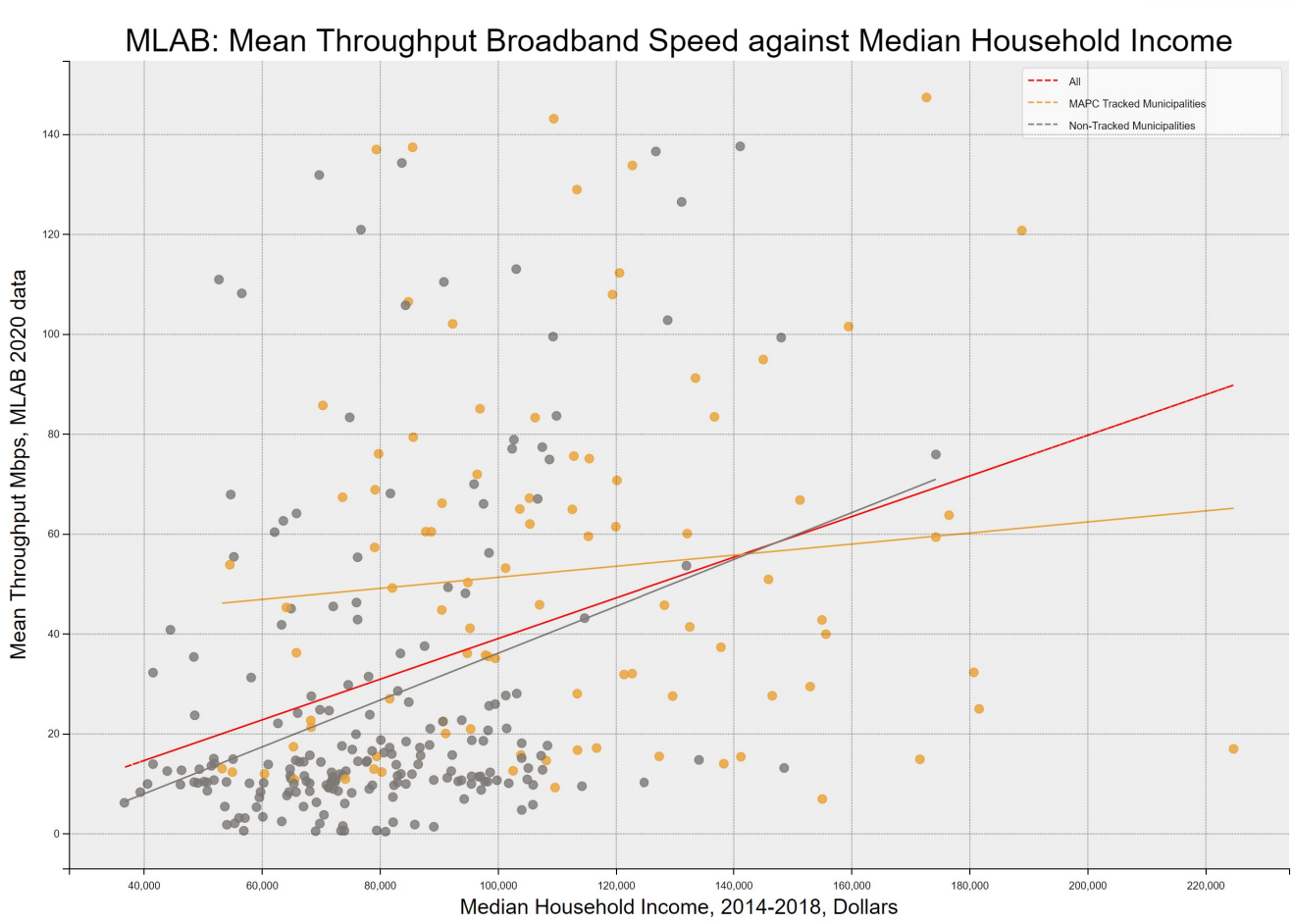
- The name of the municipality
- The measured average download speed in that municipality for the 2020 calendar year
- The measured average upload speed in that municipality for the 2020 calendar year

- The average number of tests taken in that municipality.



Here, we show a screenshot as an example of a user interacting with this HTML page, and in this case, hovering over the municipality named West Boylston. Although the presentation here is just a screenshot, we provided the full HTML webpage as a deliverable unto itself for the MAPC, for use in their continuing analysis.

MLAB Data and Median Household Income



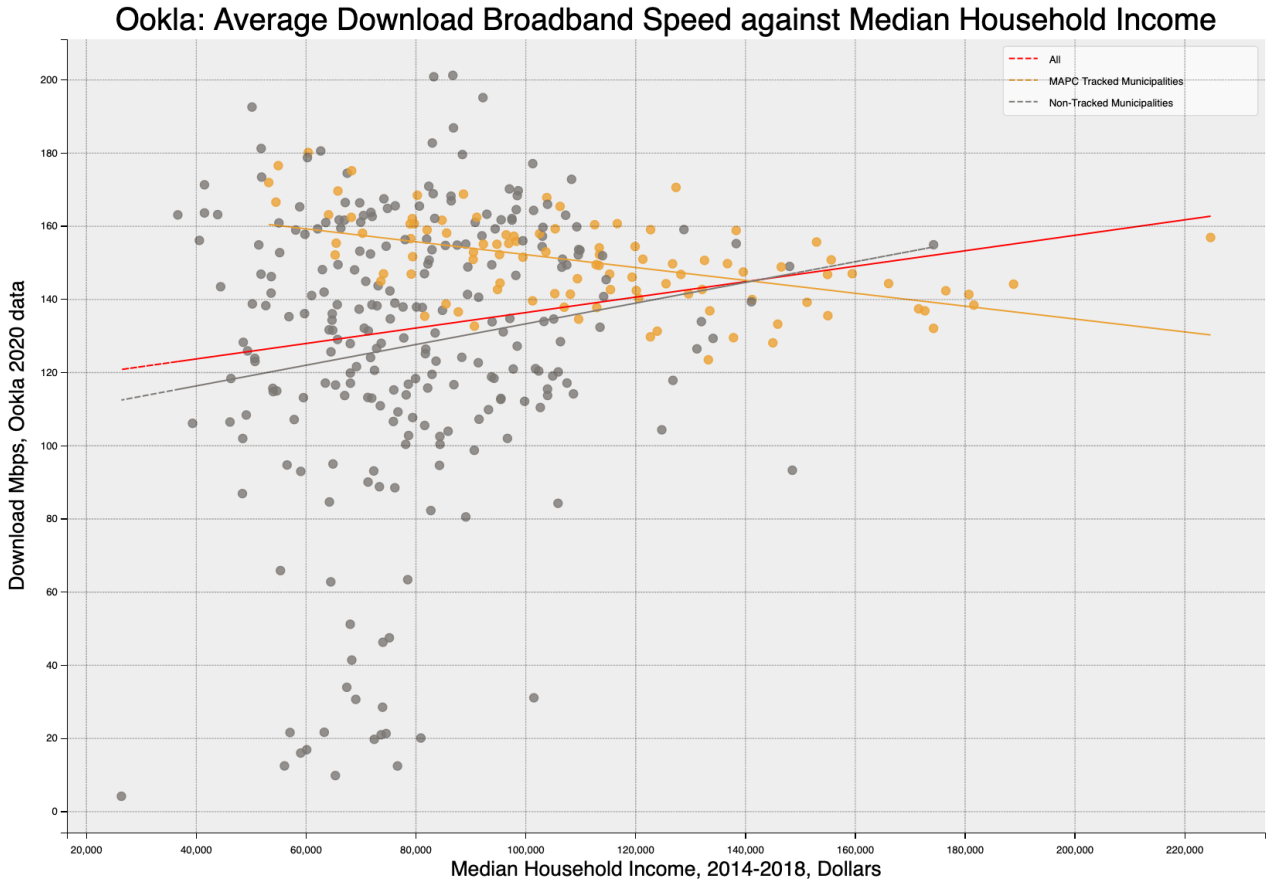
Here, we present a scatter plot of the MLAB data averaged per provider against median household income from the 2014-2018 census. Note that we used data from this census, because the 2020 census data is not yet available. In the chart, each data point is a municipality in MA; the data points colored in orange are municipalities within the MAPC's purview, and the grey data points are other cities.

There are a number of key points here:

1. The vast majority of municipalities are beneath the 100 Mbps target broadband speed, many significantly so.
2. There does exist a slight correlation between median household income and throughput Mbps
3. There numerous outliers; some low income municipalities with high throughput Mpbs, and many high income low throughput Mbbs municipalities.
4. It appears that there is a fairly dense clustering of municipalities towards the "bottom left" of the chart; interestingly, there doesn't seem to be much of a difference in terms of broadband speeds for municipalities with median incomes less than $120,000 per year.

We also generated an interactive HTML file of this data, with tooltips for labeling municipalities.

Ookla Data and Median Household Income



We also present a similar chart for the Ookla data. Here, there are a number of key points:

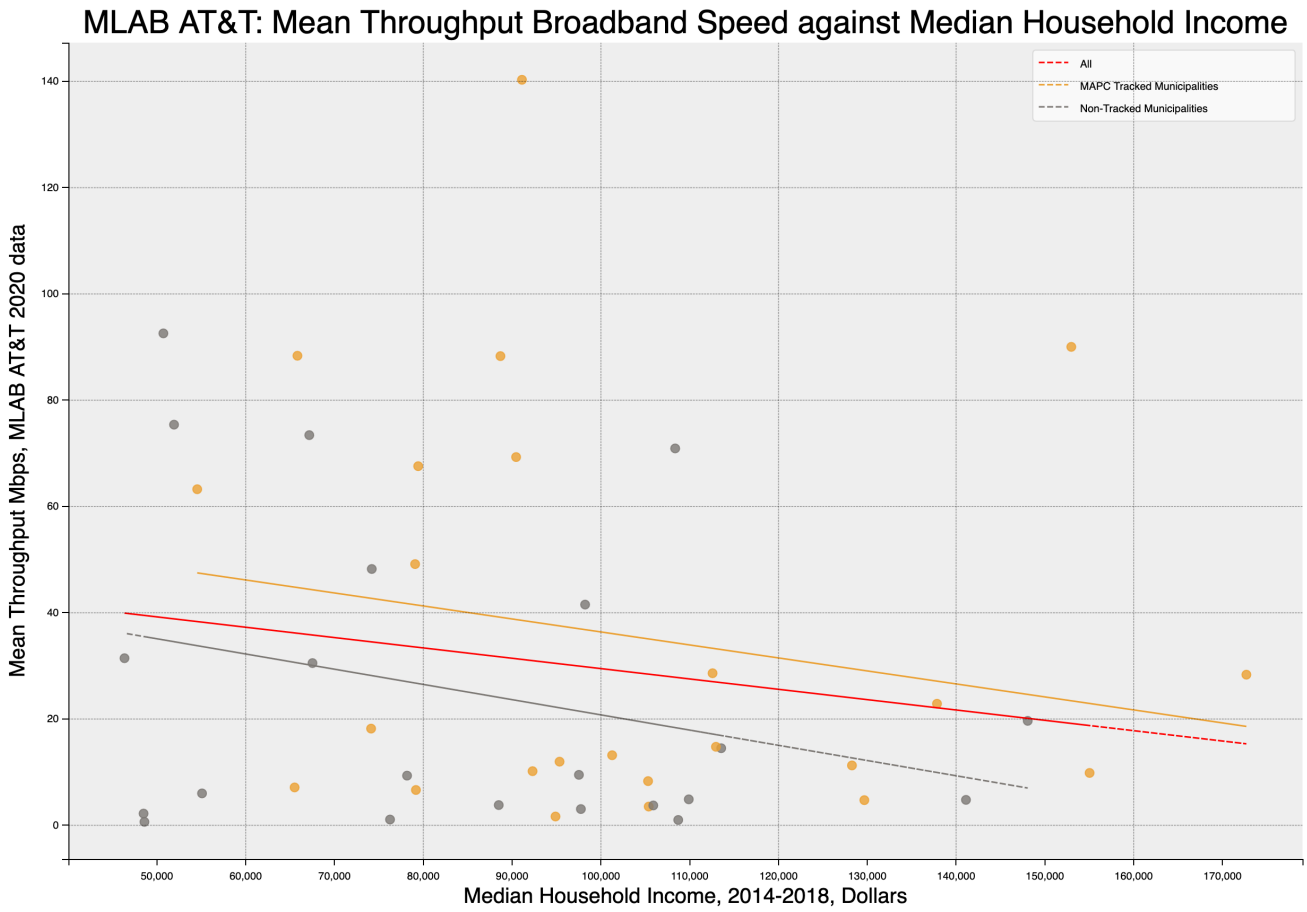
- The biggest difference is in the apparent measured speeds - measuring internet speed against a local ISP server as opposed to a larger network which includes autonomous systems yields much higher measurements.
- Among MAPC municipalities, there actually appears to be a downward trend as compared to median household income, while the opposite relationship exists for other municipalities.

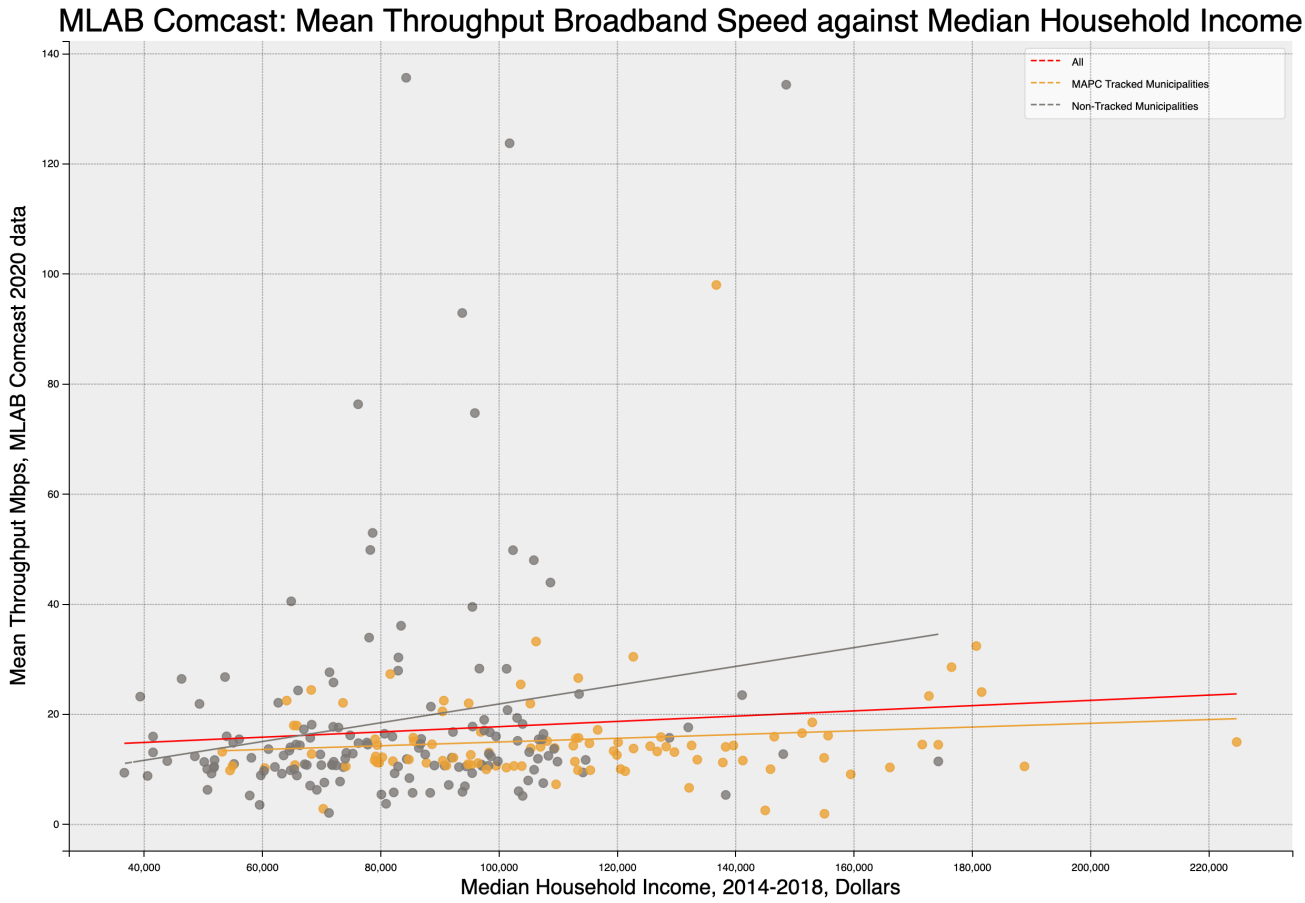
As with MLAB, we also generated an interactive HTML file of this data, with tooltips for labeling municipalities.

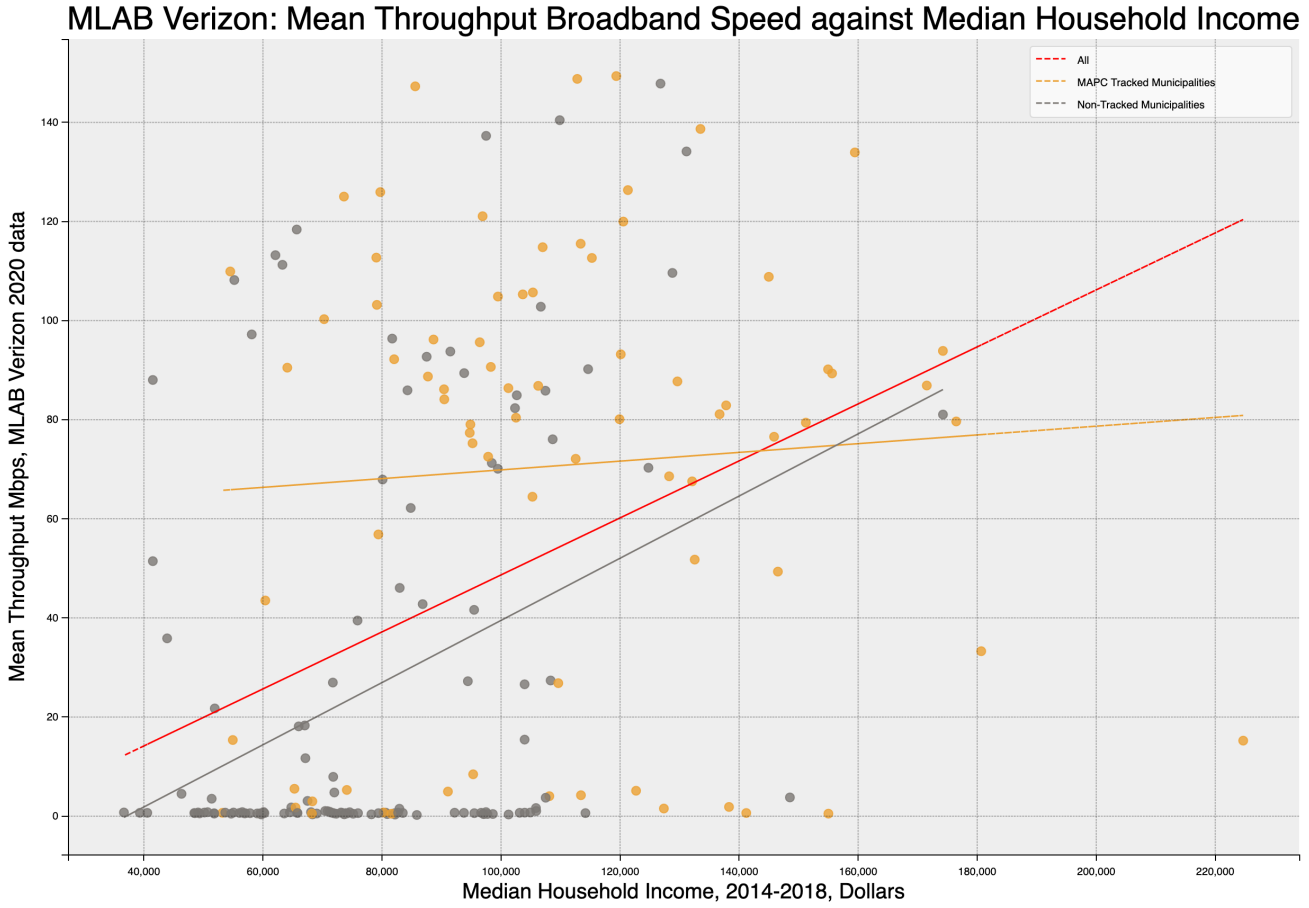
Provider Mean Throughput Mbps and Median Household Income

We also generated similar HTML files of scatter plots, but on a per Provider level. Here, we filtered the providers such that we only show data for Providers with at least 50 measurements in the dataset. Then, for each Provider, we only show municipalities where they provide some sort of broadband service.

There are 41 such providers that meet this threshold, and we generated individual scatter plots for each of them. We provide them as a deliverable to the MAPC, in their own directory for ease of access. Here, we show three of these -- AT&T, Caomcast, and Verizon -- as examples.







These three providers show some interesting trends, which in many ways are indicative of the complexity of the issues being studied here:

- There is a significant range of measured speeds; Verizon is the fastest of these three, AT&T is in the middle, and Comcast is the slowest.
- The trendlines representing the correlation between broadband speeds and median household income also show completely different trends: In Verizon's case, there does seem to an upward trend, while with Comcast there doesn't seem to be any trend (equally poor performance across all income ranges), and even a downward trend with AT&T!
- In Verizon and AT&T's case, there seem to be two clusters; one cluster at low speeds, and one at higher speeds. However, Comcast doesn't seem to have two separate clusters at all.

A complete analysis of each of the providers is out of the scope of this work. Nonetheless, we do present the associated files and the underlying data, which may be a starting point for future work. Regardless, the highlighted points above demonstrate the complexity and subtleties present when studying internet broadband.

Conclusion and Discussion

In this work, we have analyzed datasets of internet broadband speed from two sources -- Ookla and MLAB -- with regards to measured internet speeds in the 2020 calendar year in the state of Massachusetts, with a particular eye towards understanding broadband distribution on a per municipality basis. We cleaned and preprocessed the data, obtained a number of descriptive statistics, and plotted the data against median household income from the 2014-2018 census data. Further, we have provided csv files, scatter plots, HTML files, and python scripts for all the work that we have done to the MAPC.

Although this work is not a complete analysis of the data, we progressed significantly towards a better understanding of broadband availability in Massachusetts as it currently stands. In particular, there are six primary conclusions, which we present here.

Realistic Internet Measurements

Ookla and MLAB measure different things: Ookla measures the download and upload speed of a local ISP server. Effectively, it is a measure of whether or not the local ISP server is *capable* of achieving the speeds paid for by the ISP's customers in their local region. However, it is not necessarily a direct measurement of observed internet speeds.

MLAB, on the other hand, attempts to capture a more realistic measurement of internet speed - that is, how fast does a user actually observe their internet to be? In this case, the measurement includes a much larger and more comprehensive architecture than Ookla.

It is important to understand this difference, as the speeds observed by MLAB and Ookla are significantly different. A casual observer might conclude that their internet is plenty fast if they only consider speeds as measured by Ookla. This isn't to disparage Ookla - they provide an important service that allows users to get a reality check in terms of what their ISP is claiming. However, from a policy perspective, it's more important to consider realistic observed internet speeds, as this is a measure of what someone actually experiences.

A Note on MLAB Tests

As noted above, the dependent variable being measured by MLAB is not the same as that from Ookla. While we consider the tests from MLAB to be a more accurate representation of observed internet speed, this should also be taken with a grain of salt; our measurements are only for certain times of the day throughout 2020, and the providers measured are not ISPs but rather autonomous systems. It is entirely plausible that measurements may fluctuate throughout the day. In fact, we believe that there are a number of ISPs that attempt to regulate broadband speed based on timely demand, and therefore the MLAB measurements aggregated here may not be truly representative of that ISPs broadband capabilities.

In addition, the providers listed in MLAB are autonomous systems -- higher level networks that route internet traffic across the globe -- which means that some ISPs are not represented one to one. Notably, the organizations that run autonomous systems are not necessarily the same as the organizations that directly provide internet service for users. Thus, the implication is that some ISPs are limited to the capabilities of the autonomous systems they contract to for larger area internet routing. To illustrate this idea, consider an example of a user who lives in Boston downloading an article from the LA times: likely the article is on a server somewhere on the US west coast, while the user's local ISP server is located on the US east coast. The user's local ISP is likely only responsible for the initial traffic; their request will be handed off to some autonomous

system, which will handle ~90% of the route the network request takes. In this instance, the ISP cannot outperform the autonomous system, no matter how much money a user pays them for advertised speeds.

Regardless, the point of this note is simply to state that understanding broadband speeds is perhaps a more complex or subtle issue than it may seem at face value. Indeed, the best approach is probably to take both Ookla and MLAB data into account. That is, the local ISPs have invested in their servers such that 100 Mbps download speeds are attainable, but only some of the autonomous systems behind are also capable of delivering these speeds.

Ookla Download Speeds vs MLAB Mean Throughput Speeds

As noted above, Ookla and MLAB measure different things, and therefore, Ookla's average download speed is significantly higher than MLAB's Mean Throughput speed. In some sense, this is encouraging - the implication is that ISP's are in fact building servers capable of high speeds, and in fact, speeds well above the desired 100 Mbps download speed. Further, if we only consider the MLAB municipalities, it actually seems that on a local ISP server level, there is *more infrastructure investment* at lower to middle median household incomes than at high household incomes. [See "Ookla Data and Median Household Income" section]

However, from a different sense this is discouraging - despite the fact that local ISP servers are capable of handling higher speeds, the autonomous systems that back them are not. As these autonomous systems are responsible for the primary routing throughout the larger internet network, these speeds are far more realistic in terms of what a user will actually experience. We might consider this from a couple points of view:

- Customers are paying for speeds that they aren't actually truly getting
- The backbone networks are not good enough

However, it is not simply enough for a provider to claim that they can't do better because of these autonomous systems; quite the opposite, in fact, as there are a number of providers at the autonomous system level that do achieve fast internet speeds.

Failure to Reach 100/100 Mark

The 100/100 mark is an attempt to demarcate what is considered good internet speeds - 100 Mbps download speeds and 100 Mbps upload speeds. This analysis has shown that we do not achieve this definition in Massachusetts.

First, for upload speeds, we only have access to Ookla data, yet it is clear that only a few municipalities achieved 100 Mbps upload speeds. The vast majority fell significantly short.

For download speeds, we might conclude that the goal of 100 Mbps has been reached if we only consider the Ookla data. However, the MLAB data shows a completely different picture, and therefore it is not intellectually honest to conclude that this goal has been reached.

Disparity between Upload and Download Speeds

Even focused only on the Ookla data, there is significant disparity between the measured download and measured upload speeds. This concept is something that has been identified previously, and we are able to confirm that this phenomenon also exists in Massachusetts. Unfortunately, the upload speeds are nowhere close to what is necessary for the types of internet usage we have today. In particular, with much of our

business at work and coursework in education taking place via Zoom or other streaming services, it is critical that we have ample upload broadband.

Disparity amongst Individual Providers

In the MLAB data, there is a wide range of speeds from each of the different Providers. We consider this disparity to be important; it demonstrates that infrastructure investment can make a difference in terms of autonomous system performance. This is crucial, as there may otherwise be an argument to claim that only local server measurements are indicative of internet speed, but the apparent disparity in providers makes this argument a moot point.

Beyond that, however, the differences between providers make it difficult to understand broadband at a state level. As noted in the results section, the three presented providers (AT&T, Comcast, and Verizon) show some interesting trends, which in many ways are indicative of the complexity of the various issues present: There is a significant range of measured speeds, the trendlines representing the correlation between broadband speeds and median household income show completely different trends, and in Verizon and AT&T's case, there seem to be two clusters, while comcast doesn't seem to have two separate clusters at all.

Correlations Between Broadband Speed and Median Household Income

There are two important correlations to consider here:

- Ookla measurements of local servers against median household income
- MLAB MeanThroughput measurements of the larger internet network.

In the first case, among MAPC tracked municipalities, there is actually a downwards correlation - that is, as median household income increases, average download speed decreases. We take this as an encouraging sign, as the implication is that there is significant infrastructure investment at all income ranges, and perhaps even greater investment for lower income municipalities. However, among all other municipalities, there does exist an upward trend, where increased income implies an associated increase in download speeds.

In the second case, across all municipalities, there exists an upward correlation - that is, as median household income increases, mean throughput Mbps increases as well. As these measurements are likely a more indicative measure of observed speeds, we take this as a discouraging sign, that higher internet speeds are more readily available to people with higher incomes. Due to the nature of the world we live in today, with access to fast internet rapidly becoming a necessity rather than a luxury, we find that there is much room for growth here. At the very least, we expect that this data can help drive discussions on a policy level, to try and address this issue.

Final Note

As a final note, we want to say that it has been a pleasure working with MAPC and with BU, Spark! This project has been invaluable to us, and we hope that the work presented here can be the basis for effectual change and broadband accessibility progress.