Sanitizing a View of Consumer Broadband in the United States

Arun Dunna
UMass Amherst
Amherst, United States
adunna@cs.umass.edu

Zachary Bischof
IIJ Research Lab
Tokyo, Japan
z@chary.io

Romain Fontugne
IIJ Research Lab
Tokyo, Japan
romain@iij.ad.jp

***Abstract*—Since 2011, the FCC has been distributing "white-boxes" to broadband subscribers as part of their Measuring Broadband America initiative. These boxes conduct a number of network measurements that are made publicly available along with metadata on each participant (e.g., broadband provider, subscription speed, location). The FCC uses this data to publish annual reports on the state of broadband performance in the US, however, as with any study relying on crowd-sourced data, it faces difficulties in ensuring valid metadata for each vantage point. As a result, the FCC's annual reports only use a single month of measurements with validated data.**

In this paper, we present methods to accurately annotate the FCC's raw data, enabling additional types of analysis, such as longitudinal broadband studies spanning an entire year. Our methodology works by leveraging the results of the measurements themselves, as well as some additional datasets to differentiate between instances where the validated metadata can or can not be accurately applied to measurement results. We also discuss apparent issues in the data collection and sharing process that we observed in the FCC's publicly shared dataset. We make our scripts for cleaning the Measuring Broadband America data, as well as the newly annotated raw data publicly available. To illustrate the benefits of this annotated dataset, we also present a longitudinal analysis of the cost and availability of consumer broadband in the US.

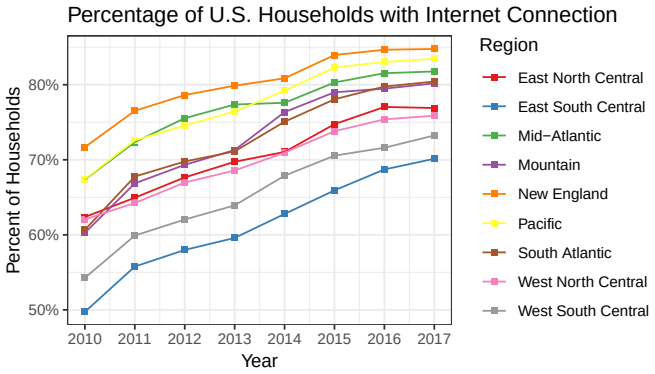

Fig. 1. Percentage of households that have a residential Internet connection, where at least one direction has an advertised bandwidth of at least 200 kbps. This is sourced from the FCC Form 477 data and the US Census 5-year American Community Survey data. This shows the growth of home broadband in the US. It is also broken down by US Census region to distinguish between rural and urban areas - showing that although there are still gaps between the regions, they have been closing over time.

I. Introduction

Over the last two decades, broadband Internet access has been of increasing importance to society, from connecting local communities to supporting national economies [1], [2]. Figure 1 shows the percentage of American households with an Internet connection with an advertised bandwidth of at least 200 kbps. It shows steady growth over the last decade across all regions in the US. While mobile broadband has seen even more rapid adoption globally over the last decade, fixed line broadband access is still a priority for many countries, and even considered to be a human right by the UN [3].

Due to the increased importance of broadband access, in 2010 the Federal Communications Commission (FCC) launched the Measuring Broadband America (MBA) project. With the MBA project, the FCC aims to improve our understanding of the performance of mobile and fixed-line Internet services throughout the US. As part of this initiative, the FCC worked with SamKnows to distribute instrumented home gateways (commonly referred to as "whiteboxes") to broadband subscribers around the country.

These whiteboxes perform regular network measurements, such as throughput, latency, and DNS queries[1] and transmit the results to an FCC server. The FCC then uses the collected data to publish annual reports on how different services and access technologies perform across various benchmarks, including throughput measurements, page loading times, and simulated video streaming.

However, such reports necessitate valid subscriber information (e.g., provider name, subscription speed, access technology, geographic location) for each participant. As with any study relying on crowd-sourced data, ensuring valid metadata for each vantage point is a laborious task. As a result, the FCC's annual reports only present a single month of measurements with validated metadata. In addition to the single month of validated data, the FCC also publishes raw unvalidated data for the entire year.

This work aims to accurately annotate the data in the FCC's raw dataset. Doing so allows us to look at trends in

[1]A full list of measurements can be found in the FCC's MBA Technical Appendix.

978-3-903176-27-0 ©2020 IFIP

broadband performance over a wider timespan and enables new longitudinal trends in consumer behavior.

Though a number of works have used data from the FCC/SamKnows deployment [4]–[15], they are sometimes limited in how much they can use the metadata. Some works [6], [10], [13] looks at performance statistics of the entire dataset as a whole, without using subscriber metadata, to compare against their own respective datasets. One work [5] specifically mentions having to take steps to validate a vantage point's service provider.

In this paper, we take the FCC dataset and study sources of error, present methods for validating metadata across the entire dataset, our process of using those methods to clean the dataset, and then evaluate the methods by comparing the resulting dataset to the original. Our methodology works by leveraging the results of the measurements themselves, as well as some additional datasets to differentiate between instances where the validated metadata can or can not be accurately applied to a measurement results. We also expand the dataset by making new inferences from combined raw and metadata.

We make our scripts for cleaning the Measuring Broadband America data, as well as the newly annotated raw data publicly available. To illustrate the benefits of this annotated dataset we also present a longitudinal analysis of the cost and availability of consumer broadband in US.

II. Dataset

In this work, the main dataset that we work with and perform the following processes on is the FCC's Measuring Broadband America dataset [16]. This dataset consists of both raw data and metadata, where raw data is measurement results collected from tests (such as traceroutes) run on gateways deployed in the home networks of consumers, and metadata about these units that is created and supplied by the FCC. Additionally, we use the FCC Form 477 data [17] in Section III for a cost and availability analysis. This dataset is separate from the MBA initiative and is comprised of responses to the FCC Form 477, a mandatory form that all telecommunications providers in the US have to complete bi-annually, and consists primarily of the maximum service speeds that each ISP offers to consumers within each geographic census block. In this same analysis, we also use the FCC's Urban Rate Survey data [18], which is an annual survey conducted on consumers to get information about their subscribed plans and the pricing of offered service plans. We release the processed data from all of these datasets at https://adunna.me/research/broadband-tma/.

In the remainder of this section, we describe the cleaning and validation processes that we apply to the primary dataset and to some portions of the supplementary datasets used in Section III. After our comprehensive cleaning and validation process, we end with reliable data from 2015-01-01 to 2018-12-01. This data is derived from 472,304,396 unique, raw tests during that timeframe. We summarize this data in Table I, though we note two limitations that impact the range of our final dataset: (1) we are unable to work with metadata from 2018, as the FCC has not yet published that data, and (2) we have metadata from 2013 onwards, but only have partial raw data provided for 2013 and 2014. As a result, we limit our data range to 2015 to 2018, and derive necessary metadata where possible, which is further elaborated in Section II-B.

A. Cleaning Process

While working with this dataset, we noticed several sources of errors and multiple inconsistencies across the data; this results in us formally categorizing these sources, and subsequently developing methods of correcting these errors, which can be applied to other datasets in the future. We call this generic process "cleaning", and in Section II-B, we follow this cleaning process with a more application-specific "validation" process. We note that with both our cleaning and validation processes, we build our processing code with generic classes that allow for easy expansion to include a new chunk of data, new type of data point, etc.

Incomplete, Invalid, and Redundant Data. With large datasets, incomplete, invalid, and redundant data is commonplace. Whether this stems from unreliable data collection resulting in missing periods in time-series data, or improper data handling resulting in invalid types for data fields, the inconsistencies need to be addressed before analyzing the data. To handle each of these issues, we either remove the relevant data points, or replace the invalid/incomplete data with an indicator that it should not be used in analysis. This replacement process applied to the metadata is summarized in Table II, where we include the percentage of units that include the given variable over time.

Field Mapping. We see multiple instances of fields not matching in type, label, and index across multiple chunks of data. This occurs in both the metadata where chunks are given in years, and the raw data in which chunks are given in months. To circumvent this issue, we apply *field mapping*. The general process relies on first identifying the data format of each chunk of data, then on merging all of these formats together. We then distill this collection of formats into the specific end formats we intend to work with in the dataset. By applying this process programatically, a scalable pipeline is developed to take in inconsistent chunks of data, and return them as chunks that follow a new, unified format. Those working with the data are then easily able to vertically merge these chunks of data into the desired sizes based on time periods (such as with longitudinal data), data size (such as when storage space is limited), etc.

Value Mapping. Often times, crowdsourced datasets contain inconsistencies or errors incurred through improper data entry by the public. One method of mitigating these errors is through intensive server-side validation when receiving the crowdsourced data, but this is not always done, whether due to infeasibility or out of ignorance. In working with this dataset,

	# Total Units			# New Units			# Tests					
Year	**Both M/R**	**M**	**R**	**Both M/R**	**M**	**R**	**Total**	**DL**	**UL**	**DNS**	**Traceroute**	**Usage**
2015	6240	1	1477	6240	1	1477	**141,778,142**	31,440,029	31,051,269	43,113,052	13,435,137	22,738,655
2016	4545	0	2705	883	0	521	**110,966,300**	22,007,271	21,768,464	35,360,672	16,845,315	14,984,578
2017	4378	0	2914	1190	0	483	**122,517,411**	22,086,750	21,885,804	40,590,027	23,837,527	14,117,303
2018	—	—	6917	—	—	1532	**97,042,543**	15,697,935	15,382,791	45,455,831	14,706,621	5,799,365

TABLE I

SUMMARY OF NUMBER OF UNITS PER YEAR, NUMBER OF NEW UNITS PER YEAR (UNITS NOT SEEN IN ANY PREVIOUS YEAR), AND RAW TESTS FOR UNITS. WE BREAK DOWN THE NUMBER OF UNITS INTO CATEGORIES *M* AND *R*, *M* REPRESENTS UNITS THAT HAVE METADATA BUT NOT RAW DATA, AND *R* REPRESENTS UNITS THAT HAVE RAW DATA BUT NOT METADATA; BOTH *M/R* REPRESENTS UNITS THAT HAVE BOTH METADATA AND RAW DATA. WE NOTE THAT 2018 METADATA IS NOT INCLUDED AS IT HAS NOT YET BEEN PUBLISHED.

Variable	**2013**	**2014**	**2015**	**2016**	**2017**
Advertised Up	95%	71%	77%	100%	100%
Advertised Down	95%	71%	77%	100%	100%
ISP	100%	80%	100%	100%	100%
Technology	99%	71%	87%	100%	100%
Longitude/Latitude	88%	73%	98%	0%	99%
Population	88%	73%	98%	0%	99%
FIPS Code	88%	73%	98%	0%	99%

TABLE II

THE PERCENTAGE OF UNITS, FOR A GIVEN VARIABLE IN A GIVEN YEAR, THAT HAVE DATA FOR THAT VARIABLE PRESENT. NOTABLY, THE FCC DOES NOT HAVE 2016 LOCATION DATA, RESULTING IN NO LOCATION DATA FOR THAT YEAR.

Reported ISP	**Corrected ISP**	**Cause**	**# Occ.**
Optimum	Altice	Subsidiary	412
Oceanic TWC	Time Warner Cable	Subsidiary	31
Clearwire	Sprint	Acquisition	71
Insight	Time Warner Cable	Acquisition	204
Qwest	CenturyLink	Acquisition	895

TABLE III

NUMBER OF INSTANCES OF AN ISP BEING INCORRECTLY REPORTED IN THE METADATA DUE TO AN ACQUISITION OR BEING A SUBSIDIARY.

as well as with the FCC's Form 477 data, we often see these entry errors. As a result, we apply the *value mapping* process in a one-pass scan over the data chunks following field mapping. This process involves first acquiring the unique instances of entered data, then generalizing those instances into classes, which will each roughly follow the types of mistakes users make when entering data (e.g. an abbreviation of an ISP's name is provided instead of the full name). Afterwards, a mapping is made from each class to the single desired end format, and then applied over the data chunks.

While a simple error where value mapping might be applicable is with a typo in user-provided data, a more complex error that we encounter in cleaning this dataset is when the provided Internet Service Provider (ISP) does not account for a merger or acquisition. In this instance, we can apply value mapping, and use the context of another field (such as the timestamp of the data point) to determine which value to map to. We show these corrections to ISPs in Table III based off of the acquisitions in Table IV and subsidiaries.

We also note that not all user-caused errors are corrected

Date	**Acquired ISP**	**Acquiring ISP**
2006-12-29	Bellsouth	AT&T
2010-06-13	Bresnan	Cablevision
2011-04-01	Qwest	CenturyLink
2012-02-29	Insight	Time Warner Cable
2013-06-01	Bresnan	Charter
2013-06-10	Clearwire	Sprint
2014-10-24	Southern NE Tel. Co.	Frontier
2015-12-21	Suddenlink	Altice
2016-04-25	Brighthouse	Charter
2016-05-18	Time Warner Cable	Charter
2016-06-21	Cablevision	Altice
2018-06-02	Hawaiian Telcom	Cincinnati Bell

TABLE IV

RELEVANT ACQUISITIONS INCLUDING THE ACQUIRED ISP AND ACQUIRING ISP, ALONG WITH DATE THAT THE ACQUISITION FINISHED.

via value mapping, as there are many instances where users submit data in the correct format, but the data itself is incorrect; for this reason, we apply the more application-specific validation process in Section II-B.

Identifier Normalization. Generally when crowdsourcing data, the order in which data points are received is not predetermined - multiple data points can be received near the same time, or even out of expected order, such as when the order in which two probes start tests is different from the order in which the data points for the two tests are received. Data points are also uniquely identified, often by numbers, but assigning these identifiers in crowdsourced data can be a hassle when determining the time of the test versus the time that the data from the test is received by the centralized server. As a result, we use the process of *identifier normalization*, where we assign a unique index to each data point based on characteristics that are important in determining the end ordering of the data points when they are analyzed.

In our case, we need to track individual units longitudinally; therefore, the primary characteristic we use is the timestamp at which the first test is performed client-side for a given unit. In this way, we can assign indexes to each unit, and use those indexes rather than the default unit identifier supplied in the dataset, therefore normalizing the given identifier. We refer to this normalized identifier in the rest of this work as the unit's *index*, and the need for it is further demonstrated in Figure 2,

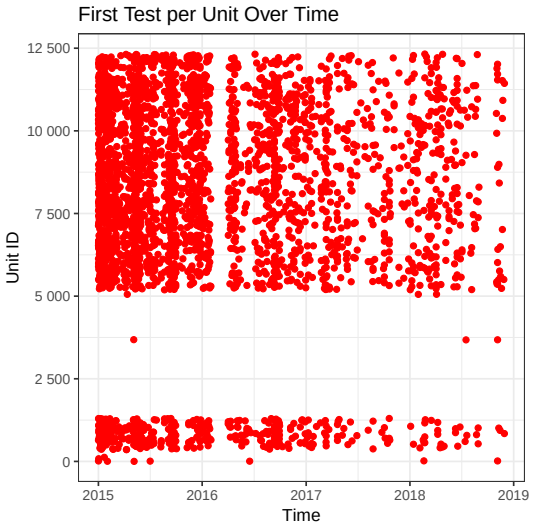

Fig. 2. Overview of time that first test was seen per unit, for any type of test, which shows the deployment of units over time and the spread of unit IDs. We can observe the clear delineation of unit deployments in groups, as well as the sparsity of unit deployment as time progresses. We also see two distinct blocks of Unit IDs, clustered at 0 to 1250 and 5000 to 12500; for this reason, we apply identifier normalization to transform this ID into an index.

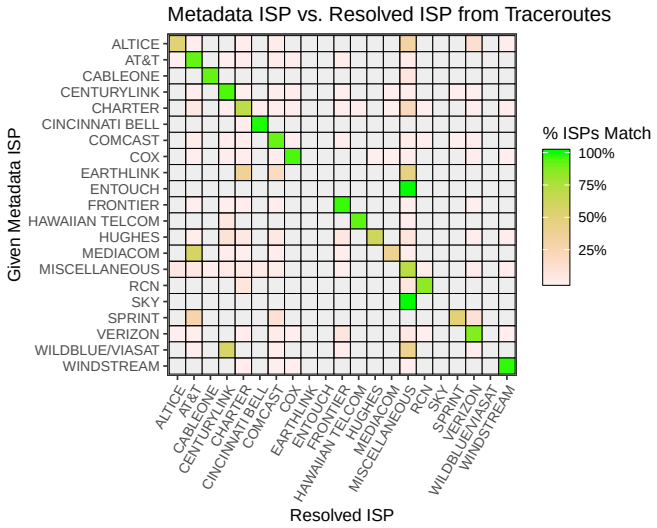

Fig. 3. Visualization of given metadata ISP compared to the ISP resolved from traceroutes. We do this by: (1) resolving the AS of the first public IP seen in traceroutes for each unit, (2) taking the closest resolved AS timestamp compared to the unit's reported metadata timestamp, (3) manually creating a map of ASes to ISPs based off of subsidiaries, acquisitions, organizations, etc. to get resolved ISP. Note that for the most part, ISPs tend to have a fairly high match rate, with some notable exceptions being Wildblue/Viasat, Altice, Earthlink, Sprint, and Mediacom.

where we see clusters of assigned unit IDs.

B. Metadata Validation

After our more generalizable cleaning process, we have a workable dataset; however, we go further by analyzing the sources of data and the available data points, and build several validation methods that can be used to check the correctness of existing data points. We next describe these validation methods, and the differences between the pre-validated data and validated data.

ISP Validation After the data cleaning process, we see a significant amount of remapping due to subsidiaries and acquisitions, shown in Table III. As a result, we also perform validation on the provided ISP in the metadata for each unit. To do so, we extract the first public IP address seen in each unit's traceroute data for each timestamp, which corresponds to the traffic transitioning into the ISP's public network. We then take these addresses and map them into ISPs, then take the ISP associated with the nearest timestamp to the metadata's timestamp. The mapping process is done by resolving the Autonomous System (AS) of the IP address using historical BGP data [19], then mapping the AS to the ISP using a manually created map based off of subsidiaries, acquisitions, and organizations. This allows us to compare the reported ISP with the observed ISP from the traceroute data.

We summarize this comparison in Figure 3, and note that in most cases, the resolved and reported ISPs have a high match rate of near 100%. Additionally, we overall have a 95.82% match rate between the validated ISP and the reported ISP. However, we note that a few ISPs, namely Viasat, Altice, Earthlink, Sprint, and Mediacom, have lower match rates than their peers, most likely due to using infrastructure of other providers or subsidiaries/parent companies, or using satellite infrastructure. Due to these disparities, we include both the validated ISP and the originally reported ISP in the distributed dataset, as the researcher can select the optimal one to use based on the application of the data.

Unit Location While IP geolocation is inherently imprecise, we are able to use it to validate the location of units to an extent. We use the geolocation to validate the US State that a unit is located in (i.e. to determine errors in location entry, such as failing to update the location after moving between states), and to infer the state of units that have no location data reported. We perform this geolocation on the first seen public IP address for each traceroute, obtained in the same method as in our ISP validation process, as the public IP addresses of units are not distributed for privacy reasons. While this has its own limitations, such as not being useful for satellite-based services, the geolocation data can be used to validate the overall integrity of the provided location data. However, we note that this method has the advantage of being less prone to changes in location caused by reassigning the IP address to a different customer in a different geographic area, as the first public IP is expected to be ISP infrastructure, which will most often have static and more stable addressing rather than the dynamic addresses assigned to customers.

Due to the accuracy limitations of geolocation, we use two sources to validate location information: MaxMind [20] and RIPE IPmap [21]. Since RIPE IPmap is more accurate but has less coverage, we first perform geolocation to the US State-

ISP	2013	2014	2015	2016	2017
Hughes	—	99 / 105	88 / 98	79 / 90	114 / 128
Mediacom	—	—	—	0 / 1	—
Viasat	65 / 71	61 / 67	46 / 52	36 / 43	31 / 38

TABLE V
RESULTS OF SATELLITE TECHNOLOGY VERIFICATION, STRATIFIED BY ISP. WE RESTRICT THE UNITS TO THOSE ONLY LISTED AS HAVING SATELLITE-PROVIDED SERVICES IN THE METADATA FOR THE GIVEN YEAR. THE NUMBERS REPORTED ARE THE NUMBER OF UNITS THAT COULD BE USING SATELLITE SERVICE BASED OFF OF THE TRACEROUTE DATA, OUT OF THE TOTAL NUMBER OF UNITS.

level using the RIPE IPmap geolocation API, and if the API does not provide a valid result, then we fall back onto the MaxMind GeoLite2 database.

The results of this process in comparison to the reported location data are shown in Figure 4. We find that overall, with the exception of satellite technology, the resolved locations align with the reported locations at a fairly high rate. There is more variance based on the ISP, which while we cannot determine for sure due to not having 100% ground truth for the unit location data, most likely is derived from a combination of shared/leased infrastructure, less localized infrastructure, and/or (to a lesser extent) imprecise geolocation. To alleviate some of this variance, we also consider border states, and show those results in Figure 4 as well.

Satellite Validation With 693 instances of satellite as the reported technology in the unit metadata, and given that the ISP validation is less accurate for satellite services, we find it important to validate the reported technology. We can find false positives of satellite as the reported technology by checking the minimum necessary RTT (based off of latency to geostationary satellites) of the reported RTT for the final hop in the unit's traceroutes. We can then extract this RTT for the closest timestamp to the reported metadata, and compare to determine whether the latency is lower than would be required for satellite-based services. To increase the reliability of this comparison, we take the median of the all the extracted RTTs from traceroutes performed in the same month as the closest timestamp to the reported metadata timestamp, then use this median to validate the technology against the minimum required latency.

The results of this validation are shown in Table V, and we stratify the results based on the three ISPs that are associated with the satellite technology in the dataset: Hughes, Viasat, and Mediacom. We find that the one unit reported as having satellite service provided by Mediacom is likely an error, which is intuitive as Mediacom provides cable services and not satellite services. We also find that the majority of units reported as using satellite-based services have the correct listed technology, but that the number of incorrect reports increases slightly over time, likely due to the field not being updated when the unit switches to a non-satellite-based service.

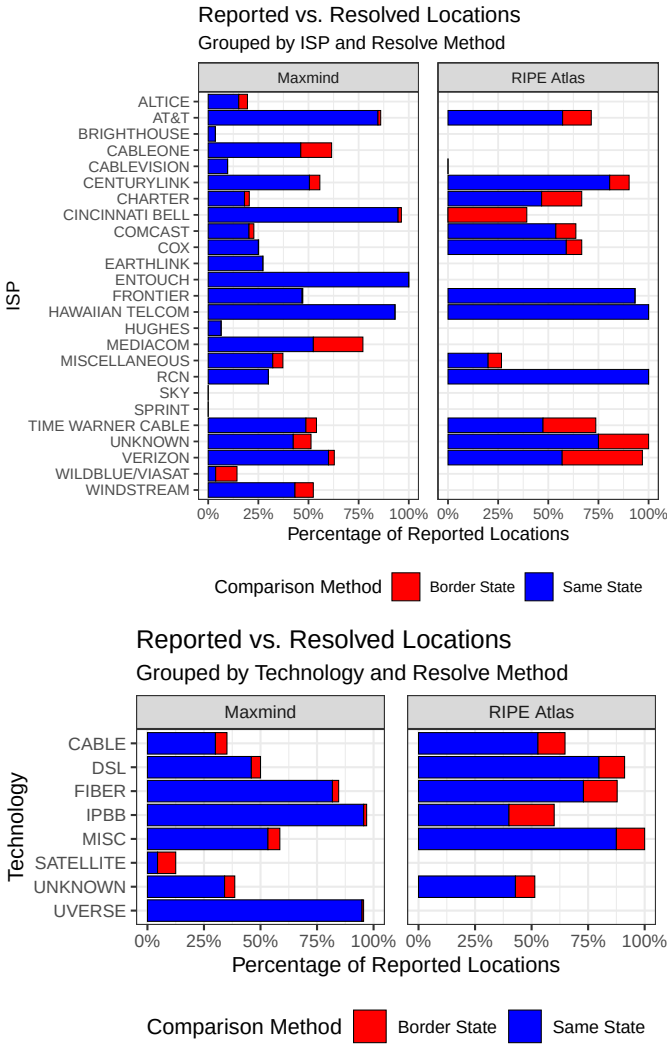

Fig. 4. Reported vs. resolved geographic locations, grouped by resolve method and technology/ISP. To do this, we: (1) for each traceroute take the first public IP, (2) geolocate the US State using that IP using the RIPE method, (3) if the RIPE method does not work, then we use the MaxMind GeoIP mapping as a fallback, (4) take the closest resolved location timestamp to the unit's reported metadata timestamp.

C. Expansion Through Inferences

While many new data points can be inferred from this existing dataset due to the copious amounts of raw test data, we derive several new datasets that that are useful for both summary statistics and for answering questions about the availability of broadband, which ties into our later analysis in this work. We next describe these derivations and summarize the results of each.

Population Data When using this base dataset in social or political applications, population data is often an important metric to have. We therefore combined the provided location data with US Census [22] and American Community Survey (ACS) data [23] to expand the metadata to include population information at the finest granularity available. While we

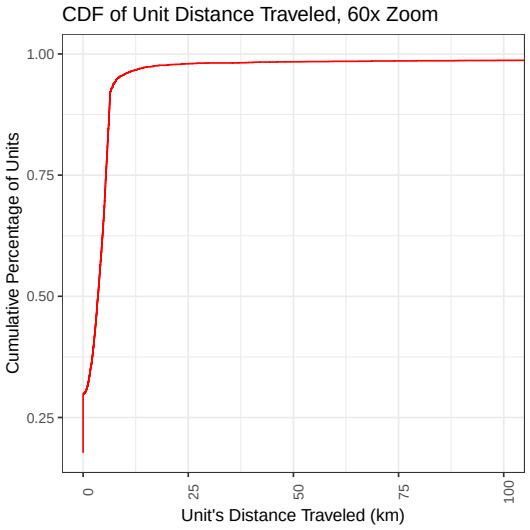

Fig. 5. CDF of the cumulative distance traveled across the lifetime of all units. We notice that almost all units do not travel much at all, with about 95% of units traveling 12.5 km or less over their lifetime. We also see a distinct knee, marking the point at which units actually "move" in contrast to inaccurate location reporting, at 6.4 km. Additionally, about 30% of units do not move at all, meaning that the same exact location data was given for all years that they were in the data; however, about 65% of units move between 0 and 12.5 km, either due to actual small moves or due to inaccurate or inconsistent location reporting (e.g. a given unit reports their census block group in 2014, but only their census tract in 2015, even though they have not relocated).

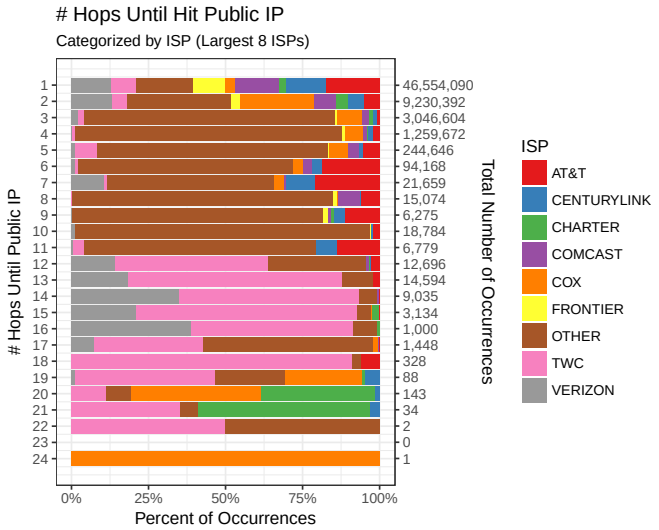

Fig. 6. The number and percentage of hops, for each traceroute, until a public IP was hit - categorized by ISP.

do not directly use this expanded dataset in this work, we include it in our published datasets for use by others and plan to explore the differences between urban and rural broadband availability in future work using this population data.

Distance Traveled One important aspect of consumer broadband choice is availability based on the locations served by different ISPs. As a result, we use the location data of units to track the movement of units over time. This allows us to derive the *cumulative distance traveled* across a given unit's lifetime, which we can use in analyzing the impacts of relocation on broadband availability, as well as in characterizing the reasons for why consumers change service providers. A summary of this metric across the unit metadata is shown in Figure 5. We can also implement a cutoff for defining whether a unit has relocated or not, which is inherently not exact due to disparities in the reported location data (e.g. a given unit reported to be 1 km away in a later year, even though the unit has not moved, due to imprecise location data), by analyzing the knee of the curve in Figure 5.

Private Infrastructure Because we have the raw traceroute data, we can infer some information about the private infrastructure, either owned by the customer or in the ISP's internal infrastructure, before traffic reaches the ISP's public infrastructure. One piece of information we can extract is the number of hops traveled until a public IP is observed, which can tell us both about the depth of the ISP's internal infrastructure before a public IP is seen, as well as about any technology such as NATs in the consumer's home network. We show a breakdown of this information in Figure 6, which is stratified by ISP for the 8 largest ISPs, with the other ISPs in their own bucket. One observation we can make is about the depth of TWC's internal infrastructure, which appears to have many more instances of larger amounts of internal routing compared to other ISPs in the dataset.

D. A Comprehensive Dataset

After our cleaning, inferencing, and validation processes, we end with a comprehensive and unified dataset. We distribute this dataset across five distinct categories: Raw, Metadata, Miscellaneous, Traceroute, and Usage. Raw pertains to raw test data, and contains reorganized raw test data as well as longitudinal summaries of all tests conducted. Metadata pertains to unit metadata, and contains the processed metadata, as well as merged US Census population data and validated GeoIP data. Miscellaneous pertains to data gathered from sources other than the primary Measuring Broadband America source, and contains the processed FCC Form 477 data as well as data about US broadband usage from the American Community Survey. Traceroute pertains to data extracted from the traceroute data, and contains first public hop data, validated unit technology data (focusing on satellite-based service), rolling statistics about the traceroutes, and supplementary AS data. Lastly, Usage pertains to data derived from the usage tests, and contains rolling statistics about usage data, such as daily bandwidth usage by units for both test-related and non-test-related network traffic.

We highlight that this dataset is available at https://adunna.me/research/broadband-tma/, and encourage the use of it by not just researchers in the measurement community, but also in other disciplines, as the dataset crosses into applications

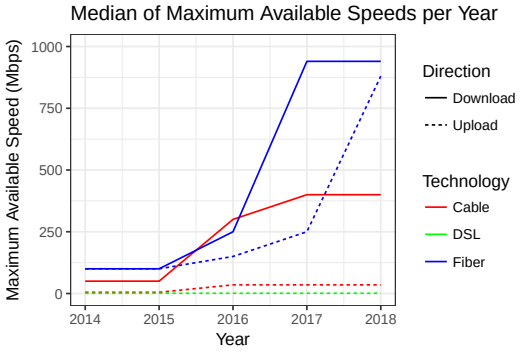

Fig. 7. The median of the maximum available speeds for all service plans in the FCC Form 477 data, delineated by year, technology, and direction of service (inbound/outbound traffic).

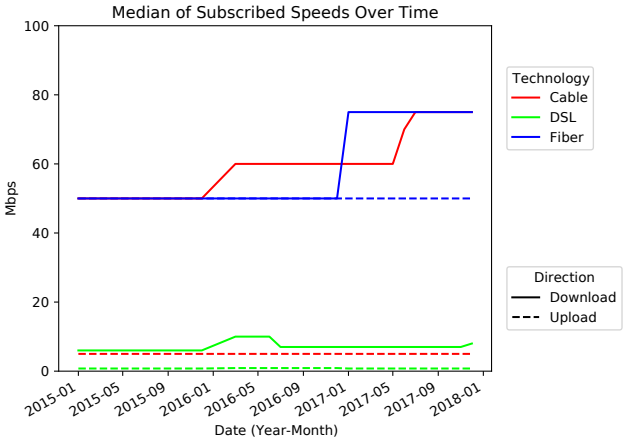

Fig. 8. The median of subscribed speeds in Mbps across all units in the original, uncleaned broadband measurement data, delineated by month/year, technology, and direction of service (inbound/outbound traffic).

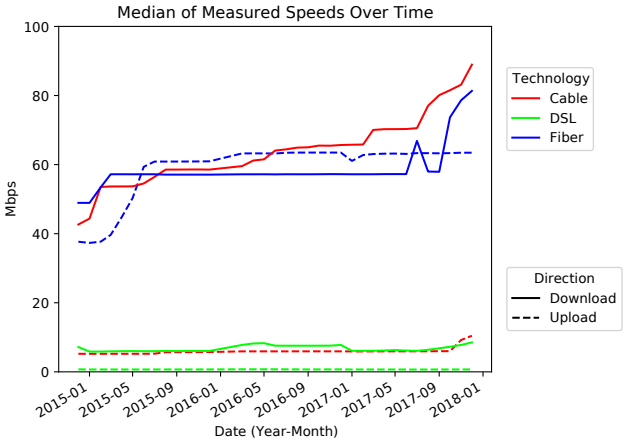

Fig. 9. The median of measured speeds in Mbps across all units in the cleaned and validated broadband measurement data, delineated by month/year, technology, and direction of service (inbound/outbound traffic).

in economics, politics, sociology, and other fields. We also provide extensive documentation along with the dataset, and the original source datasets in addition to the source code used in the transformation process of those original datasets, so that researchers can apply the same modular methods in improving existing datasets.

III. Cost and Availability Analysis

To demonstrate some of the potential applications of this dataset, we cross-reference it with FCC Form 477 data [17] and FCC Urban Rate Survey (URS) data [18] to perform a longitudinal analysis of the cost and availability of consumer broadband in the United States. Specifically, we look at how this cost and availability has changed from 2015 to 2018, as well as compare the availability to what consumers actually receive in our dataset.

Availability of Consumer Broadband We first look at the availability of broadband over time to form a basis for our cost analysis. As a first step, we process the Form 477 data to determine the available service plan speeds over time. We show this in Figure 7, and can see that the availability of higher speed service has steadily increased in both cable and fiber-based plans.

We then perform a comparison against both the unchanged broadband measurement data, shown in Figure 8, and the validated broadband measurement data, shown in Figure 9. We do this comparison both to highlight the shared trend in the increasing availability of high-performance Internet service, but also to demonstrate the differences between analyses performed using the original dataset versus our cleaned and validated version. While we see the same underlying trend, many of the nuances are not revealed until we analyze at a finer granularity with our modified dataset. Additionally, we observe substantial differences in the download speeds for fiber-based service, indicating either incorrectly reported subscribed speeds in the dataset, or a large difference between the speeds that the service provider delivers and the speeds that the customer actually receives.

In looking at the available speeds versus the actual subscribed and measured speeds over time, we see that though the available speeds have increased dramatically in the past few years for both cable and fiber subscribers, the actual speeds people are subscribed to have not increased at nearly the same rate. This is due to the conscious choice by consumers to not switch to a higher capacity plan. We surmise that a large portion of this reasoning is due to cost, which we analyze next.

Cost of Consumer Broadband By looking at the FCC Urban Rate Survey data, we can analyze the monthly cost of services over time in the United States. This data is depicted in Figure 10, which shows that the costs of both DSL and cable services have steadily risen, whereas the cost of fiber services have declined. By itself, this could be due to an increase in service plan cost; however, we cross-reference this data with the service plan speeds to gain a notion of service *value* over

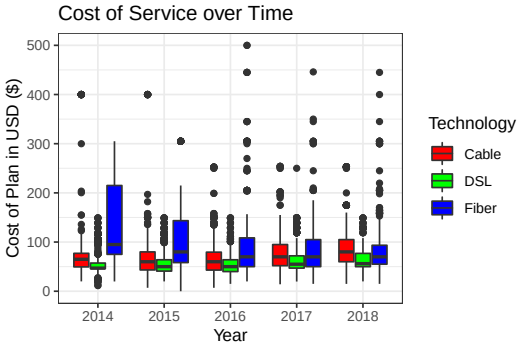

Fig. 10. The monthly cost of all service plans available in both the FCC Form 477 data and the FCC Urban Rate Survey data over time, categorized by technology.

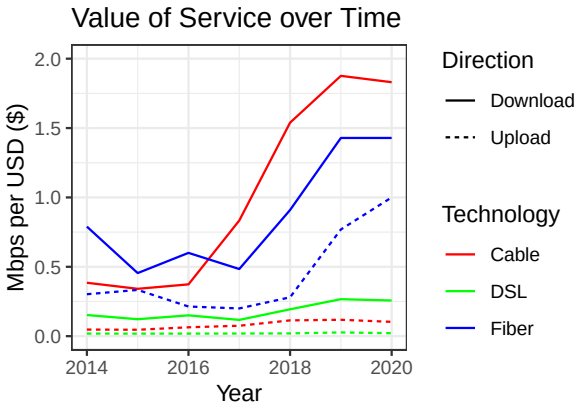

Fig. 11. The median value of all service plans available in both the FCC Form 477 data and the FCC Urban Rate Survey data over time in the format of Mbps per USD, stratified by technology and direction of service.

time. We show this in Figure 11, and see that for cable and fiber, as well as DSL to an extent, the value of service has increased significantly over time: about a 400% increase in value over the past five years. This suggests that though the availability and cost of services have both increased in recent years, the plans available to consumers offer significantly faster connections with higher value.

Figures 10 and 11 summarize the cost and value of broadband plans that consumers in the FCC's URS are subscribed to. However, beyond just validating this by using more sources, we also need to consider other important factors; notably, the *actual measured value* that consumers are getting from their selected plans.

To analyze this, we mapped units from the MBA dataset to service plans in both the FCC Form 477 data and the FCC URS data by matching units to the plan with the closest download and upload throughput rates. This provides an estimate of each participant's monthly subscription cost. However, we note that while this cost includes surcharges and other mandatory ISP charges, and does not include promotional pricing (a common tactic used in selling telecommunications services in the United States) but instead uses the regular plan pricing,

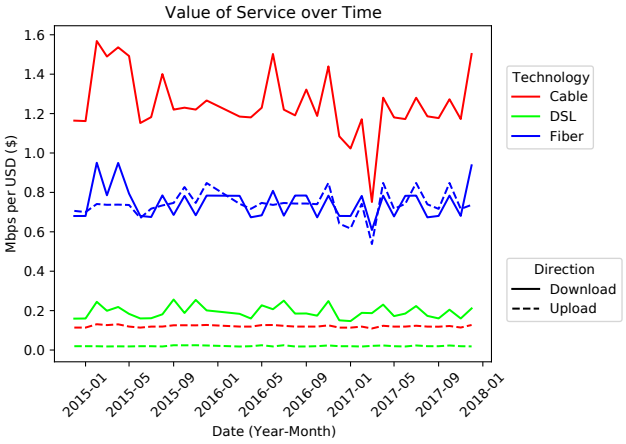

Fig. 12. The median value of the service plans of MBA participants over time in the format of Mbps per USD, stratified by technology and direction of service.

it does not include other factors such as local taxes, other local fees, or optional modem rental fees. Though this prevents us from being able to know the exact end amount paid by each individual subscriber, we do consider this to provide a reasonable estimate of the costs of broadband access.

We plot this results in Figure 12, which shows the median value of service for the MBA participants. In contrast to our findings in Figure 11, we find that the actual value of service for the median subscriber did not show any clear trends between 2015 and 2018, staying mostly constant across all technologies. This is a rather stark difference, and while we can speculate various reasons that could cause this such as certain economic factors (e.g. lack of competition), or ISPs under-delivering their services, more in-depth analysis needs to be performed in this area in the future to be certain.

IV. Conclusion

In this paper we present a set of methods to clean, annotate, and expand the FCC's Measuring Broadband America data. The resulting scripts and dataset are made publicly available, at https://adunna.me/research/broadband-tma/, so that the research community can apply these techniques to existing datasets and conduct new analysis based on unified and trustworthy datasets. We illustrate the benefits of the sanitized FCC dataset with a brief study on the cost and availability of broadband in the United States. We quantify the evolution of the value of broadband services from 2014 to 2018 and discuss its potential importance to consumers when selecting a service.

References

[1] W. Bank. IC4D 2009: Extending reach and increasing impact. [Online]. Available: http://go.worldbank.org/NATLOH7HV0

[2] ITU. The impact of broadband on the economy. [Online]. Available: https://www.itu.int/ITU-D/treg/publications/bbreports.html

[3] R. Lane. The united nations says broadband is basic human right. [Online]. Available: https://www.forbes.com/sites/randalllane/2011/11/15/the-united-nations-says-broadband-is-basic-human-right/

[4] Z. S. Bischof, F. E. Bustamante, and R. Stanojevic, "The utility argument – making a case for broadband SLAs," in *Proc. of PAM*, March 2017.

[5] Z. S. Bischof, F. E. Bustamante, and N. Feamster, "Characterizing and improving the reliability of broadband internet access," in *Proc. of TPRC*, 2018.

[6] Z. S. Bischof, F. E. Bustamante, and R. Stanojevic, "Need, want, can afford – broadband markets and the behavior of users," in *Proc. of IMC*, 2014.

[7] J. P. Rula, Z. S. Bischof, and F. E. Bustamante, "Second chance: Understanding diversity in broadband access network performance," in *In Proc. of C2B(I)D*, 2015.

[8] S. Sundaresan, W. de Donato, N. Feamster, R. Teixeira, S. Crawford, and A. Pescapè, "Broadband internet performance: a view from the gateway," in *Proc. of ACM SIGCOMM*, 2011.

[9] S. Sundaresan, N. Feamster, R. Teixeira, and N. Magharei, "Measuring and mitigating web performance bottlenecks in broadband access networks," in *Proc. of IMC*, October 2013.

[10] I. Canadi, P. Barford, and J. Sommers, "Revisiting broadband performance," in *Proc. of IMC*, 2012.

[11] N. Feamster and J. Livingood, "Measuring internet speed: Current challenges and future recommendations," 2019.

[12] S. Bauer, D. Clark, and W. Lehr, "Powerboost," in *Proc. of HomeNets*, 2011.

[13] T. V. Doan, V. Bajpai, and S. Crawford, "A longitudinal view of netflix: Content delivery over ipv6 and content cache deployments," 2020.

[14] G. Molnar and S. J. Savage, "Market structure and broadband internet quality," in *The Journal of Industrial Economics*, 2017.

[15] G. Dimopoulos, P. Barlet-Ros, C. Dovrolis, and I. Leontiadis, "Detecting network performance anomalies with contextual anomaly detection," in *IEEE International Workshop on Measurement and Networking*, 2017.

[16] FCC. FCC Measuring Broadband America. [Online]. Available: https://www.fcc.gov/general/measuring-broadband-america

[17] ——. FCC Form 477. [Online]. Available: https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477

[18] ——. FCC Urban Rate Survey. [Online]. Available: https://www.fcc.gov/economics-analytics/industry-analysis-division/urban-rate-survey-data-resources

[19] U. of Orgeon. University of Oregon Routeviews Project. [Online]. Available: http://archive.routeviews.org/

[20] MaxMind. Maxmind geolite2 ip geolocation database. [Online]. Available: https://dev.maxmind.com/geoip/geoip2/geolite2/

[21] R. NCC. Ripe ipmap: Geolocating internet infrastructure. [Online]. Available: https://ipmap.ripe.net/

[22] U. C. Bureau. Us census data. [Online]. Available: https://data.census.gov/cedsci/

[23] ——. Us american community survey. [Online]. Available: https://www.census.gov/programs-surveys/acs