# Characterizing Internet Access and Quality Inequities in California M-Lab Measurements

UDIT PAUL, UC Santa Barbara
JIAMO LIU, UC Santa Barbara
DAVID FARIAS-LLERENAS, UC Santa Barbara
VIVEK ADARSH, UC Santa Barbara
ARPIT GUPTA, UC Santa Barbara
ELIZABETH BELDING, UC Santa Barbara

It is well documented that, in the United States (U.S.), the availability of Internet access is related to several demographic attributes. Data collected through end user network diagnostic tools, such as the one provided by the Measurement Lab (M-Lab) Speed Test, allows the extension of prior work by exploring the relationship between the quality, as opposed to only the availability, of Internet access and demographic attributes of users of the platform. In this study, we use network measurements collected from the users of Speed Test by M-Lab and demographic data to characterize the relationship between the quality-of-service (QoS) metric download speed, and various critical demographic attributes, such as income, education level, and poverty. For brevity, we limit our focus to the state of California. For users of the M-Lab Speed Test, our study has the following key takeaways: (1) geographic type (urban/rural) and income level in an area have the most significant relationship to download speed; (2) average download speed in rural areas is 2.5 times lower than urban areas; (3) the COVID-19 pandemic had a varied impact on download speeds for different demographic attributes; and (4) the U.S. Federal Communication Commission's (FCC's) broadband speed data significantly over-represents the download speed for rural and low-income communities compared to what is recorded through Speed Test.

## 1 INTRODUCTION

The term "digital inequality" refers to the gap in Internet access that exists across different geographic areas and demographic variables [5]. Access to the Internet is known to impact multiple facets of human life, including economic [6], education [18], health [11], and, more recently, the ability to self-isolate to prevent spread of COVID-19 [17]. The majority of prior work on digital inequality across the U.S. has focused primarily on the availability of Internet access within a region. However, we argue that the *quality* of the Internet access is equally important. While the ability of an Internet connection to support advanced and bandwidth-intensive applications, such as video, has always been important, it has never been more so

Authors' addresses: Udit Paul, Department of Computer Science, UC Santa Barbara, u_paul@ucsb.edu; Jiamo Liu, Department of Computer Science, UC Santa Barbara, jiamoliu@ucsb.edu; David Farias-Llerenas, Department of Computer Science, UC Santa Barbara, dfariasllerenas@ucsb.edu; Vivek Adarsh, Department of Computer Science, UC Santa Barbara, vivek@ucsb.edu; Arpit Gupta, Department of Computer Science, UC Santa Barbara, arpitgupta@cs.ucsb.edu; Elizabeth Belding, Department of Computer Science, UC Santa Barbara, ebelding@ucsb.edu.

than in the post-COVID-19 world. The availability of quality Internet access now directly impacts remote learning outcomes, the ability to work at home, and the ability to use telehealth, among others [7, 8, 10, 20].

Internet access quality has received less attention than availability in part due to the dearth of reliable and granular data related to Internet quality [49]. The Federal Communications Commission (FCC), through Form 477, documents Internet coverage and maximum theoretical available download speed across the country. This documentation is done using information received from Internet service providers at the geographic granularity of the census block. The inaccuracy of this data in terms of overestimating coverage, especially in rural areas, is well documented [16, 51]. To improve access and quality of Internet, large financial investments have been made by the federal government [13], but given the underlying data used to guide these efforts is rife with errors, such investment runs the risk of being completely misdirected.

As an alternative to the FCC data source, within the past few years multiple for-profit and nonprofit programs such as Measurement Lab (M-Lab), SamKnows and Ookla have undertaken the complex task of analyzing Internet access and performance through crowdsourced measurements. For instance, Speed Test by M-Lab [35] collects Internet quality of service (QoS) metrics such as download speed, round-trip time (RTT) and packet loss rate when an user initiates a test. Google also collaborates with M-Lab and allows its users to conduct network diagnostic tests [25] using M-Lab provided infrastructure. With the aid of these measurements collected by M-Lab, it becomes feasible to dive into the problem space of determining the factors that affect the quality of Internet access across different demographics and geographical locations amongst different users who take the test. It is this topic that our work addresses.

In this paper, we combine crowdsourced measurements from M-Lab with recent demographic data from the Economic and Social Research Institute (ESRI) to characterize the effect of demographic attributes on the quality of Internet connectivity. We conduct multiple statistical and geographical aggregations of these datasets to overcome limitations imposed by crowdsourced measurements. We attempt to identify the relationship that exists between an important quality of service metric, download speed, and land and demographic factors such as type of area (rural/urban), income, education and population. In addition, as COVID-19 imposed lockdowns have significantly modified our online footprint, we explore how Internet quality changed across different demographic variables during this period. Finally, we use our analysis to highlight the amount of inaccuracy that exists in FCC data, particularly in rural and lower income areas in comparison to what is recorded through the Speed Test. We conduct this analysis for the state of California but our methodology can be extended to cover any geographical region. In summary, our paper reveals the following key factors that affect Internet quality through download speed collected from M-Lab Speed Test users:

(1) Income has the strongest correlation with download speed, followed by type of area.
(2) While rural areas record low download speeds compared to urban areas, performance gaps also exist between income groups within urban regions.

(3) The change in Internet usage patterns due to COVID-19 lockdowns coincided with a decrease in download speeds across the board, with previously high performing areas demonstrating the greatest decreases.
(4) The FCC Broadband Report highly overestimates download speed in rural and lower income group regions, more so than in urban and wealthier areas.

## 2 DESCRIPTION OF THE MEASUREMENT DATA

We begin our study by combining publicly available Internet QoS data from the Speed Test by M-Lab [37] with ESRI demographic data [27]. In the following section, we describe these datasets in more detail.

### 2.1 M-Lab Speed Test Data

M-Lab is an open-source project whose mission includes providing consumers and researchers with free information about Internet performance [36]. It has a distributed architecture with over 500 well-provisioned servers to conduct free performance measurement tests. Clients can use various tools, such as the Network Diagnostic Tool (NDT) and WeHe, to measure different aspects of Internet connectivity and quality.

Amongst their active measurement tests, we select data from NDT because it measures the performance of a TCP connection and provides summary data that includes our metric of interest: *download speed*. Measurement tests are conducted when a client initiates the measurement voluntarily, either from a web app or a browser. Once the test is initiated, the M-Lab server-selection algorithm chooses a server geographically closest to the client unless otherwise selected by the client or prohibited by factors such as network capacity and load condition of the server [32–34]. The test consists of bulk exchange of data between the client and server, as defined in IETF RFC 3148 [53]. During this single TCP connection test, a variety of information is recorded, including client and server IP addresses, download speed, upload speed, round trip time and packet loss rate. The collected data is publicly available for use [22].

To characterize the quality of Internet access for users of Speed Test by M-Lab in California, we analyze M-Lab NDT data collected in the state between 01-01-2020 and 04-30-2020. We focus our analysis on download speed—an important QoS metric. To geographically locate clients, we use a popular IP geolocation service, IPinfo [31], to obtain the location coordinates of recorded client IP addresses and information about the client's Internet service provider (ISP). Because performance in fixed networks (e.g., maximum download speed) varies from wireless networks, we separate measurement samples by access technology (fixed and wireless) to enable a fair comparison. We use the client's ISP information to separate these measurement types.

Table 1(a) displays the number of measurement samples and unique IP addresses present in our M-Lab dataset by type of access. The wireless measurements form only 3% of the measurement total. Geographic areas, as shown in table 1, can be represented by regions of varying sizes. For example, a census block is the smallest geographic area for which the Census Bureau collects and tabulates census related information [24]. On average, a group of 39 census blocks form a census block group [23] and contains between 600 and 3000

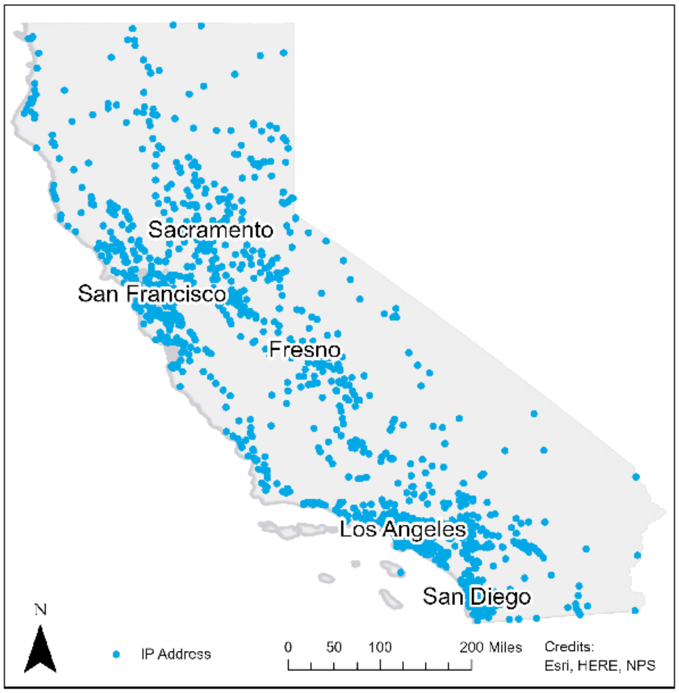


Fig. 1. Location of Unique IP addresses in the M-Lab California Data.

| **Technology** | **Total Measurements** | **Total Unique IP Addresses** |
|---|---|---|
| All | 8,666,013 | 1,133,282 |
| Fixed | 8,425,723 | 1,096,349 |
| Wireless | 240,290 | 36,933 |

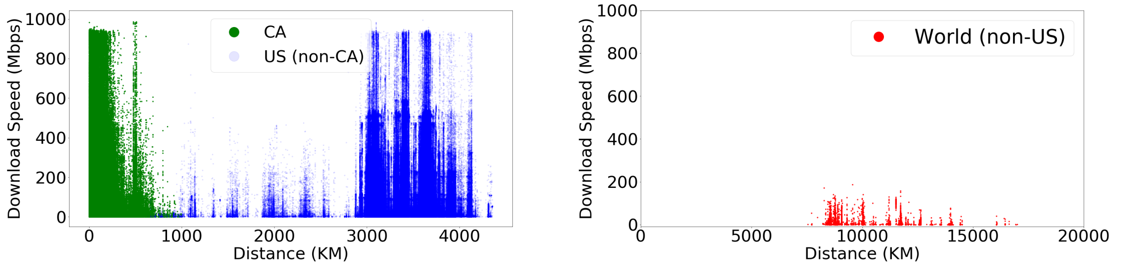
a. Access Technologies

| **Geographic Area** | **CA** | **M-Lab** | > 10 **IPs** |
|---|---|---|---|
| # of Blocks | 710,145 | 1,446 | 984 |
| # of Block Groups | 23,212 | 1,406 | 973 |
| # of Tracts | 8,057 | 1,302 | 937 |
| # of Zip codes | 1,769 | 1,158 | 844 |

b. Geographic Areas

Table 1. Breakdown of M-Lab Data.

people. It is also the smallest geographic unit for which the Census Bureau publishes sample data [41]. A census tract is formed with at least one census block group [41] and contain a population size between 1200 and 8000 people. A zip code is a U.S. Postal Service designated area. While a zip code contains an arbitrary numbers of census block groups [40], it is not considered a census unit. The shape file for each geographic area is obtained from the resources provided by the Census Bureau [1]. We map the location of the data points in each of these geographic areas within California (CA). Table 1(b) presents the total number of each geographic area present within the state of CA and the M-Lab dataset. To reduce bias in the dataset, we omit areas from which we have less than ten measurement end points. To eliminate anomalous data points, we discard measurement values that lie in the top five and bottom five percentiles of each geographic area. We then aggregate the raw samples based on the median speed value recorded within that area. Figure 1 shows the location of the IP addresses present in our California dataset.



(a) Servers located in the U.S.

(b) Servers located in the non-U.S. portion of the world

Fig. 2. Download Speed of Measurements for Different Server Locations.

The M-Lab dataset measurements can be impacted by the measurement server characteristics such as location and load conditions. There are 152 total measurement servers worldwide, with 82 within the U.S. in our dataset. Among the ones in the U.S., 11 are within California. To account for the impact of distance between clients and servers in our M-Lab dataset, we use the Haversine formula [29] to calculate the great circle distance between the client and server location for each test. Figure 2 plots the download speed for each client-server distance in our dataset. We observe that measurement tests to servers outside of the U.S. ("World (non-U.S.)") almost always recorded lower download speeds (see Figure 2(b)). Thus, we ignore them for our analysis. In contrast, using measurements to servers in the U.S. (outside CA) has a marginal impact on the download speed. Thus, we consider all measurement tests to U.S.-based servers for our analysis.

### 2.2 Demographic Data from ESRI

ESRI is a We utilize the demography data provided by ESRI's Updated Demographics [42]. ESRI curates this yearly demographic dataset using multiple sources that provide current-year estimates and 5-year projections of various demographic attributes. This is the most recent demography data available that is known to have high accuracy [28]. Using [21], we obtain the demographic variables in different geographic areas within California. For our analysis, we choose four demographic attributes: median household income, population, education, and poverty rate. We divide the category of education into three subcategories: proportion of population in an area without a high school degree (no HS), with a high school degree (HS) and with a bachelors degree (Bachelors). We also include type of area (urban/rural). Prior work [6, 18, 51] has shown that these attributes affect Internet access availability. In contrast, our goal is to explore whether these attributes affect the quality of Internet access among users of Speed Test by M-Lab.

While the ESRI data represents the most recent and granular demographic attribute data available, it is sparse at the granularity of census blocks. For example, over 25% of all blocks in California do not have a corresponding median income value in this dataset. On the other hand, at the granularity of census block group, the dataset covers all locations. This fact, coupled with sparse M-Lab data at the block level, guides us

Table 2. Summary Statistics for QoS and Demographic Variables.

| Variables | Average | Median | Standard Deviation |
|---|---|---|---|
| Download Speed (Mbps) | 40.41 | 30.44 | 38.91 |
| RTT (ms) | 24.91 | 22 | 13.72 |
| Median Income ($) | 75,536 | 63,675 | 43,009 |
| No HS (%) | 5.78 | 3.32 | 7.31 |
| HS (%) | 13.51 | 13.11 | 7.56 |
| Bachelors (%) | 14.95 | 12.92 | 14.01 |
| Poverty (%) | 5.67 | 4.04 | 5.96 |
| Population | 1790 | 1530 | 1468 |

to conduct our analysis at the granularity of the census block group. Fortunately, in 2015, the FCC classified every census block group as either urban or rural [2]. We use this data source to classify the census block groups present in our dataset. The summary statistics of the download speed and demographic attributes, at the granularity of census block group, are presented in Table 2.

### 2.3 Critique

Our data and method of aggregation has several caveats and limitations. First, the potential shortcomings of crowdsourced Internet measurements using tools such as NDT are well known [43, 46]. These crowdsourced measurements may bias the performance tests such that the observed distribution deviates from the true underlying distribution of the metrics for the population of interest. Furthermore, our approach of using IP address geolocation to obtain the physical location of the IP addresses is also prone to inaccuracies [48]. Finally, the measurements obtained from the NDT test are not uniformly distributed across all geographic areas in California. As such, we are unable to get a balanced number of samples across all types of locations, such as urban and rural areas, as well as demographic attributes such as income, education and poverty level across the state.

## 3 IMPACT OF DEMOGRAPHIC ATTRIBUTES ON INTERNET QUALITY

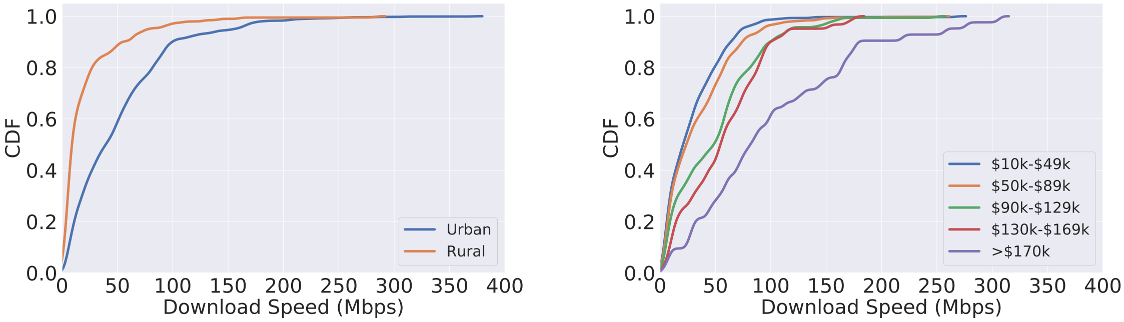
We begin by exploring the correlation between download speed with the selected demographic attributes at the granularity of the census block group. Based on our results, we then focus our analysis on area type and median income to determine their relationship to download speed.

### 3.1 Correlation between Download Speed and Demographic Attributes

We use the Pearson Correlation Coefficient (PCC) [38] as it is suitable to capture any relationship that might exist between demographic attributes and download speed. Table 3 shows the PCC metric, expressed in percentage, between the download speed and each of our chosen demographic attributes. We compute this metric separately for wired and wireless access types. Wired network samples show a higher degree of correlation with the demographic attributes compared to the wireless network samples. In particular,

Table 3. Pearson Correlation Coefficient between Download Speed and Demographic Attributes.

| Technology | Income | No HS | HS | Bachelors | Poverty | Population | Area Type |
|---|---|---|---|---|---|---|---|
| Fixed | 37.11 | -12.75 | -21.11 | 19.22 | -12.28 | 3.06 | -26.21 |
| Wireless | -1.59 | -2.26 | -3.25 | -1.75 | -7.64 | 0.8 | 3.3 |



a. Area

b. Income

Fig. 3. Cumulative Distribution Function of Download Speed by Area Type and Income.

*the median income is the most highly correlated with download speed*: growth in median income leads to an increase in the download speed. We observe a similar trend in Bachelors-level education and the overall population of the census block group. On the other hand, we observe a negative correlation between the download speed and the proportion of the block group population without or up to a high school degree. Similarly, download speed is also observed to be negatively correlated with the census block group's poverty rate. For area, we encode urban block groups as 0, rural block groups as 1, and perform special point-biserial correlation (equivalent to Pearson Correlation) [39] with download speed. This results in a negative correlation of download speed with rural areas. We observe that a census block group's population has the lowest correlation with download speed compared to other demographic attributes.

Compared to wired samples, we do not observe similar trends for the wireless measurements. This is likely attributable to the fact that, unlike in fixed networks where one can improve the download speed by opting for a more expensive subscription, higher subscription fees impact data volume instead of speed in wireless networks. Also, our dataset has many fewer samples for the wireless network. Thus, we focus on the wired network's measurement data for the remainder of our analysis.

Table 3 indicates that other than area type and income, the remainder of the attributes correlate poorly with download speed. This is due to the relative imbalance of block groups that fall within each demographic variable's categories. For example, only 8% of block groups have a poverty level of 25% or more. Given that the area type and median income have the strongest relationship with download speed, we more deeply analyze the relationship of different categories within these factors to download speed.

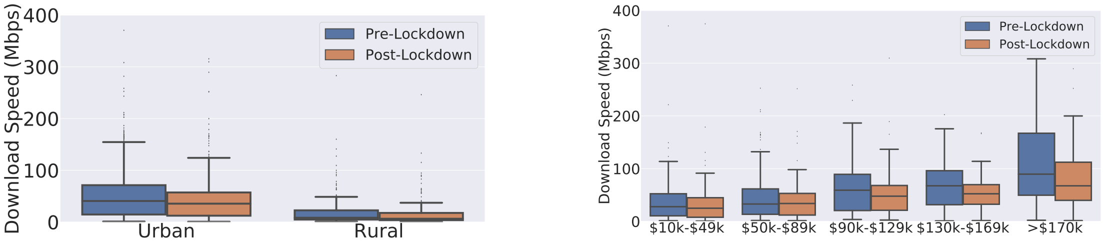
*3.1.1 Effect of Area Type.* Table 3 indicates the strong relationship between area type and download speed. There are 206 and 767 rural and urban census block groups in our data set, respectively. Figure 3(a) shows the cumulative distribution function of download speed in each of these block group categories. We note the significant difference that exists between download speed in rural areas versus urban. The average download speed recorded in rural block groups is 17.94 Mbps. This is well below the FCC definition of download broadband of 25 Mbps [3]. In comparison, urban block groups recorded an average download speed of 44.37 Mbps, almost 2.5 times the average speed recorded in rural areas. The inter-quartile range (IQR) for rural areas was 12.18 Mbps. Comparatively, the IQR for urban areas was 47.76 Mbps. 87% of the rural block groups recorded download speeds of less than 25Mbps, the broadband threshold defined by the FCC. In comparison, only 7% of urban block groups recorded less than 25Mbps of median download speed. These statistics capture the difference in quality of Internet that exists between rural and urban areas and point towards a gap in usability of Internet between these regions.

*3.1.2 Effect of Median Income.* While rural block groups may indicate a relationship between income groups and download speed, in this study we focus our income analysis on urban census block groups given the heavy skew of our dataset towards this area type.

We begin by breaking the urban census block group incomes into five bins, where each bin represents an increase in income by $40, 000 (based on the observed standard deviation of income data in the census block groups). There were 243, 288, 139, 55 and 42 census block groups in our income bins 1-5, respectively, where income bin 1 represents the lowest income group (less than $50, 000) and bin 5 represents the highest. Figure 3(b) shows the cumulative distribution functions of speed in these income bins. We can see that there is evidence of increasing download speed as the income level within these urban census block groups increases. Income bin 1 recorded the lowest average speed of 33.81 Mbps. The average download speed progressively increased to 39.52 Mbps, 53.91 Mbps, 58.34 Mbps and 93.15 Mbps for income bins 2 to 5, respectively. The corresponding IQRs for income bins 1-5 are 38.07 Mbps, 45.40 Mbps, 56.13 Mbps, 52.91 Mbps and 104.62 Mbps. respectively. This shows that even within urban areas, digital inequalities are still evident across users of Speed Test by M-Lab from different income groups. Importantly, *the average speed for the Speed Test by M-Lab users of the lowest urban income group is higher than that of the average download speed in rural block groups; however, it remains almost three times less than that recorded for the highest income group.*

### 3.2 Impact of the COVID-19 Lockdown on Download Speed

The California governor issued a lockdown/stay-at-home order on March 19, 2020 to curb the spread of the COVID-19 virus [15]. As found in a recent study [50], this COVID-19 lockdown led to changes in Internet traffic patterns nationwide; increased load in residential broadband networks have been observed as daily activities, such as work and school, shifted online. Based on this finding, our goal is to determine whether the COVID-19 lockdown caused any impact on the quality of Internet access during this period. To do so, we



a. Area Type

b. Income

Fig. 4. Download Speed before and during Lockdown by Area Type and Income.

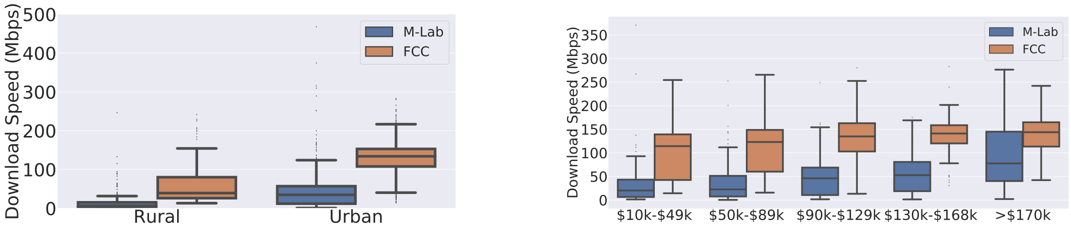
Table 4. Average Download Speed for Area Type and Income Pre- and Post-COVID-19 Lockdown.

| | **Variables** | **Pre-Lockdown (Mbps)** | **Post-Lockdown (Mbps)** |
|---|---|---|---|
| Area Type | Rural | 20.5 | 16.5 |
| Area Type | Urban | 50.38 | 41.81 |
| Median Income | $10k-$49k | 36.89 | 30.99 |
| Median Income | $50k-$89k | 44.26 | 38.38 |
| Median Income | $90k-$129k | 63.1 | 50.23 |
| Median Income | $130k-$169k | 71.09 | 53.97 |
| Median Income | >$170k | 104.91 | 86.92 |

divide our M-Lab data into two datasets to cover the pre- and post-lockdown time frames. 52% of the total M-Lab measurements in our dataset were recorded pre-lockdown, with the rest occurring post-lockdown. Figure 4 presents the speed recorded during these two periods, disaggregated by area type and urban census block group income bins.

Figure 4(a) shows the speed recorded during these two periods within urban and rural block groups. Table 4 provides the recorded average speed in these two area types during these periods. The average speed decreased by almost 20% during the lockdown in rural areas. A similar effect is observed in urban areas where, before lockdown, the average speed measured 50.38 Mbps, but reduced to 41.81 Mbps during the lockdown period. Critically, even as the average speed decreased in both location types, the average urban download speed remained 2.5 times the average rural speed.

In Figure 4(b), the download speeds recorded before and during the lockdown in urban block groups are grouped by income. From Table 4, we observe that the average speed across all income groups decreased during the lockdown period. For income bin 1, the average download speed was 36.89 Mbps before lockdown. However, this value decrease by 16% during lockdown to 30.99 Mbps. The average download speed in income bin 2 is reduced by 5.88 Mbps, while the average speed for income bin 3 decreased 20% during lockdown to 50.23 Mbps from 63.1 Mbps. *The average speed during lockdown for the two highest income groups decreased the most.* While income bin 4 shows the greatest drop (nearly 25%) in average download



a. Area

b. Income Bins

Fig. 5. Comparison of M-Lab and FCC Download Speed by Area Type and Income.

speed, income bin 5 also experienced a decrease by nearly 18 Mbps. Nevertheless, the average speed of the highest income group remained three times that of the lowest income group.

### 3.3 Discrepancy between FCC and M-Lab Download Speeds

The FCC defines “advertised" download speed as that reported by fixed service providers through Form 477 at the geographic granularity of a census block. The requirement for a service provider to claim coverage in a census block is that it can provide a download speed of at least 200 kbps in *at least one location* within the census block. Given the well-documented inaccuracy of this data [4, 51], we explore how it compares to the actual measurements collected from Speed Test by M-Lab users of different locations and income levels.

We aggregate FCC speed data at the granularity of census block groups by taking the median of the download speed of the blocks within a block group. Figure 5 compares aggregated census block group measurement values obtained from M-Lab and FCC data broken down by area type and income bins within urban block groups. The graphs clearly show that the FCC data tends to estimate significantly higher speeds, anywhere from 8 Mbps to 114 Mbps, than the M-Lab users experience across all locations and income bins. This mismatch may be explained in part by the ISP plan tier purchased by users; users may not always purchase the best/fastest plan offered by an ISP. It may also be explained in part by the timing of user Speed Tests; if users conduct Speed Tests when they are experiencing sub-par performance, then we would expect to see poorer results. On the other hand, it is also likely that in many areas providers overstate coverage speeds [51]. With the available data, it is not clear which explanation accounts for the greatest portion of the discrepancy.

To more deeply analyze the difference between FCC and M-Lab recorded download speed, for each block group within an area type and income level, we calculate an accuracy factor by taking the ratio of the download speed from M-Lab and FCC. To summarise the accuracy factor for each variable, we take the average of the accuracy factors for all block groups that belong to that variable. As seen from Table 5, *the accuracy factor is lowest in the case of rural areas, indicating that the FCC estimated download speed tends to be most different from what is recorded through Speed Test by M-lab in these regions.* While at first glance

Table 5. FCC Accuracy Factor by Area Type and Income.

| | **Variables** | **# of Block Groups** | **Factor** |
|---|---|---|---|
| Area Type | Rural | 206 | 0.32 |
| Area Type | Urban | 767 | 0.4 |
| Median Income | $10k-$49k | 243 | 0.36 |
| Median Income | $50k-$89k | 288 | 0.36 |
| Median Income | $90k-$129k | 139 | 0.42 |
| Median Income | $130k-$169k | 44 | 0.54 |
| Median Income | >$170k | 42 | 0.88 |

it appears as if the level of mismatch for both rural and urban areas are similar, accuracy factors across different income bins in urban block groups suggest otherwise. *Among income bins in urban areas, the accuracy factor is lowest for the two smallest income groups.* This points towards the FCC's record of much higher speed in these areas than what is captured in M-Lab dataset. The accuracy factor increases as the income increases, suggesting for higher income urban areas either there is i) more accurate reporting on part of the service providers and/or ii) higher purchasing power of the end users, leading to purchase of higher/better tiers of Internet service compared to the lower income areas. One shortcoming of the FCC's database is that it fails to capture the user's tier of subscription, and hence the maximum download speed, purchased by users. Further, our analysis demonstrates the discrepancy between the download speeds claimed by the service providers and what is obtained through Speed Test, thereby highlighting the need for more accurate documentation of download speeds, by both actual availability and affordability, across diverse locations and demographic attributes.

## 4 DISCUSSION AND RECOMMENDATIONS

There are several key takeaways from our analysis that can help researchers, practitioners and government officials address the factors that perpetuate digital inequality.

**Accurate Internet Measurement Data.** Given the limitations that exist in the FCC's current reliance on ISP-provided data to document available speed in a census block, coupled with the sparse geographical coverage of current crowdsourced Internet measurement tools, there is a need to develop better approaches to obtain a more accurate and complete representation of Internet availability and quality. The FCC itself has recognized the shortcomings of its current methodology and highlighted the need for higher quality data through recent initiatives [12, 30]. An added complexity is the lack of detail on available service plans, as well as the plans and data rates to which users actually subscribe. Without this critical information, it is difficult to fully understand the context behind the performance values reported through tools such as M-Lab's Speed Test.

Nevertheless, despite the fundamental limitations of crowdsourced measurement tools, our M-Lab study reveals there is a gap in Internet access quality across varied locations and demographic attributes. While some of the gap may be explained by users purchasing different service plan tiers, without further detail, it is critical to investigate more deeply the source of these disparities. Our preliminary work on service plan pricing (not presented here), and specifically our work to map download speed (and corresponding price) offered by ISPs to geographic location, has demonstrated multiple sources of digital inequality. Our current and future work attempts to quantify this disparity.

With more accurate Internet measurement data, our approach can be extended to much finer geographic granularity. Our findings also add to the body of work that has demonstrated the inaccuracies of FCC data across different area types and income levels. To address digital inequalities between communities, accurate documentation of quality metrics such as download speed is crucial. Our findings indicate rural areas and low income regions experience the greatest FCC inaccuracy. Therefore, more attention needs to be paid to these areas to accurately capture true Internet performance, as well as general Internet access availability, to guide future broadband deployment efforts.

**Fine-grained Demographic Data.** 2010 Census demographic attributes, such as poverty rate and education, are currently only available at the tract level. Hence the establishment of relationships between these variables and Internet access quality is challenging. The 2020 Census data, once fully available, is likely to be the most accurate and current demography data available within the near future. As such, the granularity of the reporting of this data needs to be finer in order to better correlate the relationship between demographic attributes and Internet access quality within smaller geographic regions.

## 5 RELATED WORK

Every year, the Census, through the American Community Survey (ACS) One Year estimates, compiles a list of cities with the worst Internet connectivity in the country [26]. However, this estimate is only done for cities with population greater than 65, 000, leaving smaller communities undocumented. Similar to our work, [9] analysed the relationship between income and download speed at the geographic granularity of zip codes in the U.S. The work utilized income data (grouped into five income bins) obtained from 2017 tax returns filed with the Internal Revenue Service. The study demonstrated a positive correlation between zip code income and download speed. Our work confirms this finding at the finer geographic granularity of census block groups in California. We also demonstrate that FCC data overestimates available speed to a greater degree in low income census block groups.

Prior research has focused on the analysis of demographic factors that affect the penetration and diffusion of Internet access in different geographic areas. In a recent study conducted by Microsoft [19], it was estimated that 162.8 million Americans did not have access to high-speed broadband, a number far greater than the FCC's estimate. The study was conducted at the granularity of zip code and, similar to our work, IP address geolocation was used to locate users within each zip code. A similar study [14] estimated 42 million Americans have Internet download speeds of less than 25 Mbps, double the estimate of FCC. Through our work, we show that in addition to overestimating the population with access to the Internet, the FCC also overestimates the quality of that Internet access, in terms of download speed; this overestimation is particularly large in lower income areas. The authors in [47] combined demographic information with Internet infrastructure data provided by the California Public Utilities Commission (CPUC). Their analysis revealed areas with low income minority population were less likely to have access to residential fiber services that provide better Internet performance. Similarly, in [52, 54–56], demographic factors such as location, race and/or income are all shown to impact Internet access. We advance this body of work and demonstrate that while areas may have Internet access, the quality of that access remains worse for lower income populations.

Finally, similar to our work, the authors of [43] used crowdsourced measurements to benchmark Internet performance across multiple metropolitan areas. In [44], cable and Digital Subscriber Line performance in residential areas of North America and Europe was characterized. Finally, cost effective deployment solutions were proposed to increase coverage in unserved areas in [45].

## 6 CONCLUSION

In this work, we analyze Internet access quality across the state of California for users of Speed Test by M-Lab. Our results study the characteristics of digital inequality that exists among the user base of M-Lab across different locations and demographic attributes within the state. Additionally, we highlight the shortcoming of the FCC's documentation of broadband speed as its current methodology significantly overestimates download speed in rural and poorer areas. Our findings point towards the need to develop more accurate Internet coverage and quality measurement tools to discover additional factors that affect Internet access availability and quality across diverse communities. We hope that our analysis can help guide the efforts of policymakers and researchers in narrowing the digital gap between communities.

## ACKNOWLEDGEMENT

We wish to thank the anonymous COMPASS reviewers for their valuable feedback on the paper. We also thank M-Lab, for their dedication to ensuring their measurement data is publicly available. This work was funded through National Science Foundation Rapid Response Research (RAPID) award NSF-2033946.

## REFERENCES

[1] 2010. TIGER Line Shapefile. Retrieved 10/18/2021 from https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Blocks

[2] 2015. A-CAM Census Block Groups PN. Retrieved 09/30/2021 from https://www.fcc.gov/document/cam-census-block-groups-pn

[3] 2016. Residential Fixed 25 Mbps/3 Mbps Broadband Deployment. Retrieved 10/23/2021 from https://www.fcc.gov/reports-research/maps/bpr-2016-fixed-25mbps-3mbps-deployment/

[4] 2018. When You Can't Trust the Data, Flaws in the Federal Communications Commission's Broadband Forms. Retrieved 10/25/2021 from https://ilsr.org/when-you-cant-trust-the-data-flaws-in-the-federal-communications-commissions-broadband-forms/

[5] 2019. America's Digital Divide. Retrieved 10/8/2021 from https://www.pewtrusts.org/en/trust/archive/summer-2019/americas-digital-divide

[6] 2020. America's Racial Gap & Big Tech's Closing Window. Retrieved 09/28/2021 from https://www.db.com/newsroom_news/

[7] 2020. Coronavirus for kids without internet: Quarantined worksheets, learning in parking lots. Retrieved 10/10/2021 from https://www.usatoday.com/story/news/education/2020/04/01/coronavirus-internet-speed-broadband-online-learning-school-closures/5091051002/

[8] 2020. COVID-19 and the rise of Telemedicine. Retrieved 09/16/2021 from https://medicalfuturist.com/covid-19-was-needed-for-telemedicine-to-finally-go-mainstream/

[9] 2020. Decoding the digital divide. Retrieved 10/05/2021 from https://www.fastly.com/blog/digital-divide

[10] 2020. During coronavirus, high-speed internet is a lifesaver - that millions lack. Retrieved 09/30/2021 from https://www.nbcnews.com/think/opinion/during-coronavirus-high-speed-internet-lifesaver-millions-lack-ncna1165321

[11] 2020. Expanding Internet Access Improves Health Outcomes. Retrieved 09/28/2021 from https://www.govtech.com/network/Expanding-Internet-Access-Improves-Health-Outcomes.html

[12] 2020. FCC Improves Broadband Data and Maps to Bridge the Digital Divide. Retrieved 10/23/2021 from https://www.fcc.gov/document/fcc-improves-broadband-data-and-maps-bridge-digital-divide-0

[13] 2020. FCC Proposes the 5G Fund for Rural America. Retrieved 10/20/2021 from https://www.fcc.gov/document/fcc-proposes-5g-fund-rural-america

[14] 2020. FCC Underestimates Americans Unserved by Broadband Internet by 50%. Retrieved 10/21/2021 from https://broadbandnow.com/research/fcc-underestimates-unserved-by-50-percent

[15] 2020. Governor Gavin Newsom Issues Stay at Home Order. Retrieved 09/28/2021 from https://www.gov.ca.gov/2020/03/19/governor-gavin-newsom-issues-stay-at-home-order/

[16] 2020. Guidelines for Broadband Data Submission. Retrieved 10/23/2021 from https://www.cpuc.ca.gov/industries-and-topics/internet-and-phone/broadband-mapping-program/guidelines-for-broadband-data-submission

[17] 2020. Social Distancing, Internet Access and Inequality. Retrieved 10/20/2021 from https://www.nber.org/papers/w26982.pdf

[18] 2020. The Results Are In for Remote Learning: It Didn't Work. Retrieved 10/16/2021 from https://www.wsj.com/articles/schools-coronavirus-remote-learning-lockdown-tech-11591375078

[19] 2020. U.S. Broadband Usage Percentages. Retrieved 10/16/2021 from https://github.com/microsoft/USBroadbandUsage\Percentages

[20] 2020. U.S. Schools Trying to Teach Online Highlight a Digital Divide. Retrieved 10/05/2021 from https://www.bloomberg.com/news/articles/2020-03-26/covid-19-school-closures-reveal-disparity-in-access-to-internet

[21] 2021. ArcGIS. Retrieved 09/28/2021 from https://www.arcgis.com/index.html

[22] 2021. archive-measurement-lab. Retrieved 10/25/2021 from https://console.cloud.google.com/storage/browser/archive-measurement-lab/ndt/ndt7/?forceOnBucketsSortingFiltering=false&project\=measurement-lab

[23] 2021. Census block. Retrieved 10/25/2021 from https://en.wikipedia.org/wiki/Census_block

[24] 2021. Census Blocks and Block Groups. Retrieved 10/25/2021 from https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf

[25] 2021. Check your connection. Retrieved 10/26/2021 from https://projectstream.google.com/speedtest

[26] 2021. Computer and Internet Use. Retrieved 09/28/2021 from https://www.census.gov/content/census/en/programs-surveys/acs/library/keywords/computer-and-internet-use.html/

[27] 2021. ESRI. Retrieved 09/28/2021 from https://en.wikipedia.org/wiki/Esri

[28] 2021. ESRI's Demographics: #1 for Accuracy. Retrieved 09/28/2021 from https://www.esri.com/library/fliers/pdfs/esris-demographics-accuracy.pdf

[29] 2021. Haversine formula. Retrieved 10/25/2021 from https://en.wikipedia.org/wiki/Haversine_formula

[30] 2021. Input Sought on Mobile Challenge, Verification Technical Requirements. Retrieved 10/25/2021 from https://www.fcc.gov/document/input-sought-mobile-challenge-verification-technical-requirements

[31] 2021. IPinfo.io. Retrieved 09/28/2021 from ipinfo.io

[32] 2021. Locate API Usage. Retrieved 09/28/2021 from https://github.com/m-lab/locate/blob/master/USAGE.md

[33] 2021. Locate API v1. Retrieved 09/28/2021 from https://www.measurementlab.net/develop/locate-v1/

[34] 2021. Locate API v2. Retrieved 09/28/2021 from https://www.measurementlab.net/develop/locate-v2/

[35] 2021. Measurement Lab. Retrieved 10/25/2021 from https://www.measurementlab.net/

[36] 2021. MLab Test Your Speed. Retrieved 10/27/2021 from https://speed.measurementlab.net/#/

[37] 2021. NDT-Network Diagnostic Tool. Retrieved 10/25/2021 from https://www.measurementlab.net/tests/ndt/ndt7/

[38] 2021. Pearson Correlation Coefficient. Retrieved 10/25/2021 from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[39] 2021. Point-biserial Correlation. Retrieved 09/28/2021 from http://web.pdx.edu/~newsomj/pa551/lectur15.htm#:~:text=A%20point%2Dbiserial%20correlation%20is,variable\%20and%20one%20continuous%20variable.&text=So%20computing%20the%20special\%20point,and%20the%20other%20is%20continuous.

[40] 2021. Relating Block Groups to ZIP Code Areas. Retrieved 10/25/2021 from http://proximityone.com/bg-zip.htm

[41] 2021. Research 101: Census Tracts vs. Census Block Groups. Retrieved 10/25/2021 from https://current360.com/research-101-census-tracts-vs-census-block-groups/

[42] 2021. Updated Demographics. Retrieved 09/28/2021 from https://doc.arcgis.com/en/esri-demographics/data/updated-demographics.html

[43] Igor Canadi, Paul Barford, and Joel Sommers. 2012. Revisiting Broadband Performance. In *Proceedings of the 2012 Internet Measurement Conference* (Boston, Massachusetts, USA) *(IMC '12)*. Association for Computing Machinery, New York, NY, USA, 273–286.

[44] Marcel Dischinger, Andreas Haeberlen, Krishna P. Gummadi, and Stefan Saroiu. 2007. Characterizing Residential Broadband Networks *(IMC '07)*. Association for Computing Machinery, New York, NY, USA, 43–56.

[45] Ramakrishnan Durairajan and Paul Barford. 2017. A Techno-Economic Approach for Broadband Deployment in Underserved Areas. *Computer Communication Review* 47 (04 2017), 13–18.

[46] Nick Feamster and Jason Livingood. 2019. Internet Speed Measurement: Current Challenges and Future Recommendations. arXiv:1905.02334 [cs.NI]

[47] Hernan Galperin, Thai V. Le, and Kurt Wyatt. 2021. Who gets access to fast broadband? Evidence from Los Angeles County. *Government Information Quarterly* 38, 3 (2021), 101594.

[48] Manaf Gharaibeh, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi, and Christos Papadopoulos. 2017. A Look at Router Geolocation in Public and Commercial Databases. In *Proceedings of the 2017 Internet Measurement Conference* (London, United Kingdom) *(IMC '17)*. Association for Computing Machinery, New York, NY, USA, 463–469.

[49] M. Hilbert. 2016. The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 1986–2014. *Telecommunications Policy* 40, 6 (06 2016), 567–581.

[50] Andra Lutu, Diego Perino, Marcelo Bagnulo, Enrique Frias-Martinez, and Javad Khangosstar. 2020. A Characterization of the COVID-19 Pandemic Impact on a Mobile Network Operator Traffic. arXiv:2010.02781 [cs.NI]

[51] David Major, Ross Teixeira, and Jonathan Mayer. 2020. No WAN's Land: Mapping U.S. Broadband Coverage with Millions of Address Queries to ISPs. In *Proceedings of the ACM Internet Measurement Conference (IMC '20)*. 393–419.

[52] Steven P. Martin and John P. Robinson. 2014. The Income Digital Divide: Trends and Predictions for Levels of Internet Use. *Social Problems* 54, 1 (07 2014), 1–22. arXiv:https://academic.oup.com/socpro/article-pdf/54/1/1/4557260/socpro54-0001.pdf

[53] M. Mathis and M. Allman. 2001. *A Framework for Defining Empirical Bulk Transfer Capacity Metrics*. RFC 3148. https://tools.ietf.org/html/rfc3148

[54] James E. Prieger. 2003. *The Supply Side of the Digital Divide: Is There Equal Availability in the Broadband Internet Access Market?* Working Papers 50. University of California, Davis, Department of Economics. https://ideas.repec.org/p/cda/wpaper/50.html

[55] James E. Prieger and Wei-Min Hu. 2008. The broadband digital divide and the nexus of race, competition, and quality. *Information Economics and Policy* 20, 2 (2008), 150 – 167.

[56] Brian Whitacre, Roberto Gallardo, and Sharon Strover. 2014. Does rural broadband impact jobs and income? Evidence from spatial and first-differenced regressions. *The Annals of Regional Science* 53 (Nov. 2014), 649–670.