University of California
Santa Barbara

Towards Bridging the Divide: Enhancing Understanding of Digital Inequity

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Computer Science

by

Udit Paul

Committee in charge:

Professor Elizabeth Belding, co-Chair,
Assistant Professor Arpit Gupta, co-Chair
Professor Amr El Abbadi

September 2023

The Dissertation of Udit Paul is approved.

Professor Amr El Abbadi

Assistant Professor Arpit Gupta, co-Chair

Professor Elizabeth Belding, co-Chair

September 2023

Towards Bridging the Divide: Enhancing Understanding of Digital Inequity

by

Udit Paul

Acknowledgements

The journey of my Ph.D. has been truly remarkable. From learning to stay motivated and doing research during the COVID lockdown, to dealing with paper rejections, I truly believe that Ph.D has taught me a great deal about myself and instilled values that I will carry with me throughout my life. In this incredible journey, I have to express profound gratitude to everyone who has offered me support, motivation, and guidance throughout my journey.

First and foremost, my heartfelt gratitude goes to my exceptional co-advisors, Elizabeth Belding and Arpit Gupta. I want to thank Elizabeth for believing in me and giving me an opportunity to realize my lifelong dream of becoming a Doctor. Every step of the way, I have felt her support and genuine care toward both my academic and personal life. I have great admiration for her exceptional patience during our meetings and her approach to comprehending intricate concepts from their fundamental foundations. I want to convey my appreciation to Arpit for consistently staying involved in refining the raw ideas and assisting me in transforming them into their most refined versions. I will cherish the memories of the regular and lengthy meetings in his office, especially before the paper deadlines. Both of my advisors have been a source of constant inspiration throughout my Ph.D. journey at UCSB. I would also like to extend my gratitude to Amr El Abbadi for his expertise and invaluable mentorship as a member of my Ph.D. committee.

I consider myself extremely fortunate to have had the privilege of being in the company of outstanding individuals throughout my Ph.D. stint. People such as Vivek Adarsh, Jiamo Liu, and Michael Nekrasov not only served as remarkable collaborators but also proved to be incredible friends. I would also like to say a big thank you to all my other amazing colleagues for their support, companionship, and the joy they brought each day.

Their presence made my Ph.D. journey truly memorable and fulfilling.

Finally, I would like to thank my parents, Dr. Amal Paul and Sabita Paul, for all the sacrifices they have made to ensure that I achieve my dreams. I am also deeply grateful to my sister, Dr. Adrita Paul, for supporting me every step of the way and always encouraging me to do more. Their unconditional love has been a constant source of motivation and driving force, pushing me to constantly strive for growth and be the best version of myself every day of my life.

Curriculum Vitæ

Udit Paul

Education

University of California Santa Barbara **Expected: September 2023**
Doctor of Philosophy (Ph.D.), Computer Science

University of Cape Town **June 2018**
Master of Science (M.S.), Electrical Engineering

Professional Experience

Graduate Research Assistant **September 2018 – Present**
University of California Santa Barbara *Santa Barbara, California*

- Developed Broadband-plan Querying Tool (BQT) that employs web scrapers for multiple internet service providers' web services to scalably obtain broadband services offer-related information such as download speed and cost of access at any given street address in the US.
- Developed Broadband Subscription Tier (BST) methodology which is a two-stage hierarchical unsupervised classification technique to identify broadband subscription tier information from crowdsourced network measurements. BST is able to classify subscription tiers of measurements with **99%** accuracy.
- Curated a data set of 17,000 tweets obtained from the social media platform, Twitter, and developed a natural language processing framework to detect and isolate power and communication outage-related tweets to assist first responders in the event of natural disasters. Implemented 22 different machine learning algorithms and achieved close to **90%** accuracy in performing the required classification task.

PhD Research Intern **June 2022 – September 2022**
IBM *Yorktown Heights, New York*

- Contributed to an Agent-Based Model (ABM) that simulates the impact of lack of good quality internet connectivity on populations of different socioeconomic statuses. Built the component of the model that estimates the effect of poor quality internet for different households. Additionally, contributed to the building of a web application that provides the whole model as a service to the public. Work under review for publication.
- Conducted a longitudinal analysis on a dataset of **30M** to understand how the internet quality has changed for different population groups. Applied statistical tests to determine the magnitude of difference in internet quality between sub-populations. Finally, deployed machine learning models to predict internet performance from demographics and infrastructure metrics. Work under review for publication.

Awards

- UCSB CS Department Outstanding Dissertation Award, 2023.
- UCSB CS Department Outstanding Publication Award, 2023.
- Best Paper Award winner, ACM SIGCOMM IMC'22.
- Best Poster runner-up, ACM HotMobile'19.

Refereed Publications

- **Udit Paul**, Jiamo Liu, Vinothini Gunasekaran, Tejas N Narechania, Arpit Gupta, and Elizabeth Belding, "Decoding the Divide: Analyzing Disparities in Broadband Plans Offered by Major US ISPs", Proceedings of the ACM SIGCOMM 2023 (**SIGCOMM'23)**.
- **Udit Paul**, Jiamo Liu, Mengyang Gu, Arpit Gupta, and Elizabeth Belding, "The Importance of Contextualization of Crowdsourced Active Speed Test Measurements", Proceedings of the ACM SIGCOMM Internet Measurement Conference 2022 (**IMC'22**). [***Best paper***].
- Tarun Mangla, **Udit Paul**, Arpit Gupta, Nicole Marwell, and Nick Feamster, "Internet Inequity in Chicago: Adoption, Affordability, and Availability", Proceedings of the 50^{th} Research Conference on Communications, Information and Internet Policy (**TPRC'22**).
- **Udit Paul**, Jiamo Liu, David Farias-Llerenas, Vivek Adarsh, Arpit Gupta, and Elizabeth Belding, "Characterizing Internet Access and Quality Inequities in California M-Lab Measurements", Proceedings of the Conference on Computing and Sustainable Societies (**COMPASS'22**).
- Vivek Adarsh, Michael Nekrasov, **Udit Paul**, Tarun Mangla, Arpit Gupta, Morgan Vigil-Hayes, Ellen Zegura, and Elizabeth Belding, "Coverage is not binary: Quantifying mobile broadband quality in urban, rural, and tribal contexts", Proceedings of the International Conference on Computer Communications and Networks (ICCCN) 2021 (**ICCN'21**).
- Vivek Adarsh, Michael Nekrasov, **Udit Paul**, Alex Ermakov, Arpit Gupta,Morgan Vigil-Hayes, Ellen Zegura, and Elizabeth Belding, "Too Late for Playback: Estimation of Video Stream Quality in Rural and Urban Contexts", Proceedings of the Passive and Active Measurement Conference (**PAM'21**).
- Vivek Adarsh, Michael Nekrasov, **Udit Paul**, and Elizabeth Belding, "Estimation of congestion from cellular walled gardens using passive measurements", IEEE Transactions on Mobile Computing, vol. 21, no. 10, 2021.
- **Udit Paul**, Alex Ermakov, Michael Nekrasov, Vivek Adarsh, and Elizabeth Belding, "#Outage: Detecting Power and Communication Outages from Social Networks", Proceedings of the World Wide Web Conference (**WWW'20**).

- Michael Nekrasov, Vivek Adarsh, **Udit Paul**, Esther Showalter, Ellen Zegura, Morgan Vigil-Hayes, and Elizabeth Belding, "Evaluating LTE coverage and quality from an unmanned aircraft system", International Conference on Mobile Ad Hoc and Sensor Systems **(MASS'19)**.

Invited Talks

- **Contextualizing crowdsourced measurements** **Oct. 2022**
  Ookla
- **Improving understanding of crowdsourced measurements** **Sept. 2022**
  Google - Measurement Lab

Teaching Assistanstship

- Introduction to Computer Networks (CMPSC 176A) - Spring, 2019.
- Computer Networking (CMPSC 176B) - Winter, 2019.
- Introduction to Computer Networks (CMPSC 176A) - Fall, 2018.

Abstract

Towards Bridging the Divide: Enhancing Understanding of Digital Inequity

by

Udit Paul

The Internet has become crucial for communication, education, commerce, and civic engagement, but not everyone has equal opportunities to benefit from it, leading to digital inequity. This inequity stems from various aspects of Internet access, such as availability, quality, and affordability. Policymakers and stakeholders must understand the presence and extent of digital inequity to develop strategies that can bridge the gaps and ensure equal Internet access for all.

Acquiring relevant data that sheds light on all aspects of digital inequity is imperative for building a complete understanding of the issue. Unfortunately, such data is currently either non-existent or too noisy to be of any use. Policymakers in the US have long relied on imprecise data obtained either from the Federal Communication Commission or through crowdsourced network measurements to estimate the availability and quality of Internet services in different regions, and allocate funding accordingly to improve Internet access. However, due to the limitations of these datasets, funding initiatives that rely on them may not achieve their intended objectives. Additionally, there are no publicly available sources of data that can provide accurate information on the cost of Internet access across the nation. As a result, it is extremely challenging to understand Internet affordability and how that contributes to digital inequity.

This dissertation aims to address these challenges in several ways. Firstly, we characterize existing Internet access datasets to gain insights into current digital inequity trends. Additionally, we develop methodology and tools that can provide comprehensive data on

various dimensions of digital inequity. Leveraging our solutions, we enhance the usability of crowdsourced network measurements to better understand Internet quality. Moreover, we curate multiple novel datasets that provide insights into Internet availability and affordability nationwide. This work is crucial in helping policymakers and organizations make informed decisions to address digital inequity and create a more equitable digital society.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the modern era, the significance of high-quality and affordable Internet cannot be overstated. The Internet is not just a network of computers; it is a fundamental cornerstone of the modern information society–an essential medium that is now as pivotal to the growth and prosperity of communities as transportation and electricity were in the 20th century. This is further evidenced by the recognition of the Internet as a basic human right by the United Nations [97]. However, despite witnessing the rapid expansion of Internet accessibility in the last decade, certain segments of the population continue to be excluded from its advantages, giving rise to the pervasive issue known as digital inequity or the digital divide [115].

Digital inequity can be succinctly described as the disparity between individuals who possess affordable access, skills, and support to actively participate in online activities and those who lack such resources. As such, it disenfranchises underserved and underprivileged communities from the benefits of an Internet-based modern economy, a problem that intensifies significantly over time. Even in the US, digital inequity remains present as highlighted by a recent Pew Research Center study [51] indicating that about 7% of the population remain non-users of the Internet, while 23% (1 in 4 people) lack access to

a home broadband connection. Digital inequity disproportionately affects specific population groups, such as those residing in rural areas or low-income households, as well as individuals from minority backgrounds living in urban areas.

There are several ways through which digital inequity can manifest. One of the significant dimensions of digital inequity is the lack of access to the Internet. The absence or insufficiency of network infrastructure can lead to a situation where people have limited or no access to the Internet. However, the mere presence of *network connectivity alone does not guarantee a usable service*. Hence, Internet access encompasses not only its *availability* but also its *quality*. Access issues can also occur through *affordability*, specifically referring to the cost of Internet services. Although a locality might have access to high-quality Internet services, its subscription may be financially burdensome for individuals, effectively resulting in a lack of access for many.

In addition to access, digital inequity can be pervasive through other dimensions such as lack of Internet adoption, skills, and support necessary to engage in the benefits provided by the Internet. While all the dimensions are critical, it is evident that access to the Internet **is a fundamental requirement to mitigate digital inequity**.

Acknowledging the significance of universal Internet access and the existence of digital inequity in the US, efforts are being made to address this issue. In 2022, the US Congress has allocated a substantial amount of **$42.5 billion** through the Broadband Equity, Access, and Deployment (BEAD) Program [187]. This project aims to improve Internet accessibility by supporting planning, infrastructure deployment, and adoption programs all across the US. This unprecedented funding initiative serves as a crucial catalyst in addressing long-standing digital inequities. Nevertheless, it is imperative to allocate these resources effectively to ensure the intended goal is achieved. Misallocation of funds could lead to mismanagement and, ultimately, failure to bridge the gap. Hence, careful and prudent allocation is essential so that the areas that currently lack proper Internet access

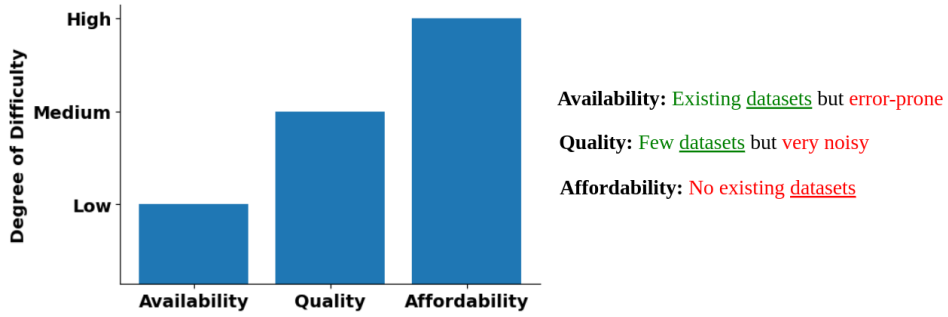


Figure 1.1: Degree of difficulty in assessing the current state of different dimensions of Internet access based on currently available data.

can reap the benefits of this initiative.

To improve Internet access, it is necessary to assess the existing state of Internet access nationwide. As mentioned earlier, Internet access encompasses multiple dimensions, including availability, quality, and affordability. Gaining a holistic perspective on the current state of Internet access involves having relevant data pertaining to each of these dimensions. However, each of these dimensions suffers from either accuracy, quality, and/or quantity issues in terms of the data required to comprehensively gain an understanding of how they contribute to digital inequity. Therefore, the degree of difficulty associated with assessing the current state of each of these dimensions vary, as depicted in Figure 1.1

Of the three dimensions of Internet access, the one that currently has the most available information is availability. Historically, this information is compiled bi-annually, at the coarse granularity of census block, by the US Federal Communication Commission (FCC). The eventual dataset of Internet availability is released through Form 477 [116]. Given the aggregate nature of this data, the limitations are widely recognized. Form 477 data overstate broadband availability because: (1) if an ISP only covers a single household in a census block, the entire block is considered covered; (2) "availability" means that the ISP does not necessarily service the area currently but could provide service

within an interval that is typical for that type of connection [77]. In response to the limitations of the imprecise Form 477 data and the recognition of the necessity for more detailed information, in 2020, the US Congress mandated the Federal Communications Commission (FCC) to create a precise map of broadband availability throughout the US [113]. This resulted in the Broadband Data Collection (BDC) by the FCC to create the National Broadband Map [78] of Internet availability in the US. Unlike Form 477, this data provides information about the Internet services available to **individual locations** (street addresses) across the country, along with new maps of mobile coverage, as reported by ISPs in the FCC's ongoing BDC. While still reported by the ISPs, this dataset serves as a massive improvement from Form 477 and provides Internet availability information at the finest granularity possible. Although in its initial stages and undergoing iterative improvements, there are reports indicating inaccuracies associated with this newly developed dataset. ISPs have been found to have provided false information [106] to the FCC giving rise to fear [121] that inaccuracies that plague the Form 477 may still be present in the new dataset.

While the existing Internet availability datasets face the challenge of inaccuracy, the limited data sources that can provide insights into Internet quality are hindered by the presence of noise. Fundamentally, measuring Internet quality on a large scale is an incredibly challenging task. The FCC operates the Measuring Broadband America (MBA) [141] project through which it longitudinally measures the Internet quality of 3k households, or 0.003% of total households, around the country. As an alternative, datasets gathered by crowdsourced speed test measurement tools such as Ookla's speedtest or Measurement Lab's speed test emerge as a potential alternative to understanding Internet quality in scale. These speed test platforms are designed to measure Internet quality and their subscribers can run these measurements at will. However, this uncontrolled setting of speed tests gives rise to a lot of noise. To interpret the results of speed tests, additional

contexts such as a subscriber's purchased subscription tier and/or local home network conditions need to be taken into account. However, current speed test tools either lack this information or only collect a few of the required parameters, leaving the datasets rife with noise and unusable to study Internet quality.

Finally, while some datasets are present to understand Internet availability and quality, there is a notable absence of a comprehensive dataset to explore Internet affordability at a large scale. The lack of granular information regarding the cost of Internet services across the country results in an inherent opacity within the Internet service provisioning industry. The absence of detailed data makes it challenging to examine significant trends regarding ISP practices, Internet service pricing, and overall Internet affordability across the country.

Taking into account the individual limitations related to each dimension, it becomes apparent that fully understanding the extensive scope of digital inequity in relation to Internet access is an exceptionally difficult undertaking. Relying on existing inaccurate and low-quality datasets to allocate funds for improving Internet access raises the risk of making uninformed investment decisions, which could ultimately impede the achievement of the intended objectives. It is therefore imperative to assess the existing state of Internet access with new and/or improved datasets pertaining to Internet availability, quality, and affordability.

This thesis centers around studying the current state of Internet access and aims to develop methodologies to enhance the usability of existing datasets. Additionally, it focuses on developing tools to contribute novel datasets that are necessary to advance our understanding of Internet availability, quality, and affordability.

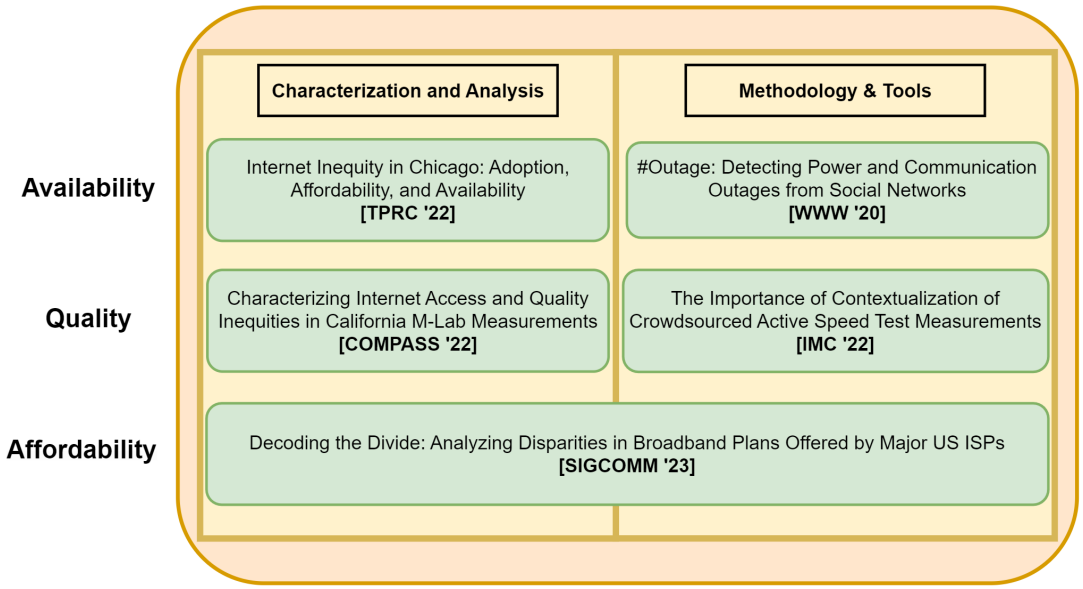


Figure 1.2: Dissertation overview.

1.1 Thesis Statement

This dissertation shows that:

Many people experience digital inequity as they do not have access to Internet services that are both affordable and of high quality. However, the extent of digital inequity remains unknown. To address this issue and bring benefits of the Internet to everyone, we must assess the presence and extent of digital inequity along three important axes: ***availability***, ***quality***, *and* ***affordability***.

Figure 1.2 presents the outline of this dissertation. In this thesis, we characterize the existing state of Internet availability [180] and quality [196] using available datasets. Additionally, we develop methodologies to enhance our understanding and utilization of these existing datasets [197]. Finally, we develop tools to contribute novel datasets that provide additional insights into the prevailing state of Internet access [194, 193].

1.2 Key Contributions

In this section, we discuss our broad contributions and their implications based on the outline shown in Figure 1.2.

1.2.1 Internet Availability

To enhance understanding of Internet availability, we i) analyze existing datasets and ii) create new methods and tools, including utilizing non-traditional sources like Twitter, to identify areas with limited Internet access.

- **Characterization and Analysis:** We combine multiple existing datasets to understand the digital divide/inequity in Chicago and the contributing factors [180]. The focus of this study is on Chicago as the city is known to have economically/demographically diverse communities.

  Contributions: To gain insights into the patterns of Internet availability and adoption within the city of Chicago, we employ two existing datasets as primary sources of information: i) the FCC Form 477, which provides data on broadband availability, and ii) demographic data obtained from the US Census American Community Survey (ACS), which offers valuable insights into the characteristics of the population. The study provides a comprehensive analysis of the existing digital inequity in Chicago, and identifies the specific geographical areas that require attention and resources. Additionally, the study demonstrates the relationship between various demographic factors (e.g., age, income, and education) and the adoption of Internet services in the city.

- **Methodology & Tool:** We explore the possibility of utilizing unconventional data sources, like the widely-used micro-blogging platform Twitter, to detect In-

ternet outages [193]. Our focus is primarily on pinpointing instances of network disruptions during natural disasters.

Contributions: In addition to conventional Internet availability data, this dissertation brings a unique perspective by exploring an unconventional source—Twitter data—for understanding and detecting Internet outages. The utility of social media as a vast, real-time information source is well recognized. Still, its potential for aiding Internet outage detection and, by extension, illuminating aspects of Internet quality and access is relatively untapped. We develop machine learning models that can be employed on real-time human experiences shared on Twitter to automatically detect Internet availability issues such as outages. Our methodology and tool have versatile applications, including providing crucial information to first responders about possible areas facing communication disruptions during natural disasters, or identifying regions that generally endure subpar connectivity.

1.2.2 Internet Quality

Due to the scarcity of large-scale datasets offering insights into Internet quality, we initially analyze available datasets to characterize digital inequity along the dimension of Internet quality. Additionally, we devise an innovative methodology that augments these existing Internet-quality datasets with valuable contextual information, thereby enhancing their interpretability and overall applicability.

- **Characterization and Analysis:** The availability of Internet services does not guarantee usability. It is therefore important to assess the extent of digital inequity in terms of Internety quality. Crowdsourced active measurement platforms such as Measurement Lab (M-Lab) Speed Test allows us to explore the relationship between the quality, as opposed to only the availability, of Internet access and demographic

attributes of users of the platform.

Contributions: In this study [196], we use network measurements collected from the users of Speed Test by M-Lab [57] 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 in the state of California. The study demonstrates that tests conducted in urban areas record higher download speeds compared to rural areas. Furthermore, even within urban areas, the study finds digital inequity in terms of Internet quality along the demographic dimension of household income. Finally, the study also contrasts the information reported by the FCC Form 477 with actual network measurements and identifies considerable exaggeration in the Form 477 data, especially in rural locales and urban areas with low income. The study also discusses the potential limitations of existing crowdsourced measurement datasets.

- **Mehodology & Tool:** Although the few existing crowdsourced network measurement, or speed test, datasets can provide insights into trends in Internet quality, they come with several significant constraints. Primarily, these measurements only provide a snapshot of the user's current network condition. For example, if a test reports a download speed of 10 Mbps, it doesn't inform if the low speed was a byproduct of poor Internet quality or other factors that are hyperlocal to the test taker themselves such as their subscription tier and/or poor home network condition. Such contextual information is missing from the existing crowdsourced measurement datasets. To that end, we design a novel methodology that is able to associate crowdsourced measurements with critical contextual information of subscription tier.

  Contributions: To infer the subscription tier of a speed test, we develop the Broad-

band Subscription Tier (BST) methodology which is a two-stage hierarchical unsupervised classification technique [197]. We evaluate the efficacy of BST on the MBA dataset. Results reveal that BST can infer speed test's subscription tier with over **96%** accuracy across multiple ISPs. We then apply BST on existing Internet quality datasets gathered from crowdsourced network measurements platforms such as Ookla's speedtest and M-Lab's Speed Test. We quantify the potential impact of various test taker's specific factors such as access medium (wired or WiFi), WiFi band, WiFi signal strength, and device memory on the ultimate Internet quality reported by the speed tests. Additionally, our results show that the choice of speed test platform can also affect the reported performance. This work is timely given the recent focus on crowdsourced speed test measurements for policy-related decision-making. Our analysis shows how the lack of context contributes to misleading conclusions and offers a set of recommendations for speed test vendors and the FCC to contextualize speed test data and correctly interpret measured performance.

1.2.3 Internet Affordability

As mentioned previously, unlike the dimensions of availability and quality, there exists no comprehensive dataset that allows the characterization of Internet affordability in the US. To tackle this challenge, we develop a tool that can extract the information of ISP provided Internet plans at the street address level granularity around the country. Subsequently, using the data gathered by our tool, we are able to analyze and characterize broadband affordability in the US by focusing on the nature of broadband plans offered by major ISPs.

Contributions: To curate a novel dataset of ISP provided Internet plans at street

address level granularity, we develop the Broadband plan Querying Tool (BQT). BQT takes a street address as input and determines the ISPs that provide services to that address. Subsequently, BQT extracts the set of plans provided by each ISP to that particular address. Using this tool, we have curated a dataset of Internet plans from seven major US ISPs, covering over a million residential addresses in the US [194].

The dataset gathered using BQT brings a revolutionary change to the study of Internet affordability. It allows for a comprehensive examination of trends concerning Internet access costs and the practices of Internet Service Providers (ISPs) nationwide. Analysis of the dataset reveals several interesting findings such as strong spatial clustering of similarly valued plans, benefit of competition between ISPs offering similar value services, and digital redlining amongst various communities. Additionally, this dataset is well suited to verify the accuracy of the FCC's newly developed National Broadband Map [78]. Finally, we make BQT and the dataset publicly available, thereby empowering researchers and policymakers to utilize and build upon our work to deepen our understanding of Internet affordability around the country.

1.3 Broader Impacts

In addition to peer-reviewed publications and presentations to academics, this work was impactful to the larger community.

- We have partnered with multiple state and non-profit entities such as California Public Utilities Commission (CPUC) [112], California Community Foundation (CCF) [111], and Oakland Undivided [119] to assist in providing the datasets to better understand communities who remain disenfranchised from the benefits of the Internet.

- Our work successfully convened stakeholders from academia, industry, and policymaking domains in a dedicated workshop [108], fostering a collaborative environment for discussing broader concerns related to digital inequity. Through this platform, we explored the development of methodologies, tools, and algorithms aimed at enhancing our comprehension of Internet inequity and devising effective strategies for its mitigation. The workshop facilitated meaningful dialogue and knowledge exchange, enabling participants to collectively contribute to the advancement of research, policy, and industry practices in addressing the challenges of digital inequity.

- Through our work, we were able to provide technical assistance to the investigative reporting conducted by the Markup[98], which revealed trends of digital redlining through an examination of ISP-offered services across 43 metropolitan cities in the US. The subsequent news article [73] garnered significant attention from various sectors, facilitating extensive discussion and scrutiny of the US broadband service provisioning sector.

- This work has generated considerable interest in the industry as evidenced by our invited presentations in 2022 to Google's Network Analytics team, as well as to Ookla's Data Science team, where we shared our findings.

- The work that constitutes Chapter 5 of this dissertation received the Distinguished Paper Award (Long Paper) at the 2022 Internet Measurement Conference.

1.4 Dissertation Outline

The remainder of this dissertation is organized as follows. The dissertation is comprised of three main sections: Internet availability, Internet quality, and Internet affordability.

Related to Internet availability, in Chapter 2, using existing datasets, we characterize the current state of Internet availability in Chicago. We propose a novel approach of using data from a popular microblogging platform, Twitter, to detect Internet outages in Chapter 3. In terms of Internet quality, Chapter 4 presents an analysis of Internet quality in California using the data generated from the speed test platform, Speed Test by Measurement Lab. Chapter 5 introduces an innovative approach that links data points from crowdsourced network measurement platforms with the essential context of the subscription tier. The methodology enables a better understanding and interpretation of network performance data and Internet quality in general. Regarding Internet affordability, in Chapter 6, we first present a scalable data collection tool, that enables the curation of the most comprehensive cost of Internet access dataset in the US. Subsequently, we analyze the data collected using the tool to capture different trends pertaining to Internet affordability around the country. Lastly, Chapter 7 discusses our findings and provides some recommendations to various stakeholders to further our understanding of digital inequity.

Part I

Internet Availability

Chapter 2

Internet Inequity in Chicago: Adoption, Affordability, and Availability

2.1 Introduction

Internet access has become critical to ensure equitable opportunity in workforce participation, education, health, and other domains, especially in the post-pandemic world. Historically, the question of Internet equity generally has focused on bridging the "digital divide" between urban and rural areas; that is, on bringing broadband service to the (mostly) rural areas that have none. For example, the recently approved Federal Broadband Equity, Access, and Deployment (BEAD) funding, which promises $42.5 billion to expand high-speed Internet access, with priority given to the connection of "unserved" areas, many of which are rural [189]. More attention, however, is also needed to understand and address Internet equity in areas that already meet the FCC standard of "served", because many residents of these mostly urban areas do not, in fact, have broadband

connections at home. It has long been known that there exist many barriers to Internet adoption, from access to affordability [200]; thus, policy efforts to bridge the digital divide must also focus on areas that are technically "served" by Internet infrastructure, even in large urban centers.

Accurately identifying the urban digital divide and the contributing factors is an important first step towards mitigating it; doing so can subsequently inform the policy interventions that might reduce the divide. As cities around the United States are working towards understanding gaps in Internet equity, the city officials in Chicago asked us for input into how to quantify the digital divide in Chicago. Specifically, they asked us: *What is the Internet connectivity across different neighborhoods in Chicago and how does it relate to socio-economic factors as well as broadband availability?*.

Recent work has pointed towards the existence of a digital divide in urban areas, but an analysis that quantifies this at a neighborhood level is missing. Given that income, unemployment, institutional resources, and social capital are known to be unequally distributed at the neighborhood level—and the influence these characteristics have on both individual and collective outcomes—having an understanding of neighborhood-level inequity in Internet connectivity is paramount [210].

Towards this goal, our work seeks to understand the following questions: (1) What is the current state of Internet inequity in Chicago; specifically which geographies need the most attention and resources?; (2) How does Internet adoption relate to population characteristics including age, income, and education?; and (3) How does broadband availability relate to adoption rates? To answer these questions, we use two datasets in this chapter: (1) the American Community Survey (ACS), a household-level survey containing spatially aggregated information about broadband adoption and key population characteristics (e.g., income, education, occupation), and (2) FCC Form 477 fixed broadband data, a semi-annually collected dataset that indicates availability of ISPs at

a census block level. We combine these datasets to answer the above questions. We find the following:

- Broadband adoption rates vary greatly across neighborhoods in Chicago: adoption rates range from 58% to 93% depending on neighborhood.
- The neighborhoods with the lowest adoption rates are concentrated on the South and West Sides of the city, in majority-Black areas that reflect Chicago's historical patterns of racial residential segregation.
- Adoption rates also correlate with Hispanic population concentration, low income, low educational attainment, and a higher proportion of elderly population.
- Nearly all census blocks (90%) have at least one high-speed broadband ISP present. The number of ISP options available varies greatly by census block, and 50.6% of census blocks have only one high-speed broadband ISP available (as defined by 100/20 Mbps).

Our findings quantify the extent of home Internet inequity in Chicago, highlighting the neighborhoods that would benefit the most from attention towards increasing connectivity rates. While our analysis stops short of establishing causal relationships, the high correlation between income and adoption rate suggests the importance of affordable Internet access either through subsidy programs (such as the United States federal government's Affordable Connectivity Program [154] or the City of Chicago's Chicago Connected program for students in Chicago Public Schools [142]) or measures to support deployment of community networks. The correlation of adoption with age highlights the importance of digital literacy programs such as digital navigators, as well as efforts to innovate on inclusive technologies, including in regards to privacy protection. Finally, the

spatial disparities in ISP availability, including the number of ISP options and newer access technologies such as fiber and DOCSIS 3.1, highlight potential inequities in Internet infrastructure.

2.2 Background and Related Work

In this section, we provide a brief background on the datasets we used in our analysis; we then provide an overview of related work.

2.2.1 Datasets

Census American Community Survey (ACS)

This dataset is a nation-wide demographic survey conducted by the US. Census Bureau. Every year, the Census samples approximately 3.5 million addresses (roughly 1% of the United States population) and gathers population characteristics such as employment, income, and household characteristics. In 2013, the Census also included two questions around broadband adoption inquiring about access to Internet and mode of access The data is made public and is used by governments, communities, and private entities for many purposes (e.g., allocating funds). The Census shares spatially aggregated information from the ACS samples. For analysis in a smaller region, the census recommends using five-year aggregates; for larger geographies, yearly aggregates suffice. This is because of the sampled survey and aggregation across years ensures there are enough responses within a smaller geography. In addition, the Census also shares individual responses through Public Use Microdata Samples (PUMS), which are anonymized to preserve privacy. The PUMS data is available at a larger spatial granularity called Public Use Microdata Areas (PUMAs). In this paper, we use the five-year aggregate

data, unless otherwise specified.

Limitation: The data combines information from last five years to obtain estimates and is thus not always reflective of the most current circumstances. Other limitations include potential errors in the estimates due to sampling and errors in survey response.

FCC Form 477 Data

The Federal Communication Commission (FCC) mandates that Internet Service Providers (ISPs) provide information about areas where they provide service. ISPs need to file their offerings at a census-block level; an ISP can include a census block in its offerings if it can provide Internet service to at least one household in the census block. Along with each census block, the ISPs must also file information about the maximum advertised download and upload speeds in the census block, the access technology (e.g., cable, fiber-to-the-home), and whether the service plan is for consumers or business. This data is one of the key datasets used to decide whether an area is unserved. We use the latest form 477 data to understand broadband availability in Chicago.

Limitation: Form 477 data may overstate broadband availability because: (1) if an ISP only covers a single household in a census block, the entire block is considered covered; (2) "availability" means that the ISP does not necessarily service the area currently but could provide service within an interval that is typical for that type of connection [150]. From a practical perspective, such service delays may act as a barrier to broadband adoption.

2.2.2 Related Work

Previous research has investigated the presence and extent of digital divide/inequality amongst various communities within the US. In [196], the authors utilize the crowd-

sourced speed test measurements from Measurement Lab [56] in California to identify the location and demographic factors that impact download speed. Their results show speed test performance positively correlates with household income and urban areas. Paul et al. conduct statistical analysis on publicly available datasets from another popular speed test vendor [195] (Ookla [191]). The objective of this work was to identify states where digital inequality in terms of internet performance exists between urban/rural and low/high-income areas across all states in the US. Their results once again confirm the presence of digital divide in the majority of the states in the dimensions of location and household income. A similar performance trend between communities using the same dataset was also captured in several previous studies [132, 22].

Other work found that that a major reason behind digital inequity in the urban regions stems from households opting against purchasing high-quality of internet even when it is available [1] . This finding is further reinforced in the survey conducted by Liu et al. [175], who found in a study across 978 US. households that the surveyed population expressed less willingness to invest in Internet speeds exceeding 100 Mbps. Another study conducted in Detroit, Michigan finds while lower-income communities want to purchase high-quality of Internet, the higher associated cost proves to be a major barrier and creates digital divide [205]. An analysis of the urban region of San Antonio, Texas reveals the cost of deployment of new technologies by ISPs and geographical disparities primarily contribute to the digital divide between urban communities [138] .

Combining zip code level demographic information with its own data, Microsoft [35], estimated that adoption of high-speed Internet was lacking for 162.8 million Americans, a number far greater than the FCC's estimate. Galperin et al. identify the low-income minority population as a group likely to be disenfranchised from having access to residential fiber services that provide better Internet performance [158].

2.3 Adoption

In this section, we analyze the broadband adoption rates as reported by the Census ACS data. We consider adoption rates at both the census tract and neighborhood levels.

2.3.1 Method

We use the latest five-year data spanning 2016–2020 from the ACS survey to obtain broadband adoption rates. The survey asks residents the following two questions regarding Internet access:

- *At this house, apartment, or mobile home – do you or any member of this household have access to the Internet?*. The response can be either *no* or *yes*. For households with access, the survey also differentiates whether the access is through a paid subscription or otherwise.

- If the answer to above question is yes with an Internet subscription available, the survey also asks about the mode of Internet access. *Do you or any member of this household have access to Internet using a:* The options include cellular, broadband, satellite, dial-up Internet service, or others.

We consider households who responded yes to having a broadband connection at the household. Note the ACS broadband definition is not the same as the FCC's definition which uses specific speed thresholds. Rather, ACS specifies broadband Internet as high speed Interent (without any speed limits) and provides few examples of access technology which include cable, fiber optic, and even DSL service. We define adoption rate as the percentage of total households that responded yes to having Internet access through a broadband subscription.

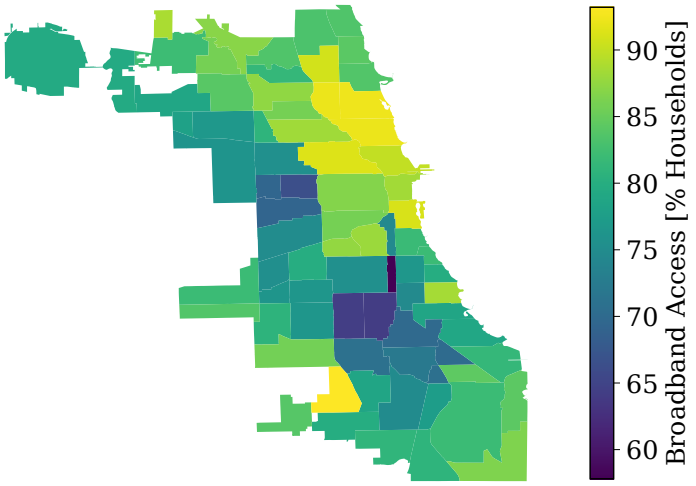


Figure 2.1: Internet Access across Community Areas in Chicago

To calculate adoption rates, we use the data in Table *B28002* containing spatially aggregated responses to the above survey questions. We consider the finest spatial granularity, i.e., census tract. The table contains estimate of number of households with broadband access along with its 90% confidence interval denoted as margin of error (MOE). To compare adoption rates across geographies, we calculate the percentage of households with broadband access by dividing the estimate with the total estimated households in the census tract. We also obtain margin of error for the derived percentage as follows:

$$MOE(\hat{P}) = \frac{1}{\hat{Y}}\sqrt{[MOE(\hat{X})]^2 - (\hat{P} \times [MOE(\hat{Y})])^2} \tag{2.1}$$

Here, $\hat{X}$ and $\hat{Y}$ are the estimated households with broadband access and total households, respectively. $\hat{P}$ is estimated percentage of households with broadband access and is simply

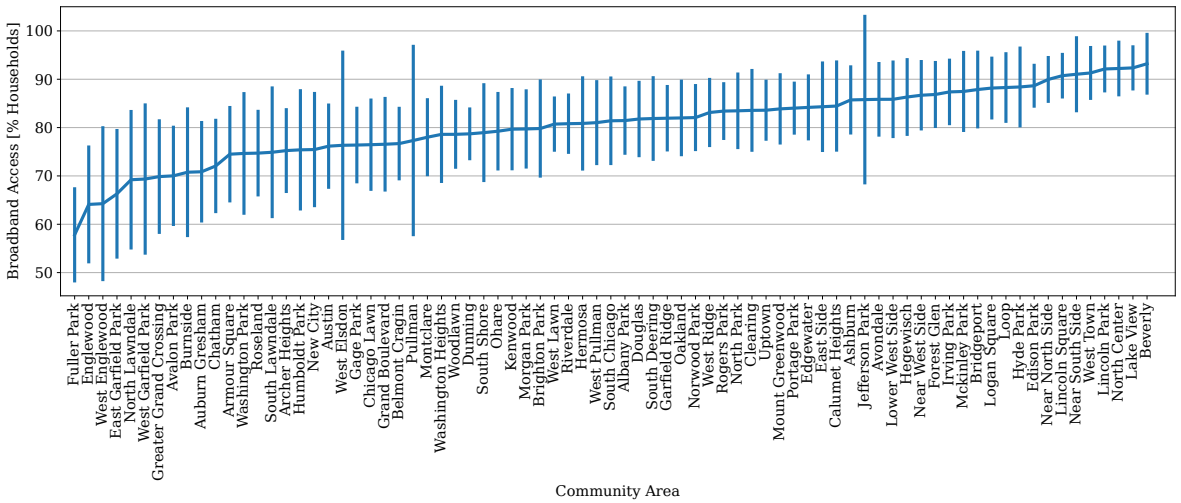


Figure 2.2: Adoption rates along with Margin of Error

$\frac{\hat{X}}{\hat{Y}}$. $MOE(\hat{X})$, $MOE(\hat{Y})$, and $MOE(\hat{P})$ denote the respective margin of errors. In case the expression under the square root is negative, we sum the two expressions under the square root instead of subtracting them, as recommended by the Census [137]. This leads to a more conservative estimation of margin of error.

As Chicago has distinct community areas, we also repeat the analysis at community-area level. We first map each census tract to the respective community area. If a census tract overlaps with multiple community areas, we associate it with the community with which it has the highest-area overlap. We then aggregate the census tract adoption rates within a community to obtain community-level adoption rates. We sum the tract-level estimates to obtain estimates of total households and households without Internet access. To obtain the margin of error, we use the ACS table containing the successive difference replication values [136]. This calculation provides a more accurate margin of error during aggregation.

2.3.2 Broadband adoption across community areas

We first explore the community-level broadband adoption rates. Figure 2.1 shows the fraction of households without Internet access across community areas in Chicago. We observe large disparities in Internet access across community areas. Communities with lowest adoption rates include Fuller Park (58%), Englewood (64%), West Englewood (64%), and East Garfield Park (67%). By comparison, the areas with the highest adoption rates, Beverly and Lake View report 93% and 92% of households with broadband access, respectively. Most of the areas with the lowest adoption rates are located in the South and West sides of Chicago, areas containing neighborhoods that are historically marginalized, consisting of immigrant and lower-income residents who settled away from the central business districts and wealthier, lakeside areas in the northern sections of Chicago [190].

We also analyze the margin of error in adoption rates across community areas as shown in Figure 2.2. The margin of error varies from 4.5% to 19.8% with a median of 8.1%. The margin of error is generally less than 10% of households for almost 90% of the community areas. We found the MOE to be high for three communities, i.e., West Elsdon, Pullman, and Jefferson Park. One reason is that the MOE approximation formula defined in Equation 2.1 could not be applied for these neighborhoods as the expression under the square root was negative. We instead compute a more conservative MOE based on census recommendation which leads to higher MOE. Even considering the margin of errors, the adoption rates are significantly different between community with the highest and lowest adoption rates.

Takeaway : The ACS data provides evidence of stark Internet equity in Chicago. Although the underlying reasons of the divide can be different between urban and rural areas (e.g., affordability vs. connectivity), the data highlights that even urban areas

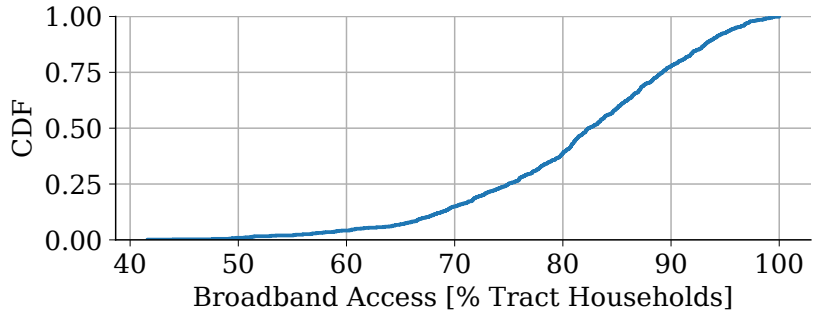


Figure 2.3: Households without Internet Access at tract level.

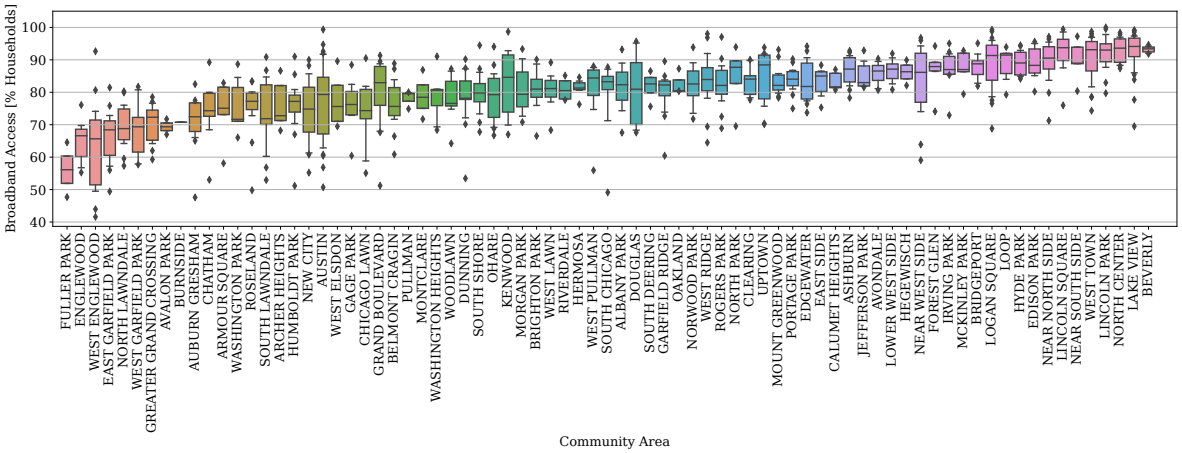


Figure 2.4: Distribution of adoption rates within community areas.

require attention at the local, state, and federal levels—and efforts by communities and governments alike to bridge these gaps.

2.3.3 Broadband adoption at the census-tract level

Although community areas have a social meaning for both residents and local government administrators, there is merit to doing analysis at census-tract levels because census tracts are more fine-grained than community areas. Although a community may have a high adoption rate overall, there may be some smaller regions with lower adoption. Figure 2.3 shows the cumulative distribution function (CDF) of the percentage of tract

	Hispanic	Black	White	Asian
Correlation with broadband adoption	-0.06	-0.49	0.58	0.26

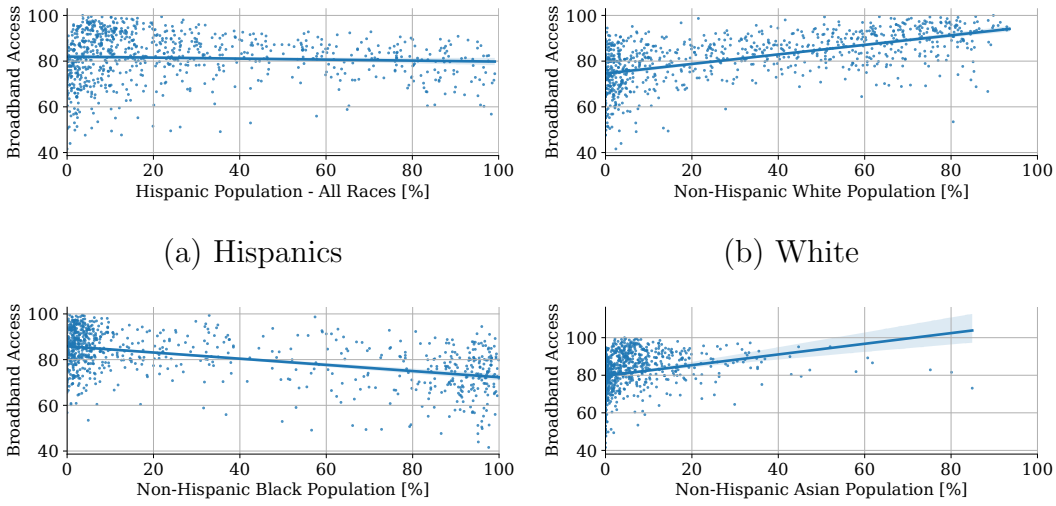
Table 2.1: Correlation of broadband adoption with Race/Ethnicity

households with broadband access across tracts. The adoption rate varies from 42% to 100%; in the median tract, 82% of the households lack broadband access. The analysis shows clear a disparity in adoption rates across census tracts. Next, we consider whether tracts within a community show disparity in adoption rates at a census-tract level. We group tracts based on the community area and show a box plot of distribution of tract adoption rates. We find significant variance within communities. For example, the 10th and 90th percentile tract adoption rates in the *Near West Side* community are 75% and 95%, respectively. This finding indicates that targeted interventions may be required at sub-community area levels.

Takeaway : In addition to disparity at community-area level, there are disparities *within* some communities, indicating community areas are not homogeneous in terms of broadband adoption rates. Thus, more micro-level approaches (e.g., at block level) may be needed to address issues of low adoption within certain individual community areas.

2.4 Correlation with Population Characteristics

In this section, we study how various socioeconomic factors correlate with adoption rates. We first consider the relationship between adoption and race/ethnicity at the level of census tract. Next, we consider three major factors: income, education, and age. These three factors have been shown to correlate with Internet adoption in previous studies [149]. We examine whether these relationships also hold in Chicago.



(a) Hispanics (b) White

(c) Blacks (d) Asians

Figure 2.5: Scatter plot of broadband adoption vs race/ethnicity constitution

2.4.1 Adoption Rates and Race/Ethnicity?

For this analysis, we use the Data Profile Tables from the ACS Census data. The Data Profile Tables or Data Profiles contain a variety of socio-economic and demographic information at a tract level. We select the following estimates at census tract level: (1). **Ethnicity**: percentage of Hispanics of all races (Table $DP05\_0071PE$), (2) **Race:** percentage of Non-Hispanic Blacks (Table $DP05\_0078PE$), Non-Hispanic Whites (Table $DP05\_0077PE$), and Non-Hispanic Asians (Table $DP05\_0080PE$), We do not include other races (e.g., Native Americans, Mixed race) as there is only a small proportion of people who fall into these categories in most tracts. We compute the Pearson correlation coefficient between the race/Hispanic ethnicity percentages and broadband adoption rates across census tracts (see Table 2.1). We find that the percentage of Black residents in a tract correlates negatively with broadband adoption rates, while the percentage of White and Asian residents has a positive (though weaker) correlation with adoption rates. The correlation between percent Hispanic residents and broadband adoption is negative but

small.

We also show a scatterplot of tract racial composition and adoption rates along with a trend line obtained by fitting a linear regression (see Figure 2.5). Of particular interest is Figure 2.5c, showing the adoption rates and percentage of Black population in a census tract. This graph recalls the high levels of residential segregation between Black and non-Black populations in Chicago, and shows that broadband adoption rates are similarly divided. Most tracts have either few Black residents or a majority of Black residents. In the latter tracts, we find low adoption rates. In contrast, tracts with few or no Black residents have high adoption rates. This finding is corroborated by Figure 2.5b, which shows high adoption rates in tracts with a majority White population.

2.4.2 Adoption Rate and Population Characteristics

We next study the correlation between adoption rates and key population characteristics. We focus on three characteristics: income, education, and age. These three characteristics may have a causal effect on adoption. Income can affect the ability to get a broadband connection; education has an effect on both employability (and hence income) as well as digital literacy and hence the perceived utility of the Internet; age may also affect broadband adoption with older population less likely to adopt due to multiple reasons such as difficulty of using digital technology, privacy concerns, or perceived lack of utility. With the available datasets, demonstrating causality is challenging; thus, we restrict our analysis to correlation.

For this analysis, we again use the Data Profile Tables from the ACS Census data; we use the following tables:

- **Income**: We select two metrics for income. The median household income (Table DP03_0062E) and percentage of families below poverty level (Table DP03_0119PE).

Population Characteristic	Correlation Coefficient
Log median household income	0.72
Percentage of families below poverty level	-0.58
Percentage of population above 25 with Bachelor Degree or high	0.66
Percentage of population above 25 with a high school degree or higher	0.52
Percentage of population above 65	-0.38
Percentage of single-person householders above 65	-0.52

Table 2.2: Correlation of population characteristics with broadband adoption

The latter metric normalizes the income by the number of people in the household. We further take the log of the median household income as existing work suggests using log income for modeling [148].

- **Education**: We consider two metrics for education, namely percentage of population above 25 with a high school degree or higher (Table DP02_0015PE) and percentage of population above 25 with a bachelors degree or higher Table DP02_0068PE).
- **Age**: We consider the percentage of older adults (above 65) in the population. We consider two metrics, percentage of population above 65 (DP05_0024PE) and percentage of single-person households with the householder age above 65 (Table DP02_0009E and DP02_0013E for males and females, respectively). The second metric can more strongly show the association between age and broadband adoption. For the second metric, we obtain a single number by adding the male and female householders.

We obtain the above metrics at the census-tract level and compute the Pearson's correlation coefficient with broadband adoption rates (see Table 2.2. Looking first at metrics of income, we find that broadband adoption rate is positively correlated (in fact most correlated) with median log income and negative correlation with percentage of families with income below the poverty level. This result is expected since low income may make it difficult to purchase broadband Internet access. The prices of popular ISPs like Comcast and AT&T do not vary based on geography within Chicago; as such, we can

compare broadband affordability by comparing income across neighborhoods. High per-capita income tracts may generally find broadband to be more affordable and thus they also may have high broadband adoption rates.

We also find a high correlation between broadband adoption rates and education. Broadband adoption rates are more correlated at the tract level with percentage of people with a bachelors degree than with percentage of people with a high-school degree. Education is also highly correlated with income. As mentioned before, education can impact adoption through income as well as through households' perceived utility of Internet. Future work can consider isolating the impact of the latter by controlling for income.

Finally, we find a weak negative correlation (-0.17) between adoption rates and percentage of population above 65 in a tract. The negative correlation is stronger (-0.42) when we consider single person households with householder above age 65. This relationship suggests that the older population may have lower adoption rates due to low perceived utility of the Internet, may also be related to the difficulty of using technology. Increasing high-speed broadband adoption rates could, however, be critical for these households (e.g., with remote telehealth opportunities during and after the COVID-19 pandemic).

Takeaway: When examining bivariate correlations, we find that broadband adoption in Chicago is most correlated with income and education. In terms of policy, the association suggests the need to make broadband more affordable for the lower-income population (which is also more likely to comprised people of color) in Chicago. We also find a negative correlation between adoption and percentage of single-person households above 65 years of age. Community-based programs such as digital navigators, which aim to enhance digital literacy, may be useful for increasing the adoption rate, especially among

Access Type	Technology	ISP	Advertised speeds [Mbps]	% Census Blocks Present
Wireless	Satellite	ViaSat	35/3	100
		HNS	25/3	99.71
		VSAT	2/1.3	99.71
	Terrestrial	T-Mobile	25/3	44.2
		Verizon	300/50	0.16
		Google Webpass	100/100, 200/200, 500/500, 1000/1000	0.4
		Everywhere Wireless	25/10,..,2000/2000	0.75
Wired	ADSL, ADSL2, ADSL2+, VDSL, SDSL, Fiber	AT&T	0.42-0.42, .., 1000/1000	89.59
	DOCSIS 3.1	Comcast	1000/35, 2000/2000	88.62
	DOCSIS 3.0	WOW	1000/50	20.23
	DOCSIS 3.1, 3.0, 2.0, 1.1, 1.0, Fiber	RCN	25/4, 500/20, 1000/20, 1000/1000	12.59

Table 2.3: ISPs in Chicago by access technology.

the elderly.

2.5 Availability

In this section, we analyze broadband availability in Chicago. We specifically consider variability of technology, speeds, and number of ISP options within Chicago. We use the latest FCC form 477 data from December 2020 which contains ISP-provided availability information at a census block level.

2.5.1 ISP Availability by Access Technology

We filter the form 477 data for census blocks within Chicago and characterize the ISPs based on the access technology. Table 2.3 shows the major ISPs, including their access technology, advertised speeds, and percentage of census blocks covered.

Satellite wireless : These include ISPs that use satellites (e.g., Low Earth Orbit Satellites) to provide Internet access. Consumers can obtain Internet access by installing a satellite receiver antenna. The satellite ISPs are characterized by low speeds and high

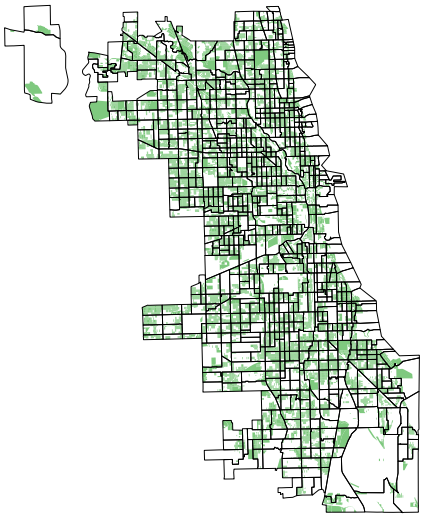


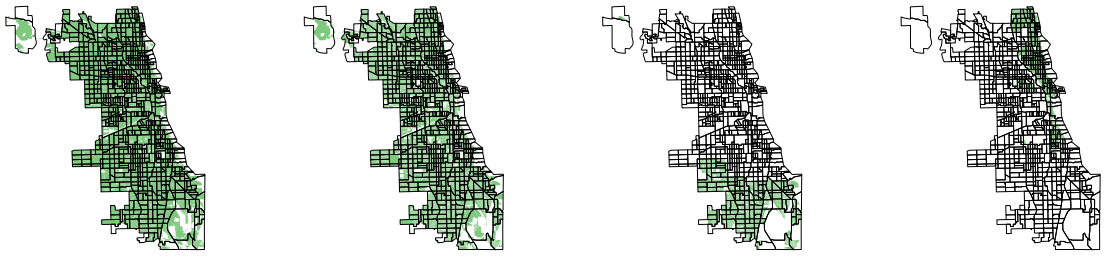
Figure 2.6: Spatial coverage of T-Mobile fixed wireless offerings.

last-mile latency compared to the wired ISPs. These satellites are mostly suited for rural or remote contexts where it is challenging or not profitable for wired ISPs to provide access. We find three satellite providers with residential offerings in Chicago. All three ISPs span nearly all of the census blocks. However, we do not include satellite providers in our analysis because of two reasons: (1) Internet speeds from these ISPs are typically slow. Two ISPs provide sub-broadband speeds (less than 25/3 Mbps) and the other two provide speeds of 25/3 and 35/3 Mbps. (2) The plans are expensive compared to fixed broadband plans. For instance, the least expensive broadband plan from one of the ISPs is 99$/month. As a result, satellite networks are likely not feasible options for broadband users in Chicago which is mostly urban with less expensive, high-speed, and low-latency terrestrial connectivity options.

Terrestrial wireless : These ISPs provide access using a terrestrial wireless system. Typically, most of these ISPs have a last-mile wireless link with the upstream links being wired. For instance, ISPs that provide fixed broadband using cellular technology would

be categorized as a terrestrial wireless ISP. We find 12 fixed terrestrial wireless providers with residential plans in least one census block within Chicago. Among these, 8 provide offerings in fewer than 30 census blocks with 6 serving only up to two census blocks. We exclude these providers from our analysis. We examine the speeds and coverage of the remaining four ISPs (see Table 2.3). Two providers, T-Mobile and Verizon, are cellular, and provide home Internet using 5G or LTE technology. T-Mobile is the largest terrestrial wireless provider, covering 20,498 census blocks. In terms of speed offerings, it reports only a single speed tier of 25/3 Mbps, which barely meets the FCC broadband standard. Verizon, on the other hand, reports speed offerings of up to 300/50 Mbps but covers only 0.16% of the census blocks. The other two providers are Everywhere Wireless and Google (doing business as Webpass) with presence in 0.4% and 0.75% census blocks. Both of these providers report up to gigabit symmetric speed offerings. This seems surprising given these are fixed wireless ISPs. However, based on the description online, these ISPs likely use fiber in most of their network with a single wireless hop in the last- mile. For instance, Webpass uses a rooftop antenna to receive wireless Internet at the building. It likely uses the Google Fiber infrastructure for upstream connection.

We next examine the spatial coverage of these ISPs. As shown in Figure 2.6, T-Mobile service seems to be evenly distributed across Chicago. In comparison, the other three ISPs have sparse coverage, with offerings mostly in the northern neighborhoods and business district areas of Chicago. This characteristic may result from some of the following factors: (1). these companies have fiber-based infrastructure in these regions, (2) these areas generally have high-occupancy buildings, with occupants having relatively higher income, thus providing better potential return on investment, especially for Google Webpass and Everywhere Wireless.



(a) AT&T (b) Comcast (c) WOW (d) RCN

Figure 2.7: Spatial coverage of fixed wired ISPs. Green indicates availability

Technology	**% census blocks present**	**Download speed**	**Upload speed**
ADSL2,ADSL2+	86.2%	0.77-25	0.38-3
ADSL	17.3%	0.77-6	0.26-0.51
SDSL	0.0%	0.42	0.42
VDSL	58.0%	18-100	1.5-20
Fiber	23.7%	1000	1000

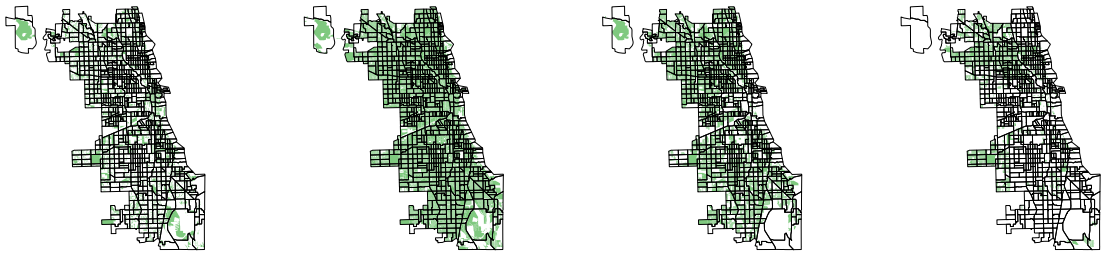
Table 2.4: AT&T: Advertised speeds and coverage [% of census blocks] by technology

Wired: Twelve wired ISPs providing residential offerings in at least one census block in Chicago. Among them, eight ISPs are present in fewer than 25 blocks, with 5 ISPs serving only up to 2 census blocks. We focus on the remaining four ISPs which have a significant footprint in Chicago, also summarized in Table 2.3. Figure 2.7 shows the spatial coverage of these providers across Chicago. AT&T and Comcast are the largest wired providers with presence in more than 88% of census blocks in Chicago. The census blocks that are not covered are likely the blocks with zero residential population. In future, we plan to validate this using once the data from 2020 Census becomes public. The other two ISPs have a limited footprint. WOW provides offerings in far south side of Chicago while RCN has offerings in the Downtown and Northern lakeside areas of Chicago.

In terms of access technology, three ISPs— Comcast, RCN, and WOW—mostly use

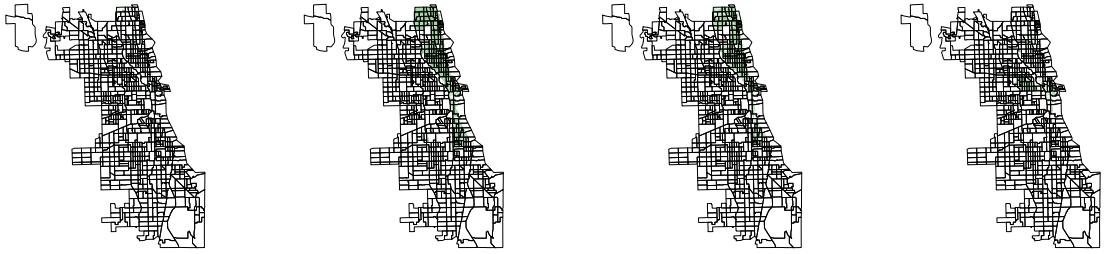
DOCSIS to provide Internet connectivity over hybrid-fiber-coaxial (HFC) networks. Each version of DOCSIS varies in channel configurations and throughput for upstream and downstream links. DOCSIS 3.1, the latest standard, can support up to 10 Gbps downstream and 1 Gbps upstream throughputs. Comcast uniformly supports DOCSIS 3.1, the latest DOCSIS standard, across all census blocks. This also reflects in the Comcast's speed offerings, as it uniformly reports maximum advertised speed of 1000/35 Mbps across all census blocks. In comparison, WOW supports an older cable standard, DOCSIS 3.0, across all census blocks. However, in terms of advertised speeds, WOW reports the same maximum advertised speed of 1000/50 Mbps across all census blocks. The advertised upload speeds are higher compared to Comcast, despite supporting an older cable standard. Finally, RCN reports different version of DOCSIS technologies in different census blocks. Among the 8,495 census blocks it serves by cable, 37% support DOCSIS 3.1, 60% support DOCSIS 3.0, and 3% support older version of DOCSIS (i.e., 1.0, 1.1 or 2.0). We also find difference in advertised speeds across the three standards, with the speeds being 1000/20 Mbps, 500/20 Mbps, and 25/4 Mbps for DOCSIS 3.1, 3.0, and 2.0 or older versions, respectively. RCN also provides access using fiber in 1044 census blocks with maximum advertised speeds of 1000/1000 Mbps. Figure 2.9 shows the spatial map of RCN offerings coded by the access technology. Most Fiber offerings are centered around a single region, slightly north-west of the downtown Chicago.

The fourth major fixed wired ISP, i.e. AT&T, uses a mix of Digital Subscriber Line (DSL)-based and Fiber as the access technology. Among the 41532 census blocks served, it supports Asymmetric DSL (ADSL) 2 and ADSL2+ in 96.2% blocks, Very high speed DSL or VDSL in 64.7% blocks, Fiber to the home or fiber in 26.4% blocks, ADSL in 19.3% blocks, and Symmetric DSL in 0.02% blocks. Figure 2.8 shows the spatial distribution of the different access technologies. There is no clear pattern in the spatial distribution of DSL technology. The fiber, however, is mostly concentrated in the West



(a) ADSL (b) ADSL 2, 2+ (c) VDSL (d) Fiber

Figure 2.8: AT&T: Spatial coverage of different Access technologies



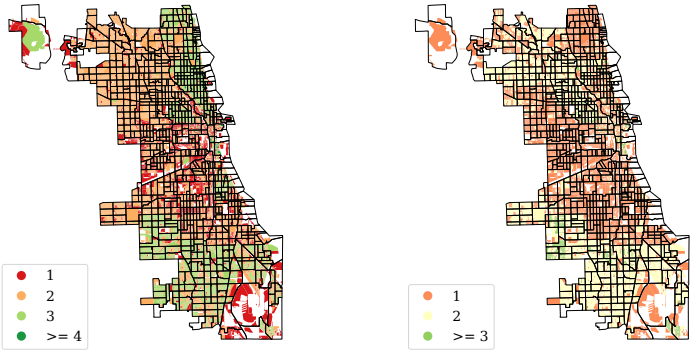
(a) DOCSIS 1.0, 2.0 (b) DOCSIS 3.0 (c) DOCSIS 3.1 (d) Fiber

Figure 2.9: RCN: Spatial coverage of different Access technologies

and North western parts of Chicago. In terms of speed, the maximum advertised speed over Fiber is 1000/1000 Mbps across all census blocks. The speeds over the same DSL technology varies across census blocks. This is likely because DSL performance depends on the distance between the subscriber and Central Office (CO). Table 2.4 summarizes the different advertised download and upload speed pairs for different DSL technologies.

2.5.2 Number of ISPs in a Single Census Block

We next study the number of ISPs that are available in a census block. Note that having multiple ISPs in a census block does not necessarily imply more competition as each ISP may serve a disjoint set of addresses. It can be considered as a necessary but not sufficient condition to indicate competition. We consider two speed thresholds while counting



(a) 25/3 Mbps (b) 100/20 Mbps

Figure 2.10: Number of ISPs available across census block

the number of ISPs, (1) 25/3 Mbps, the speeds used by the FCC to define broadband, and (2) 100/20 Mbps, the newly proposed minimum speeds for broadband [151]. Figure 2.10 shows the number of ISPs available within a census block based on different speed thresholds. We see a disparity in the number of options available across different regions. While, the lakeside areas and the far south parts have 3 broadband options available (25.8% blocks), the remaining areas have only two options (59% blocks) available with some pockets (14.1% blocks) having only one broadband ISP option. For high-speed broadband options, 50.6% blocks have only one ISP option, 45.9% have two options, and the remaining 3.5% blocks having 3 or more options.

Are adoption rates correlated with number of ISPs? We now analyze if adoption rates are correlated with number of ISPs available in an area. We first calculate the number of ISPs with broadband offerings (greaten than 25/3 Mbps) available in each census block. The ACS data, however, is available at census tract level. To match the two datasets, we compute the average of number of ISPs available across census blocks in a census tract. We then compute the pearson correlation coefficient between average number of ISPs and adoption rates across census tracts. We observe a weak

positive correlation of 0.27 between availability and adoption. It is not clear however if there is a causal relationship. There could be other confounding factors such as legacy infrastructure and higher population density that may impact availability in an area.

Takeaway: AT&T and Comcast are available evenly across census blocks in Chicago. AT&T, however, varies in terms of access speeds due to their dependence on DSL technology in certain areas. Their fiber offerings are also available in selected regions. Among the terrestrial wireless ISPs, only T-Mobile has a city-wide presence. The remaining five ISPs (two wired and three wireless) have a more limited footprint. Four of these ISPs are concentrated in North Lakeside and Downtown areas of Chicago, with only one ISP providing sevice to the far south side of Chicago.

2.6 Conclusion

In this chapter, we analyzed Internet equity in Chicago across three dimensions: adoption, affordability, and availability. We find disparity in adoption rates across community areas in Chicago. The areas with lowest connectivity also exhibit low income, thus indicating that adoption may result from a lack of affordability, although future work could aim to firmly establish this causal relationship. In addition, we also observe low adoption in households with elderly populations or low education, possibly indicating issues related to perceived utility of broadband potentially impacting adoption. Finally, our broadband availability analysis using FCC form 477 data shows that most regions in Chicago have at least one broadband (and even high-speed broadband) option, yet different regions do exhibit variability in terms of the number of ISP options that are available.

Chapter 3

#Outage: Detecting Power and Communication Outages from Social Networks

3.1 Introduction

Users post content on social media platforms such as Twitter, Reddit and Facebook for a variety of purposes, including to report real-time situational incidents such as loss of electricity, internet connectivity and telecommunications [209]. During the onset of a natural disaster, situational information is posted by the affected individuals in real-time, including, increasingly, cries for assistance when 911 lines are overloaded [163]. First responders are responsible for carrying out rescue operations to help affected people during such emergency situations. Real-time social media posts can therefore provide critical information about the situation on the ground so that first responders can be most effective. Researchers have previously analyzed the usefulness of online information in timely crisis response and management [165, 171]. A key challenge is to extract

valuable and actionable information such as missing or injured people and damaged utilities and infrastructure from all other content that appears online. It is therefore critical to develop information extraction tools that are capable of cutting through the noise and quickly filtering out vital information that authorities can use in their search and rescue operations.

Twitter has emerged as an ideal platform for information retrieval tasks due to the concise nature of the posts (tweets) [125]. Crisis informatics researchers have studied how to identify different types of sub-events, such as loss of lives and damage to infrastructure, from user generated posts [208, 166]. However, most of the developed algorithms focus on extracting information related to a wide spectrum of events, rather than a specific type of event [212, 216]. Since every type of event is not equally tweeted about by the users, some categories are classified with poor precision and recall as they represent only a small percentage of the entire dataset [188]. Additionally, in a recent study [225], it was reported that the majority of the existing frameworks that aim to provide situational awareness to responders during a crisis do not meet the immediate informational requirements of specific responders. For example, information related to power outage would be more useful to responders responsible for restoring damaged utilities than responders in charge of locating trapped people. As such, there is an urgent need to develop highly domain specific information extraction tools to properly assist responders during emergencies.

In this work, we study the viability of the use of tweets to detect power and communication outages during natural disasters, with a specific focus on hurricanes. We begin by collecting tweets based on carefully selected keywords, and subsequently curate a *raw* dataset. We label a sample from the raw dataset to generate an *annotated* dataset that contains tweets related to power-outage, communication-outage and power-communication outage. Our goal is to first analyse characteristics, such as commonly used words, hashtags and sentiment, associated with the tweets that convey outage-related in-

formation during natural disasters. Then, we evaluate the performance of simple machine learning algorithms, neural network and transfer learning models to create a classification framework that is capable of determining whether or not a tweet is about outages. Once identified as an outage-related tweet, we perform an information extraction task to filter further information such as whether the tweet is about a power outage, a communication outage or both. To the best of our knowledge, no previous study has analyzed Twitter data in-depth to perform information extraction to detect both power and communication outages.

While prior work has shown that people often use Twitter as a platform to report power and communication outages [128], our study observes that over 75% of the tweets that contain outage-related keywords do not mention an actual outage. Hence, it is not enough to simply filter tweets based on keywords as this results in noisy dataset. Among tweets that actually mention an outage, we determine that the majority of tweets are made about power outages, followed by communication outages, and then both power and communication outages. Our analysis reveals that actual outage-related tweets carry more negative sentiment than tweets that contain outage-related keywords but that do not actually report an outage. As we attempt to classify these tweets, we are faced with the challenge of low numbers of usable tweets, as well as the inherent noise that is present in data gathered from Twitter. In spite of that, we observe that simpler models such as boosting and support vector machine are able to identify tweets that contain outage-related words with close to 100% accuracy. Furthermore, by applying state-of-the-art text classification techniques such as transfer learning, we are able to identify tweets that not only contain these keywords, but specifically report power and communication outages with very high accuracy, precision and recall scores. In summary, this work presents the following contributions:

- We curate a dataset of 18,097 unique tweets containing outage-related keywords, posted during seven major hurricanes that made landfall in the USA between 2012 and 2018.
- We present an in-depth analysis to determine features such as commonly used words, hashtags and sentiments associated with the tweets that mention power and communication outages.
- We use machine learning algorithms to perform multiple levels of information extraction to detect tweets that contain information about power and communication outages.
- We show that using simpler models such as SVM, tweets that contain outage-related keywords can be quickly detected with very high accuracy. Furthermore, employing transfer learning models such as BERT, we show that different types of outage-related events can be identified with high precision and recall scores.

3.2 Related Work

Information extraction from textual data is a very popular application of natural language processing. Previous work has been conducted to detect power outages using tweets as a source of information. Researchers have also focused on using data from social media to detect other types of events during natural disasters. In this section, we present the related work in two categories.

3.2.1 Power Outage Detection from Tweets

There has been some work that focused on detecting power outages using posts available on Twitter. In [169], the authors gathered a dataset and applied several machine learning

algorithms to detect power outages from tweets. Their analysis showed that a multi-layer perceptron model is capable of detecting tweets related to power outages with reasonable accuracy, precision and recall. The authors in [128] used active learning, standard learning and Kleinberg's burst to detect real-time power outages using tweets. Supervised topic modelling was employed in [218] to detect power outages from tweets. Nightlight satellite imagery and tweets were used in [164] to identify locations of power outages. Specific keywords were used in [129] to gather a datset and then use classification algorithms to detect whether a tweet is about a power outage.

The primary focus of these studies was to make the binary distinction of whether or not a tweet refers to a power outage. Further, while these studies detect power outages from tweets using machine learning algorithms, they each utilized datasets that contained equal numbers of outage-related and unrelated tweets. In contrast, in this work, we maintain the ratio of tweets in each category that we observe during the analysis of our raw datset. Critically, in our work, we only consider a tweet to be relevant to an outage if it mentions an actual outage and not simply if it contains outage-related keywords. Finally, in addition to identifying power outages, we also carry out detailed analysis to identify tweets that mention communication outages. To the best of our knowledge, our work is the first to perform detection of tweets that identify actual power and communication outages as well as to discern tweets that identify outages from tweets that simply contain outage-related words.

3.2.2 Sub-event Detection from Tweets

Prior work has attempted to identify information from social media during crisis scenarios [171, 204]. In [126], the authors attempt to use tweets to identify users in need of resources during or post natural disaster and match them with others who claim to

have the needed resources. In [188], deep neural networks are used to identify useful tweets during crisis situations and categorize useful tweets. Tweets related to damaged infrastructure and utilities formed 8% of their dataset . The authors of [221] applied matching and learning based methods to identify tweets that provide situational awareness during natural disasters. In [206], the authors used integer linear programming to identify several types of events from tweets made during some natural disasters. In addition to natural disasters, sub-event detection from tweets has been performed in other fields. A recent paper [162] used keyword volume to identify specific events that belong to a category such as protests. The authors in [207] used tweets to gather information during epidemics. In each of these studies, different machine learning frameworks were employed to extract/classify information from a large number of data points gathered from Twitter. However, none of this work employed multiple levels of classification to obtain fine-grained information about power and communication outages.

3.3 Data and Annotation

To achieve our goal of identification and classification of outage-related tweets, we first curate a raw dataset using specific keywords. Once we obtain the raw dataset, we manually perform annotation to generate an annotated dataset for detailed analysis and classification. Figure 3.1 presents our overall framework for this study.

In this section, we describe our process to collect the raw dataset. We then explain the annotation procedure used to generate the annotated datset.

3.3.1 Dataset Curation

The volume of tweets related to infrastructure damage increases during natural disasters [167]. Unlike other natural disasters such as earthquakes that occur within a short

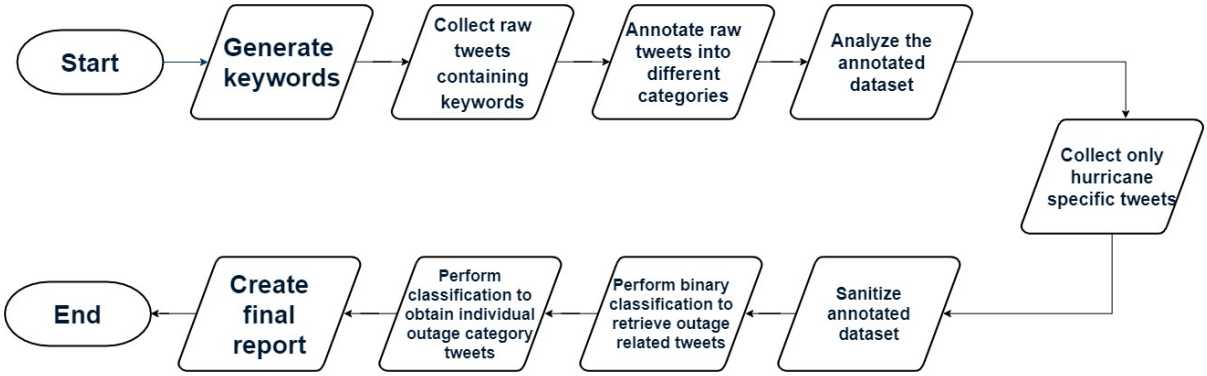


Figure 3.1: Proposed framework to detect power and/or communication outages from Tweets.

span of time, hurricanes pass through an area over a much longer period, typically hours or days. As such, to curate our dataset, we selected the seven major hurricanes that made landfall in the USA between 2012 and 2018. We used Crimson Hexagon [10], a social media firehose with access to 100% of the Twitter stream, to collect tweets that appeared online in the time period from when each of these hurricanes made landfall to when they dissipated [15]. To collect tweets of interest, we generated two sets of keywords: Hurricane-specific keywords and Outage-specific keywords.

Hurricane-specific keywords: Similar to [206], to obtain hurricane-specific tweets, we filtered tweets using keywords such as, but not limited to, HurricaneMaria, harvey storm, hurricanematthew and #HurricaneSandy. In all cases, our keywords contained either the word "hurricane" or "storm", as well as the name of the storm. This resulted in a total of thirteen keywords per storm (91 total for the seven storms), each a permutation of these words and name combinations with different capitalization (i.e. we used each of michael storm, Michaelstorm, and MichaelStorm as a keyword). These formed our hurricane-specific keywords that we used to identify tweets related to these natural disasters.

Outage-specific keywords: In order to generate keywords to obtain tweets related to power and communication outages, we employed the semi-supervised topic modelling algorithm Latent Dirichlet Allocation (LDA) [168]. We began by scraping news articles

that mention power and/or communication outages and formed a document containing the keywords mentioned in those articles. These keywords from the articles were obtained using the keywords class of the Newspaper3k [16] library provided by Python.To generate a diverse set of keywords, we applied LDA, with various combinations of numbers of topics and keywords, to this document. Five sets of topics, each having 15 keywords were heuristically determined to generate keywords of desired quality. Upon completion, we manually selected the words that we considered to be most relevant to obtain the required tweets. For example, keywords such as *blackout*, *outage*, *spotty*, *reception* and *damage* emerged from LDA as related keywords. We also added joined keywords such as *no power*, *can't call* and *call drop* to retrieve relevant tweets. Furthermore, to improve the quality of data, we collected tweets that had geo-location information and originated in the specific areas at the time the hurricanes passed through. The areas were determined from [15] and the geolocated tweets were collected using the location feature provided by Crimson Hexagon.

Table 3.1 presents the number of tweets that contained only hurricane-specific keywords as well as outage-specific keywords. The query containing outage-specific keywords also contained the hurricane-specific keywords for each hurricane. Among all tweets that have one or more hurricane-related keyword, only about $1-4$ percent of those also contain outage-related keywords. We observed that the overall number of tweets that contain tagged geolocation is on average 10 times less than the un-tagged tweets. Interestingly, the percentage of geo-tagged tweets that contain outage specific keywords in the total set is larger than those present in the un-tagged tweets.

Among the hurricanes, Hurricane Sandy contained the greatest number of tweets with outage-specific keywords by volume. This hurricane caused over 8 million people to lose power, far greater than any other hurricane we studied [160]. Hurricane Maria also caused extensive power outages, leaving over $80,000$ households without power [11]. In

Table 3.1: Number of tweets generated during hurricanes that contain keywords with and without geo-location.

Hurricane	Tweet extraction period	Keywords	Non-Geo-Tagged	Geo-tagged
Michael	10/06/2018-10/17/2018	hurricane-specific	387,617	62,191
		outage-specific	15,909	3,300
Florence	08/30/2018 -09/20/2018	hurricane-specific	718,414	69,262
		outage-specific	25,155	3,231
Maria	09/15/2017-10/03/2017	hurricane-specific	483,195	34,740
		outage-specific	26,509	1,594
Irma	08/29/2017-09/14/2017	hurricane-specific	1,761,869	252,082
		outage-specific	58,102	13,944
Harvey	08/16/2017-09/03/2017	hurricane-specific	1,372,863	193,965
		outage-specific	18,643	4,141
Matthew	09/28/2016-10/11/2016	hurricane-specific	1,202,774	175,941
		outage-specific	35,367	6,841
Sandy	10/22/2012-11/02/2012	hurricane-specific	1,903,552	250,936
		outage-specific	75,349	14,209

Table 3.2: Number of tweets per hurricane in the dataset.

Hurricane	Total Number of tweets selected
Michael	3,005
Florence	2,742
Maria	2,597
Irma	3,136
Harvey	1,208
Matthew	2,209
Sandy	3,200

terms of communication outages, Hurricane Maria destroyed over 88% of the cell sites in Puerto Rico alone [152]. In comparison, cell phone infrastructure experienced less damage during Hurricane Sandy [4]. This could explain the larger number of outage-related tweets that appeared online during Hurricane Sandy than during Hurricane Maria; when faced with both cellular and power outages, many residents of Puerto Rico probably found themselves unable to post on Twitter. Hurricane Michael, on the other hand, had the fewest outage-related tweets, possibly because it also had the shortest duration among the hurricanes. In terms of percentage of outage-related tweets (percentage of tweets that contained outage keywords among all hurricane related tweets), Hurricane Maria contained the greatest number. When comparing geo-tagged tweets, we notice that Hurricane Sandy contained the greatest number of outage-related tweets, both by

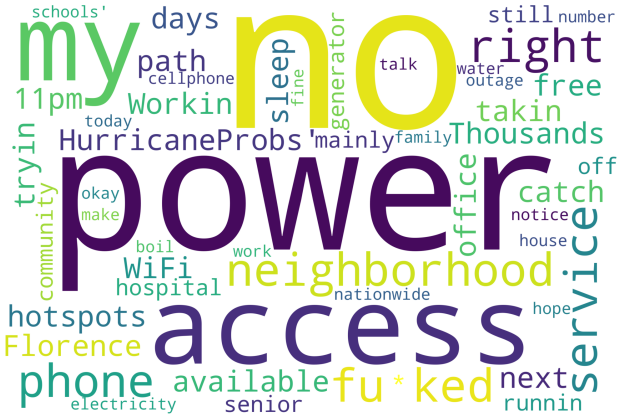


Figure 3.2: The salient words associated with power and communication outage tweets. A larger font for a word signifies high frequency of occurrence of that word in the dataset.

volume and percentage.

To curate our dataset, a sub-sample of the raw tweets that contained one or more of our outage-related keywords from each hurricane was selected. The sub-sampling strategy involved selecting a greater number of tweets that originated from the locations where the hurricanes made landfall. To ensure that our dataset was not dependent on one particular hurricane event, we incorporated roughly equal numbers of tweets from each hurricane. The smallest number of samples were drawn from Hurricane Harvey and Hurricane Matthew as they contained the smallest percentage of outage-related tweets (both geo-tagged and un-tagged) in their datasets. Table 3.2 presents the total number of tweets selected from each hurricane. The gathered dataset consisted of $18,097$ outage-related tweets. The salient words[1] present in the tweets is shown in Figure 3.2.

3.3.2 Dataset Annotation

To identify the different categories of tweets present in our raw dataset, we proceeded to annotate the dataset. We first attempted to perform the annotation process using Amazon Mechanical Turk (AMT) [133]. However, the annotated results obtained from

[1]inappropriate language has been modified with the '*' character

AMT were unreliable; they contained many incorrectly labeled tweets, and in many cases multiple annotators labeled the same tweet differently. We therefore discarded these annotations. The annotation was then instead performed by 80 closely supervised volunteer upper division computer science students using Labelbox [14]. A pair of students were assigned the same subset of the *raw* dataset. The labels for the data points that did not match were further annotated by one of the authors. The annotators were provided detailed guidelines and asked to tag each tweet into one of the following four categories:

Not relevant: A large number of tweets in the raw dataset contained outage-related keywords but did not convey actionable outage-related information. For example, many tweets mention losing power in the future and thus do not provide any actionable information about current outages. As such, any tweets that do not contain current outage information are categorized as Not Relevant.

Power-outage: This category of tweets was reserved for tweets that mention power outages. In addition to directly reporting an outage, many tweets were informational in nature. They either contained a first-hand account by a person about a power outage in an area, or they contained a news article with information about areas currently experiencing an outage. Tweets that contained information about power restoration after a period of outage were also included in this category.

Communication-outage: Similar to the power-outage category, the category of communication-outage represents tweets that report communication outages. This category also consists of tweets that provided information about a related outage in an area/locality as well as tweets that reported regaining communication facilities after an outage.

Power-Communication-outage: We observed a small number of tweets that mentioned both power and communication outages, and placed those tweets in this category. Note that tweets in this category do not necessarily indicate that both power and communi-

Category	Example Tweet
Not Relevant	Hurricane Sandy - please don't take out my power and wifi.
Power-outage	Day 2 of no power..Thanks for everyone's concern.#HurricaneMatthew
Communication-outage	Internet just went out #hurricaneirma
Power-communication-outage	No access to my neighborhood right now no power and no phone service #f*cked #HurricaneSandy #HurricaneProbs

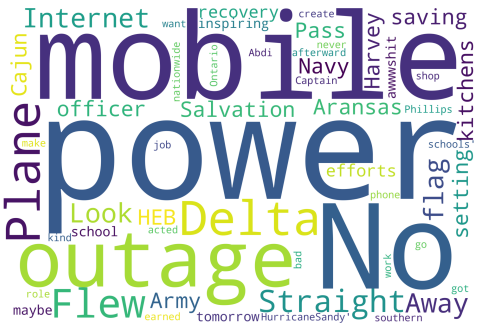
Figure 3.3: Example tweet per category.

Table 3.3: Number of annotated tweets per category.

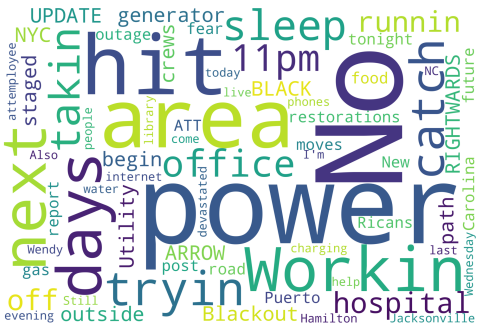
Category	Number of Tweets
Not Relevant	13,957
Power-outage	2,791
Communication-outage	1,000
Power-communication-outage	349

cation are out; instead, they provide information about the status of both utlity types.

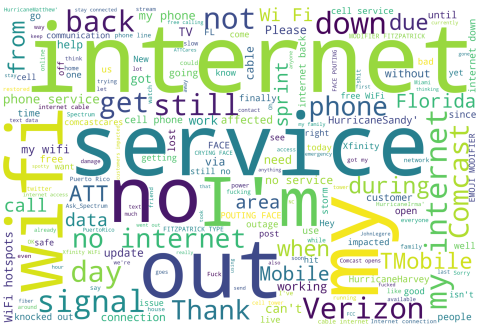
Figure 3.3 shows an example tweet from each category; the total number of annotated tweets per category is presented in Table 3.3. Surprisingly, a large portion of the tweets belong to the Not Relevant class even though the tweets were carefully extracted using domain-specific keywords. The reason behind this is the tendency of people to use words such such *outage* and *blackout* to mention an anticipated outage in the future rather than using these words to report an active outage. Because we only annotated tweets about active outages in the outage-related categories, a large portion of the tweets ended up in the Not Relevant category.



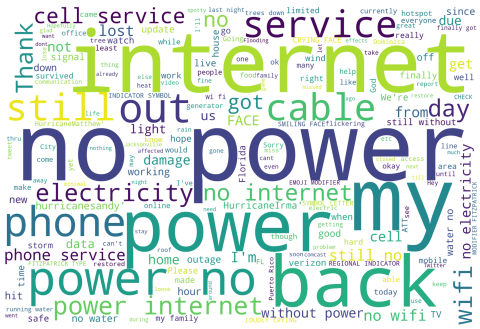
(a) Not-relevant



(b) Power-outage



(c) Communication-outage



(d) Power-communication-outage

Figure 3.4: The salient words in each tweet category.

3.4 Dataset Analysis

In this section, we analyse the annotated dataset to better understand the nature of tweets that contain outage-related keywords. Our goal, through this analysis, is to highlight the differences that exist between the not-relevant class and others as well as between individual outage-related classes. In particular, we determine the inherent features such as popular words, bi-grams, tri-grams, hashtags and sentiments that are present in the tweets in each category. In Figure 3.4 we present the salient words associated with each of these four categories. We note that the not-relevant category consists of many of the same words that are present in other categories. However, as mentioned previously, the tweets in this category do not actually identify an outage. The salient words present in other categories are consistent with the names of the categories. Below we first perform

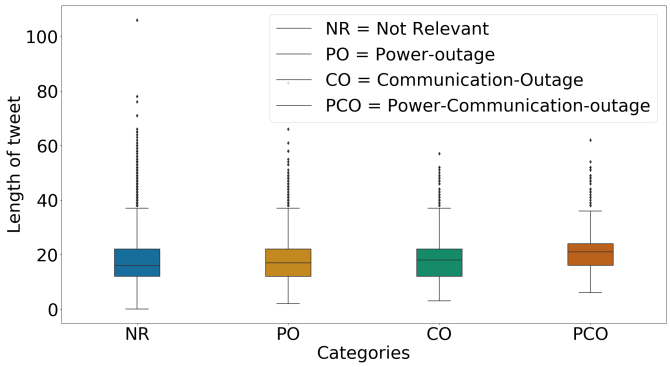


Figure 3.5: Length of tweets in each category.

lexical analysis to detect features such as single words, bi-grams, tri-grams and hashtags that are prevalent in each category. We then proceed to analyse the sentiments that are associated with the tweets by category.

3.4.1 Lexical Analysis

The lexical analysis of each category is presented below.

Not-relevant: This category consisted of over 75% of the total annotated tweets. In addition to investigating the most commonly occurring words in this category, as shown in Figure 3.4a, we evaluated the predominant bi-grams and tri-grams. Among single words, *power*, *mobile*, *No*, *outage* and *plane* were the five most frequent words in this category. For bi-grams, the words *my phone*, *no power*, *power outage*, *cell phone* and *phone call* appeared most frequently. From the frequently occurring words and bi-grams, it is not yet apparent that the tweets in this category do not convey any specific outage-related information. However, the most common tri-grams in this category, which include *get radio play*, *uncut internet station*, *charge my phone*, *got my phone* and *my phone off* shed more light on the nature of these tweets. Additionally, we investigate the frequent hashtags of the tweets from this category. The top three hashtags are *#mobile*, *#news* and *#tech*. Figure 3.5 shows the length of tweets in this category. On average, each

tweet contained 18.6 words. We note the presence of a large number of outliers in the length of tweets in this category compared with others.

Power-outage: This is the second most popular category among the annotated data, containing 13% of the entire dataset. The category includes tweets that reported either power outages or restoration of power after an outage. The most frequently appearing words in this category are shown in Figure 3.4b. The top five common words are *No*, *power*, *hit*, *area* and *days*. Outage-related words such as *outage* and *blackout* also appear in the tweets in this category. The popular bi-grams in this group of tweets include *no power*, *still no*, *power outage*, *no electricity* and *without electricity*. Some of these words are also present in the commonly occurring tri-grams, which include *still no power*, *no power no, no power thanks*, *no power my* and *no power since*.

Further analysis of the 2,791 tweets in this category determined that 4% of the tweets mention power restoration after an outage. 20% of the power outage tweets were observed to be informative in nature, providing useful information about an outage. These informational tweets reported areas experiencing outages and in many cases included live updates from news organizations that stated the number of people experiencing outages in affected areas. The majority of tweets in this category, 76%, directly reported an outage during the time of the outage. Popular hashtags in this category are *#blackout*, *#nopower* and *#lightsout*. Figure 3.5 shows the distribution of the length of the tweets in this category. Tweets in this category contained 18.3 words on average.

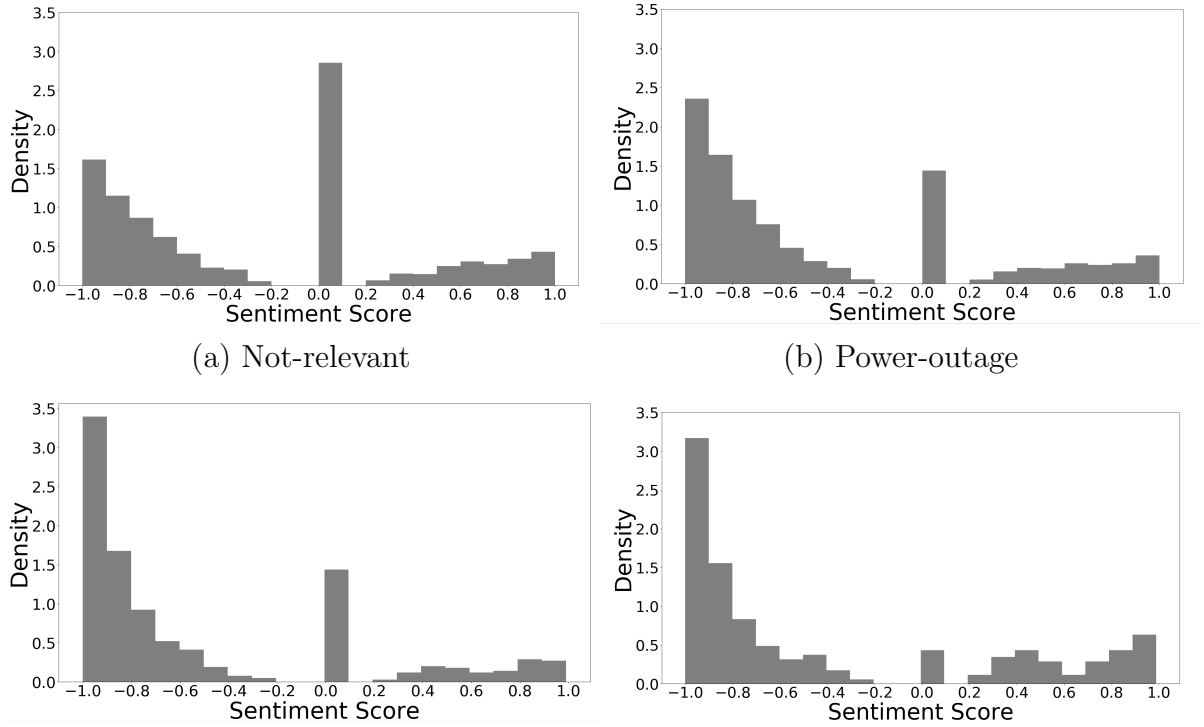
Communication-outage: This category of tweets represents 5.5% of the overall number of tweets in our annotated dataset. Tweets in this category either inform about or report an active communication outage or mention having some form of communication capabilities returned after their loss. Popular words in this category include *internet*, *service*, *wifi*, *no* and *out* and are shown in Figure 3.4c. One interesting observation is that specific

provider names, such as *verizon*, *tmobile* and *xfinity* appeared frequently in the tweets of this category. This could be as a result of users being more familiar with the names of their telecommunication service providers. The most popular words pairs in the tweets of this category include *my internet*, *no internet*, *phone service*, *no service* and *internet down*. Tri-grams such as *cell phone service*, *my internet down*, *mobile networks knocked*, *networks knocked out* and *still no internet* emerged as the most common. The collection of these words indicate that when reporting a communication outage, people tend to use the word *service* together with *down*. Power outages are reported using *outage* and *blackout* in addition to *out*. Similar to the power-outage category, we subdivide the communication-outage related tweets into three subcategories. 9% of the tweets belong to the sub-category of tweets that mention restoration of communication service after an outage. 24% and 67% of the tweets in the communication-outage category inform or report about a communication outage, respectively. Hashtags *#wifi*, *#att* and *#internet* are the three most frequently used in this category. Unlike the most popular hashtags in the power-outage category, hashtags in this category do not inherently convey information related to an outage. Tweets in this category have an average length of 18.9 words as seen in Figure 3.5.

Power-communication-outage: This category contained the fewest tweets, about 2% of the overall annotated dataset. Because this category consists of tweets that must mention both power and communication outages, the average length of a tweet in this category, shown in Figure 3.5, is 22 words long. As can be seen from Figure 3.4d, these tweets combine the keywords from both the power-outage and communication-outage categories. Popular keywords include *power*, *internet*, *no*, *back* and *service*. Common bi-grams are *no power*, *power no*, *power internet*, *no internet* and *cell service*. We find that the tri-grams *still no power*, *power no cell*, *no cell service*, *no electricity no* and *no power internet*

appear most frequently. As the tweets in this category are longer than the rest on average, we also determine the commonly occurring four-grams. These include *no power no cell*, *no power no internet*, *power no cell service*, *no power no wifi* and *no power cell service*. We observe that in addition to reporting about experiencing both power and communication outages, a number of tweets reported either having power but no communication or vice versa. Some tweets also provided information related to power and communication outages. When we analyze the nature of the tweets in this category further, we find that 10% of the tweets mention having power while experiencing some form of communication outage. Similarly, 10% of the tweets mention having communication capabilities while suffering from power outage. 15% of the tweets mentioned getting back both power and communication services after an outage. Informative tweets such as those providing locations and number of people experiencing power and communication outages formed 7% of the tweets in this category. Finally, 58% of the tweets reported experiencing both power and communication outages. The top three hashtags are *#poweroutage*, *#electricity* and *#finallygotpowerback*.

The lexical analysis of these categories highlights various salient features present in each category. The popular words and bi-grams of the not-relevant category are similar to those of the actual outage-related categories. In spite of the similarity between keywords, further analysis of the tri-grams and hashtags of these categories shows that the contents of the tweets in the not-relevant category do not report an outage. It is also noticed that during hurricanes, users tend to anticipate experiencing power and communication outages and post on Twitter before such outages actually occur. In addition to reporting an outage, tweets often mention restoration of services after an outage as well as provide meaningful information such as number of people experiencing an outage.



(a) Not-relevant

(b) Power-outage

(c) Communication-outage

(d) Power-communication-outage

Figure 3.6: Distribution of sentiment scores of each category.

3.4.2 Sentiment Analysis

To better understand the inherent traits of the tweets that are present in these categories, we perform sentiment analysis [192] using the sentiment analysis API provided by IBM Watson [12]. IBM Watson analyzes the sentiment associated with a statement and assigns it a score between -1 and 1. A score closer to -1 conveys extremely negative sentiment while a score closer to 1 signifies more positive sentiment. Figure 3.6 shows the distribution of the sentiment scores of each of the four categories. The average sentiment score of not-relevant, power-outage, communication-outage and power-communication outage categories is calculated to be -0.26, -0.42, -0.51 and -0.40, respectively. As seen from Figure 3.6a, the sentiments associated with tweets in the not-relevant category are more neutral as they hover around 0. In contrast, the sentiment scores of the rest of the cate-

gories are more concentrated in the negative side of the scale with communication-outage tweets having the most negative sentiment. Overall, in the not-relevant category, 29% of the tweets had sentiment score of 0, while 51% of the tweets attained negative scores. The percentage of tweets with negative sentiment score increased for the other categories. The power-outage, communication-outage and power-communication-outage categories had 68%, 72% and 70% of their tweets with score below 0, respectively.

3.5 Outage-Specific Classification

In this section, we design a two-stage classification framework to automate the process of detecting outage-related tweets. Before performing the first level of classification, we collect a new set of tweets, using Crimson Hexagon, that occurred during the seven hurricanes. This dataset is comprised of tweets that contain only hurricane-specific keywords and not outage-specific keywords. These tweets are then added to the previously annotated dataset to form two separate classes of tweets. We first perform a binary classification to quickly extract all tweets that have our outage-specific keywords in addition to hurricane-specific keywords from the rest of the tweets. Before performing classification, we clean and pre-process the dataset. Once we have identified the outage-related tweets, we then perform the second level of classification, only on the annotated dataset, to automatically place tweets into the categories we established in the previous sections. Below we present the details associated with the pre-processing and classification tasks.

3.5.1 Pre-processing Dataset

Tweets are typically not properly grammatically structured and are likely to contain abbreviations, rendering them incomplete and noisy. In order to sanitize the dataset, we employed multiple text pre-processing steps. We removed URLs, non-ASCII characters

and non-English characters. We also removed hashtags, user names and date and time strings. Emoticons were converted to UNICODE strings. To reduce the feature space, we converted all words to lower case.

We next created a custom set of stop words to ensure that we preserved the context of our tweets while eliminating unnecessary repeated stop words. For example, the stop words library provided by Python's NLTK contains 179 words such as *as*, *they*, *himself*, *out*, *down* and *not*. Removing the words *out*, *down*, *off*, *no* and *not* from our tweets could leave outage-related tweets meaningless. Hence we excluded these words from the stop words library. Additionally, we removed occurrences of event-specific words, such as hurricane, sandy and irma from the training dataset. This was done to ensure that the classifiers did not become dependent upon such words while identifying information that we require.

We used popular word embeddings frameworks to perform word vector initialisation. To generate word tokens, we first used term-frequency-inverse-document-frequency (tf-idf). Tf-idf is used to obtain the most important words within the tweets. These tokens from tf-idf were subsequently vectorized using GloVe [199]. We choose GloVe over another widely used word embedding framework, Word2Vec, due to the former's ability to take the ratio of the co-occurrence probabilities of consecutive words to establish semantic meanings for those words. For binary classification, we employed nine state-of-the-art classifiers such as logistic regression, support vector machine and K-nearest neighbors. These simpler classifiers were implemented using the scikit-learn 0.21 [198] library of Python. To extract various classes of our outage-related events, we used popular neural network models such as convolutional neural network (CNN) and recurrent neural network (RNN), in addition to the simpler models. These more sophisticated models were implemented using Keras with Tensorflow backend [13], as this platform contains the packages that are required to run these algorithms. Additionally, we implemented

an emerging technique of text classification known as transfer learning to perform classification of our categories.

3 All the classifiers were run on Google Cloud Compute powered by a 16GB NVIDIA Tesla V100 GPU. In addition to using a categorical cross entropy loss function with our neural network models, we also employed focal loss [174], which has been proven to be effective in classifying minority samples in image classification tasks. Next we present details associated with the two types of classifications we conducted and various methods we implemented to achieve better classification success.

3.5.2 Binary Classification

The goal of binary classification is to quickly isolate the domain-related tweets to conduct further information extraction. Specifically, we want to separate the tweets that contain hurricane-specific keywords from those that also contain outage-related keywords.

To perform binary classification, we first create a training dataset of a roughly equal number of samples that contain only hurricane-specific keywords (but not outage-specific keywords), comprising class 0, and tweets that contain both hurricane-specific and outage-specific keywords (our annotated dataset), forming class 1. We collected equal numbers of geo-tagged and un-tagged tweets that contained hurricane-specific keywords but excluded our outage-specific keywords using [10]. The training set consisted of $10,007$ tweets, of which $5,236$ samples belonged to class 0 and the rest to class 1. The distribution of the two classes in the test set, however, was kept similar to what we observed while curating the original dataset in Section 3. Because outage-related tweets only comprised a very small fraction of the overall tweets that contained hurricane-related keywords, our test set contained $2,326$ tweets, of which $2,203$ belonged to class 0 and 123 belonged to class 1 (making up roughly 5% of the dataset). This small number of tweets of class 1 ensures consistency with what is observed during a real scenario. However, this

results in difficulty in identifying these tweets with very high precision and recall. To perform this layer of classification, we only employed the simple classifier models as they are computationally inexpensive and capable of producing results with high accuracy.

3.5.3 Category Classification

Once we successfully filter tweets that contain outage-related keywords, we then attempt to further classify these tweets into the four major categories we established in Section 3. This is done to obtain more fine-grained information about different outage-related events. We first create a training set by selecting $3,500$ random samples of the not-relevant class (class 0). The rest of the training set is formed of $2,295$ randomly selected tweets from the power-outage category (class 1), 828 tweets from the communication-outage category (class 2) and 306 tweets from the power-communication outage category (class 3). As with the binary classification task, we kept the distribution of categories in the test set similar to the original dataset. In our test set, we selected $1,500$ tweets from the not-relevant category, 496 tweets from the power-outage category, 172 from the communication-outage category and 43 tweets from the power-communication outage category.

In addition to the simple classifiers, we employed neural network and transfer learning models to extract tweets of each category in this layer of classification. To address the imbalance problem in our dataset, we applied the sampling technique SMOTE [139] and various sampling ratios amongst the classes. These techniques, however, fell short in improving the classification performance while detecting outage-related tweets, as they failed to adopt to the feature space that exists in our tweets. Therefore, because this is a multi-class classification problem, we instead first use a categorical-cross entropy loss function with a softmax layer in our neural network models. The categorical-cross

entropy loss function can be mathematically defined as:

$$H(y, \hat{y}) = -\sum_{j=0}^{M}\sum_{i=0}^{N}(y_{ij})(log(\hat{y}_{ij}) \tag{3.1}$$

where H is the loss function, y is the actual label of the i^{th} observation of the j^{th} class and $\hat{y}$ is the predicted label for the observation made by the softmax layer of the neural network. An issue that arises with this loss function is that in a skewed dataset, it fails to properly penalise the classifier when it predicts the majority class. Because we are dealing with a dataset that exhibits class imbalance, we incorporate a loss function, known as focal loss, with our neural network classifier. Focal loss has proven to increase classification accuracy in datasets that suffer from the imbalance problem between classes [203]. Focal loss can be represented as:

$$FL(p_j) = \alpha(1 - p_j)^{\gamma} log(p_j) \tag{3.2}$$

where FL is the focal loss function and p_j is the softmax probability of the j^{th} class for a particular observation. α and γ are two regularizing parameters. This loss function adds more importance when the network predicts a minority sample as opposed to the overly represented sample. This makes it ideal for performing classification on an imbalanced dataset.

We choose a number of neural networks that have proven effective in text classification to perform this level of classification. To determine the ideal hyper-parameter configuration for each neural network, we use Grid Search [18] starting with multiple numbers of configurations. We train the CNN model using 100-word long embedding vectors alongside 512 convolutions filters of sizes $2, 3, 4, 5$. To avoid over-fitting, we use a dropout of 0.5 while training with the Adam gradient descent optimizer [170]. The

CNN model was run for 10 iterations with a batch size of 32. We also evaluated the performances of both LSTM and GRU-based bi-directional RNN. These RNN models were further incorporated with an attention layer to improve performance. We trained the RNN models containing 100 neurons for 20 iterations. We then employed Hierarchical Attention Network (HAN) [224] with 200 LSTM based word encoders and 250 sentence encoders. Finally, we tested the performance of Bidirectional Encoder Representations from Transformers (BERT) as a transfer learning model for the classification task [145]. Transfer learning models are pre-trained on a very large corpus and then fitted to perform classification on a smaller number of domain-specific data points. We used BERT-Large, Uncased (Original) model as the pre-trained model due to its ability to produce good results while remaining computationally inexpensive [17].

3.6 Results

In this section we first present the results obtained after applying different classifiers to detect tweets that contain outage-related words. We then present the performance of the classification models in identifying specific outage categories. We compare the performance of the classification models by measuring the per-class precision, recall and f-score that each of these models produce. In addition, we compare the overall accuracy of each model as well as the time it takes for the model to perform the classification task. Because our goal is to classify the outage-related tweets quickly, the runtime for each algorithm presents us with important information we need to select the right model to perform the classification. Our goal is to determine the model that is able to quickly detect outage-related tweets with high accuracy, precision and recall scores.

Table 3.4: Performance comparison of the binary classifiers.

	Not-outage-related			Outage-related				
Methods	Precision	Recall	F1-score	Precision	Recall	F1-score	Accuracy	Runtime(seconds)
Bagging	0.96	0.92	0.94	0.21	0.38	0.27	0.89	5.3
Boosting	0.99	1	1	0.99	0.94	0.97	0.99	1.94
Decision Trees	0.99	0.97	0.98	0.62	0.94	0.75	0.96	0.59
K-nearest neighbors	0.97	0.71	0.82	0.1	0.6	0.18	0.7	0.71
Logistic Regression	0.99	0.98	0.98	0.67	0.94	0.78	0.97	1.56
Multinomial Naive Bayes	0.99	0.87	0.93	0.28	0.9	0.43	0.87	0.12
Nearest Centroid	0.99	0.92	0.95	0.36	0.82	0.5	0.91	0.22
Random Forest	0.99	0.99	0.99	0.8	0.94	0.87	0.98	2.78
Support Vector Machine (SVM)	0.99	0.99	0.99	0.84	0.95	0.89	0.99	0.1

3.6.1 Binary Classification

Table 3.4 presents the results we obtained after applying each of the nine classifiers on a curated dataset that contained only hurricane-specific keywords (class 0) as well as hurricane-specific and outage-related keywords (class 1). Almost every model performs exceptionally well in identifying tweets that contain only hurricane-specific keywords. The precision, recall and F1 scores of these models are very close to 1 when classifying members of class 0. In comparison, only a small set of models are able to identify samples of class 1 with good precision, recall, and f1-score. The boosting algorithm identifies class 1 tweets with the highest precision, recall and f-score values. Because the boosting algorithm has a hierarchical tree structure, where a new tree learns from the results of the previously trained tree, it is able to perform better than other simple classifiers when performing binary classification. The SVM and random forest models achieve the second and third best performance in classifying the samples from class 1, respectively. K-nearest neighbor performs poorly when classifying class 1 samples. This occurred as a result of the insensitivity of the distance function of K-nearest neighbor towards small but meaningful differences between tweets. In addition to performing the overall classification task reasonably well, SVM also recorded the fastest run-time.

Table 3.5: Accuracy and runtime of the models used to perform outage-related categories classification.

Model	Accuracy	Runtime(Seconds)
Bagging	0.77	3.38
Boosting	0.84	9
Decision Trees	0.79	0.71
K-nearest neighbors	0.76	0.47
Logistic Regression	0.86	1.68
Multinomial Naive Bayes	0.82	0.11
Nearest Centroid	0.79	0.11
Random Forest	0.84	3.57
Support Vector Machine	0.86	0.16
CNN	0.65	645.05
CNN-Focal	0.84	650.89
RNN-LSTM	0.83	2309.47
RNN-GRU	0.84	1907.36
RNN-Attn-LSTM	0.84	2541.71
RNN-Attn-GRU	0.8	2216.57
RNN-LSTM-Focal	0.83	2239.26
RNN-GRU-Focal	0.83	1953.92
RNN-Attn-LSTM-Focal	0.83	2409.45
RNN-Attn-GRU-Focal	0.84	2261.31
HAN	0.85	2335.13
HAN-focal	0.82	2342.31
BERT	0.88	87

3.6.2 Outage-category Classification

Table 3.5 presents the accuracy and run-time of the classification models. Table 3.6 presents the classification performance achieved by these models in detecting not-relevant, power-outage, communication-outage, and power-communication-outage tweets.

When comparing the accuracy of the models, Table 3.5 indicates that of the simpler models, boosting, logistic regression, random forest and SVM achieve accuracy scores above 0.8. These models also record low run-times, ranging from 0.16 to 9 seconds. The simpler models classify tweets from the not-relevant category with high precision and recall. In categorizing power-outage related tweets, the simpler models perform reasonably

Table 3.6: Classification performance of the models in detecting tweets per category.

Methods	Not-relevant			Power-outage			Communication-outage			Power-Comm-outage		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Bagging	0.84	0.87	0.86	0.65	0.72	0.68	0.47	0.31	0.37	0.33	0.09	0.14
Boosting	0.91	0.87	0.89	0.76	0.92	0.83	0.59	0.44	0.5	0.54	0.57	0.56
Decision Trees	0.9	0.8	0.85	0.73	0.87	0.79	0.41	0.56	0.47	0.45	0.61	0.52
K-nearest neighbors	0.85	0.85	0.85	0.65	0.69	0.67	0.41	0.39	0.4	0.21	0.09	0.13
Logistic Regression	0.92	0.9	0.91	0.77	0.92	0.83	0.63	0.52	0.57	0.65	0.25	0.36
Multinomial Naive Bayes	0.84	0.93	0.89	0.76	0.83	0.79	1	0.02	0.03	0	0	0
Nearest Centroid	0.92	0.8	0.86	0.8	0.82	0.81	0.34	0.62	0.44	0.34	0.68	0.45
Random Forest	0.91	0.88	0.9	0.72	0.93	0.82	0.66	0.47	0.55	0.69	0.2	0.32
Support Vector Machine	0.94	0.86	0.9	0.77	0.94	0.85	0.57	0.66	0.61	0.59	0.39	0.47
CNN	0.96	0.54	0.69	0.67	0.92	0.78	0.25	0.84	0.39	0.26	0.77	0.39
CNN-Focal	0.91	0.87	0.89	0.76	0.92	0.83	0.53	0.45	0.49	0.63	0.6	0.62
RNN-LSTM	0.95	0.81	0.87	0.76	0.91	0.83	0.49	0.79	0.6	0.48	0.7	0.57
RNN-GRU	0.93	0.84	0.88	0.75	0.91	0.82	0.56	0.64	0.6	0.63	0.79	0.7
RNN-Attn-LSTM	0.92	0.85	0.88	0.79	0.85	0.82	0.52	0.7	0.59	0.59	0.74	0.66
RNN-Attn-GRU	0.94	0.77	0.85	0.75	0.9	0.81	0.42	0.81	0.55	0.6	0.74	0.67
RNN-LSTM-Focal	0.93	0.82	0.87	0.76	0.89	0.82	0.48	0.73	0.58	0.56	0.84	0.67
RNN-GRU-Focal	0.94	0.81	0.87	0.76	0.92	0.83	0.48	0.74	0.59	0.59	0.79	0.67
RNN-Attn-LSTM-Focal	0.92	0.83	0.87	0.77	0.9	0.83	0.47	0.64	0.54	0.6	0.7	0.65
RNN-Attn-GRU-Focal	0.93	0.85	0.89	0.77	0.9	0.83	0.54	0.69	0.6	0.6	0.58	0.59
HAN	0.93	0.86	0.89	0.79	0.87	0.83	0.57	0.71	0.63	0.56	0.84	0.67
HAN-focal	0.96	0.77	0.86	0.74	0.96	0.84	0.46	0.77	0.57	0.47	0.88	0.61
BERT	0.93	0.9	0.91	0.83	0.89	0.86	0.67	0.66	0.66	0.69	0.84	0.7

well, with logistic regression and boosting models achieving an f-score of 0.83. The performance of the simpler models, however, drops significantly while detecting samples from the two minority categories: communication-outage and power-communication-outage. This occurs as a result of these models' inability to learn classes with a small number of samples in an unbalanced dataset. Logistic regression, random forest and SVM are the only three models that produce an f-score greater than 0.50 when identifying tweets in the communication-outage category. In classifying tweets from the power-communication-outage category, among the simpler models, only the boosting and decision tree models achieve an f-score above 0.50.

As expected, the run-times of the neural network models are significantly greater than their simpler counterparts as shown in Table 3.5. CNN models execute fastest whereas RNN models take the longest. In terms of accuracy, except for the CNN model with categorical cross-entropy loss function, every other neural network model achieves accuracy scores around 0.80. The models also perform fairly similarly when classifying samples from the not-relevant categories. As seen in Table 3.6, the precision recorded by the neural network models exceeds 0.90 in detecting not-relevant tweets. Except for CNN with categorical cross-entropy loss function, all other models achieve recall scores of around 0.80 in detecting tweets of this class. Precision scores between 0.75 and

0.79 are reached by the neural network models when identifying power-outage tweets. The difference in performance between the simpler models and neural network models is seen when detecting communication and power-communication-outage tweets. The neural network models achieve higher recall scores in detecting communication-outage tweets while reaching better f-scores than the simpler models when identifying power-communication-outage tweets. Focal loss outperforms categorical-cross entropy loss when used with CNN across all four categories. It also records a 10% increase in f-score when used in conjunction with the RNN-LSTM model compared to the categorical cross-entropy loss function when detecting power-communication outage tweets.

The best performance in all the considered metrics is achieved by BERT in this classification task. From Table 3.5, we can see that though it takes longer to execute than the simpler models, it is able to achieve the highest accuracy; further, its run-time is faster than all the neural network models. It records the best f-scores when detecting tweets that belong to both the not-relevant class and outage-related categories. Because BERT is already pre-trained on a large corpus of texts, it is able to identify the tweets with very good performance, making it an ideal candidate to perform this classification task.

3.6.3 Remarks

With the aid of our annotated dataset and machine learning algorithms, we are able to detect outage-related events from a large stream of tweets that appeared online during recent hurricanes. The binary classifier is able to separate outage-specific tweets from others quickly, thereby reducing the amount of time needed to extract domain-specific tweets. Once these outage related tweets are detected, they can be further classified into different groups with the aid of an advanced learning model such as BERT. Using the ideal model to perform each level of information extraction will result in rapid classification of

tweets, which first responders can then use to take immediate action or aid planning of additional operations.

3.7 Conclusion

In this work, we take an in-depth look at the tweets that originate during hurricanes and convey important outage-related information. We first determine keywords that are commonly used during power and/or communication outages. We use these keywords to obtain tweets that were posted during the seven major hurricanes in the USA between 2012 and 2018. These tweets were then annotated and placed into one of the four categories based on the outage information they contained. We perform a detailed analysis to better understand the type of tweets that belong to each of these categories. Finally, we apply various state-of-the-art machine learning models to first detect tweets that contain our outage-specific keywords and subsequently place them in their respective categories. Results show that computationally inexpensive models such as SVM and logistic regression can be used to filter out tweets that mention words related to outages. Through use of transfer learning models such as BERT, such outage-related tweets can be detected with high accuracy, precision and recall. Our framework can be implemented to provide first responders with outage related information during natural disasters. In our future work, we will build a user interface that incorporates classification models to perform real-time detection of outage-related tweets. Another next step is to conduct a deeper level of information extraction to sub-classify the tweets within each outage-related category. For example, with enough samples, we can train a classifier to automatically identify tweets that mention restoration of services or other useful information.

Part II

Internet Quality

Chapter 4

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

4.1 Introduction

The term "digital inequality" refers to the gap in Internet access that exists across different geographic areas and demographic variables [9]. Access to the Internet is known to impact multiple facets of human life, including economic [19], education [34], health [24], and, more recently, the ability to self-isolate to prevent spread of COVID-19 [33]. 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 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 [23, 20, 21, 36].

Internet access quality has received less attention than availability in part due to the dearth of reliable and granular data related to Internet quality [161]. 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 [30, 179]. To improve access and quality of Internet, large financial investments have been made by the federal government [26], 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 [57] 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 [43] 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 chapter, 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, this chapter 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.

4.2 Description of the Measurement Data

We begin our study by combining publicly available Internet QoS data from the Speed Test by M-Lab [58] with ESRI demographic data [45]. In the following section, we describe these datasets in more detail.

4.2.1 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 [57]. 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 [54, 55, 53]. The test consists of bulk exchange of data between the client and server, as defined in IETF RFC 3148 [184]. 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 [39].

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 [52], 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

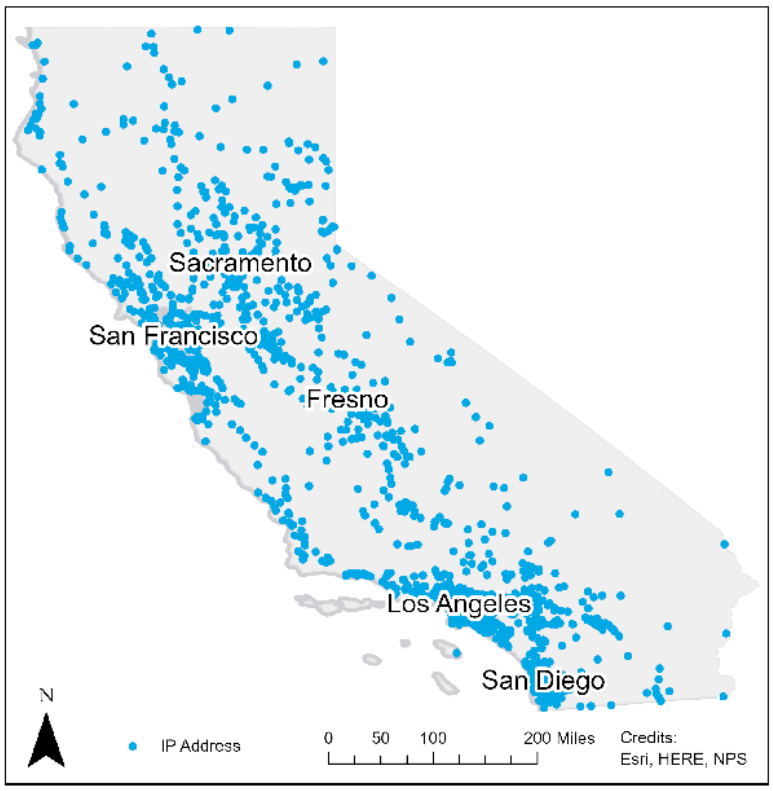


Figure 4.1: Location of Unique IP addresses in the M-Lab California Data.

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

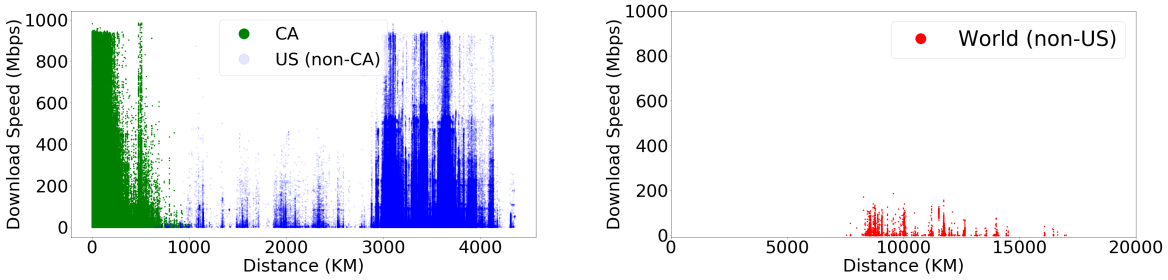
Table 4.1: Access Technologies

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

Table 4.2: Geographic Areas

client's ISP information to separate these measurement types.

Table 4.1 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 4.1, 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 [42]. On average, a group of 39 census blocks form a census block group [41] and contains between



(a) Servers located in the US.

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

Figure 4.2: Download Speed of Measurements for Different Server Locations.

600 and 3000 people. It is also the smallest geographic unit for which the Census Bureau publishes sample data [62]. A census tract is formed with at least one census block group [62] and contains a population size between 1200 and 8000 people. A zip code is a US. Postal Service designated area. While a zip code contains an arbitrary number of census block groups [61], it is not considered a census unit. The shape file for each geographic area is obtained from the resources provided by the Census Bureau [3]. We map the location of the data points in each of these geographic areas within California (CA). Table 4.2 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 endpoints. 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 4.1 shows the location of the IP addresses present in our California dataset.

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 US. in our dataset. Among the ones in the US., 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 [49] to calculate the great circle distance between the client and server location for each test. Figure 4.2 plots the download speed for each client-server distance in our dataset. We observe that measurement tests to servers outside of the US. ("World (non-US.)") almost always recorded lower download speeds (see Figure 4.2(b)). Thus, we ignore them for our analysis. In contrast, using measurements to servers in the US. (outside CA) has a marginal impact on the download speed. Thus, we consider all measurement tests to US.-based servers for our analysis.

4.2.2 Demographic Data from ESRI

ESRI is a We utilize the demography data provided by ESRI's Updated Demographics [64]. 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 [46]. Using [38], 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 the 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 [19, 34, 179] 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

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

Table 4.3: Summary Statistics for QoS and Demographic Variables.

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 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 [5]. 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 4.3.

4.2.3 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 [153, 134]. 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 [159]. 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.

4.3 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.

4.3.1 Correlation between Download Speed and Demographic Attributes

We use the Pearson Correlation Coefficient (PCC) [59] as it is suitable to capture any relationship that might exist between demographic attributes and download speed. Table 4.4 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, *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

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

Table 4.4: Pearson Correlation Coefficient between Download Speed and Demographic Attributes.

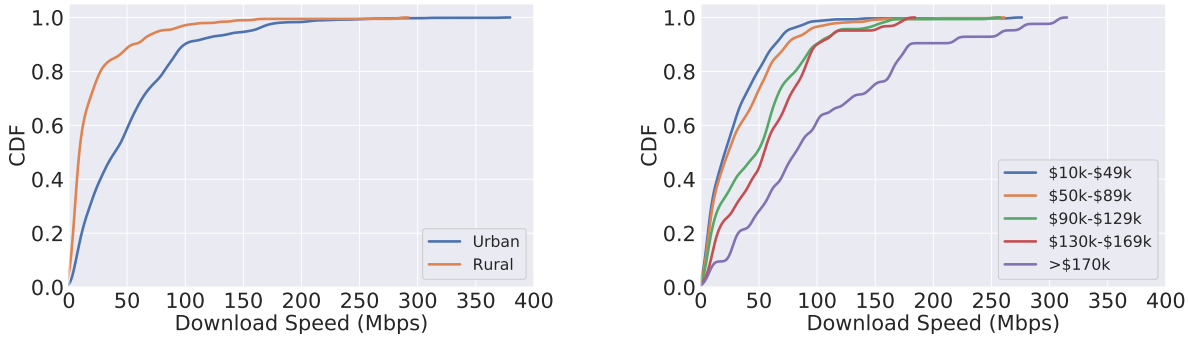
group's poverty rate. For area, we encode urban block groups as 0, rural block groups as 1, and perform special point-biserial correlation (equivalent to Pearson Correlation) [60] with download speed. This results in a negative correlation of download speed with rural areas. We observe that a census 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 4.4 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.

Effect of Area Type.

Table 4.4 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 4.3(a) shows the cumulative distribution function of download speed in each of these



a. Area b. Income

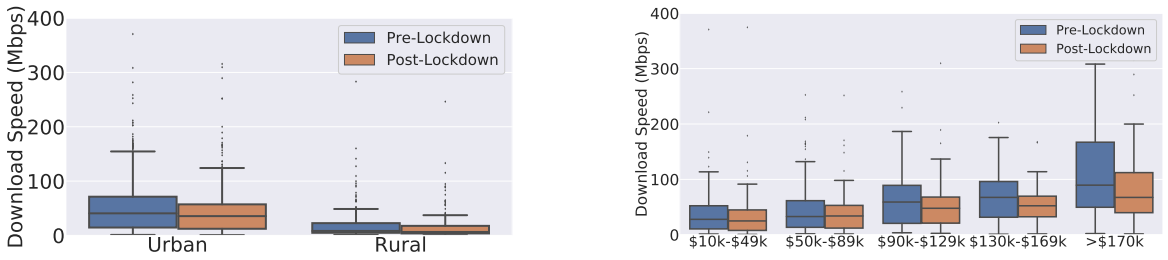
Figure 4.3: Cumulative Distribution Function of Download Speed by Area Type and Income.

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 [6]. 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.

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



a. Area Type b. Income

Figure 4.4: Download Speed before and during Lockdown by Area Type and Income.

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 4.3(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.*

	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

Table 4.5: Average Download Speed for Area Type and Income Pre- and Post-COVID-19 Lockdown.

4.3.2 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 [29]. As found in a recent study [177], 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 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 4.4 presents the speed recorded during these two periods, disaggregated by area type and urban census block group income bins.

Figure 4.4(a) shows the speed recorded during these two periods within urban and rural block groups. Table 4.5 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

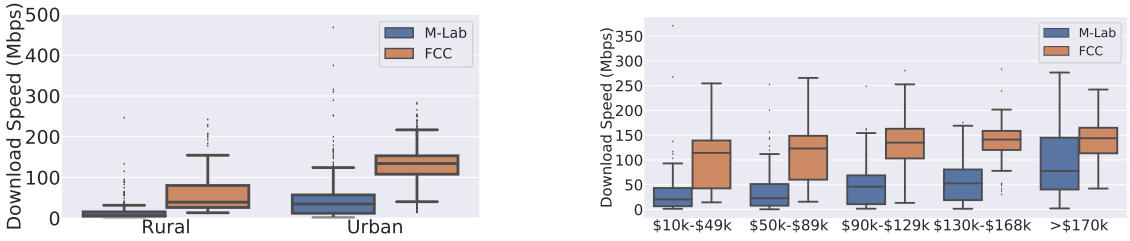
urban download speed remained 2.5 times the average rural speed.

In Figure 4.4(b), the download speeds recorded before and during the lockdown in urban block groups are grouped by income. From Table 4.5, 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 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.

4.3.3 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 [179, 8], 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 4.5 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



a. Area b. Income Bins

Figure 4.5: Comparison of M-Lab and FCC Download Speed by Area Type and Income.

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 [179]. 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 4.6, *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 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*

	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

Table 4.6: FCC Accuracy Factor by Area Type and Income.

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.

4.4 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 [25, 50]. 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.

4.5 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 [44]. However, this estimate is only done for cities with population greater than $65,000$, leaving smaller communities undocumented. Similar to our work, [22] 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 [35], 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 [28] 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 [158] 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 [222, 202, 201, 183], 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 [134] used crowdsourced measurements to benchmark Internet performance across multiple metropolitan areas. In [146], 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 [147].

4.6 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.

Chapter 5

The Importance of Contextualization of Crowdsourced Active Speed Test Measurements

5.1 Introduction

The challenge of mapping fixed broadband Internet access was brought to the forefront during the stay-at-home orders of the Covid-19 pandemic. Suddenly, individuals without high-quality Internet access could not participate in the remote schooling, work, and telehealth that these orders required [19, 34, 24]. Further, federal money for Internet infrastructure improvement was made available through the Bipartisan Infrastructure Investment and Jobs Act [84]; however, a key challenge remained: knowing where high-quality Internet access was lacking [67, 89]. While the Federal Communications Commission (FCC) has long compiled annual Broadband Reports that map provider-reported access at the census block level, these reports are known to overstate access availability and speed, particularly in rural and under-served urban areas [179, 96, 70].

Crowdsourced active network measurements have emerged as a powerful tool to map fixed broadband access more accurately. These "speed tests" provide a critical snapshot of the network state from the vantage point of the end users. Because they are active measurements, they provide data on actual performance instead of the theoretical maximum performance reported by the providers. Popular network speed test platforms, such as Ookla's speedtest.net [93], Measurement Lab's speed.measurementlab.net [57], FAST [76] and Xfinity's speed test [103], are utilized by Internet users worldwide to conduct these measurements. For instance, Ookla claims over 40 billion user-initiated tests since its inception [95]. Because of the inherent benefits, numerous governmental initiatives (e.g. [131, 72, 82, 102, 85, 69]) have come to rely on crowdsourced speed test data to map broadband access. With this data, local governments, community organizations, and others can attempt to discern where to make the economic investment in infrastructure to address digital inequality. *Perhaps most critically, the FCC itself has recently specified a challenge process [79], whereby individual users and communities can gather active measurement data to challenge provider-reported coverage claims.*

However, despite the broad use of crowdsourced active network measurements and the call for their usage by the FCC, the data generated through these speed tests suffer from several key limitations, which must be addressed before drawing meaningful conclusions about fixed Internet performance. More concretely, *we argue that speed test measurements must be contextualized to accurately interpret the measured performance.* The challenge here is understanding *what a speed test measures* and *how it compares to expected speed values.* For example, many fixed broadband plans offer rates as high as 1 Gbps download and 35 Mbps upload. If a speed test measures performance significantly less than these values, is it because the access network is under-performing, the user has purchased a lower-tier plan, or the user's home WiFi network is misconfigured or experiencing interference? It is critical to determine the source of the under-performance. If the

under-performance is attributable to issues in the access network, then the problem could be reported to the Internet Service Provider (ISP) to challenge coverage claims in an area. In contrast, if the under-performance is attributable to local factors, such as channel interference or poor signal quality, the user can address it directly. If the user simply purchased a lower-tier plan, then perhaps the speed test is measuring the paid-for speed. Finally, the methodology of the test itself can impact performance results, adding anothe layer of complexity [153, 140].

In this chapter, we utilize more than 1.5M total measurements from Ookla and M-Lab speed tests to demonstrate the critical need for contextualization of these measurements. We start with an analysis of aggregate performance, as represented by this data, across four major metropolitan cities in the US. To demonstrate the importance of subscription plan context, we propose a novel approach called the Broadband Subscription Tier (BST) methodology that determines, with over 96% accuracy, the subscription plan associated with a group of speed test measurements. We evaluate the accuracy of this methodology on over 60k Measuring Broadband America (MBA) data points, for which we have subscription ground truth. After applying the methodology to our M-Lab and Ookla datasets, we show that the majority of the speed tests in a city originate from the lower subscription tiers. This implicit bias in the data skews the overall results for metrics such as download speed to lower throughputs.

Second, we incorporate the subscription tier context to Ookla measurements to quantify the impact of factors such as access type (WiFi vs. Ethernet), WiFi spectrum band, RSSI and device memory. We find that side effects of these local factors can lead to performance that only achieves half the data rate of the subscribed plan. We also evaluate the impact of the time of the test on the measured performance and, interestingly, discover minimal impact. Finally, we evaluate the performance of M-Lab versus Ookla speed test results for each subscription tier and demonstrate that M-Lab tests consis-

tently achieve lower download speeds than Ookla tests, at times by as much as a factor of two.

In summary, our work makes the following contributions:

- We develop a novel methodology (BST) that maps crowdsourced speed test results to the residential broadband subscription plan at the test location. We demonstrate over 96% accuracy on 60k MBA data points, for which we have ground truth.
- We apply this methodology to $1.5M$ Ookla and M-Lab speed test measurements in four US. cities and show that the majority of data points originate from lower subscription tiers, thereby skewing throughput results.
- We quantify the impact of access type, WiFi characteristics, device memory, and time of day on Ookla measurements.
- We quantify the performance difference of Ookla versus M-Lab measurements for the same subscription tiers, cities, and ISPs that stem from the differing measurement methodologies.
- Based on our results, we put forth a set of recommendations for speed test vendors and the FCC to contextualize speed test data and correctly interpret measured performance.

5.2 A Motivating Example

We begin by illustrating the challenge and inaccuracy of interpreting crowdsourced active measurement (e.g. "speed test") data at face value. We base our initial analysis on 745k Ookla measurements from the primary fixed broadband ISP in four major US. cities during 2021. The median download speed of each of these four cities is roughly 115 Mbps. In

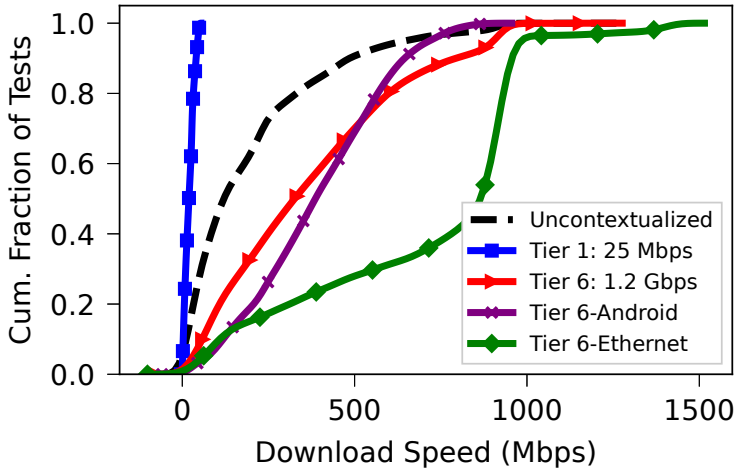


Figure 5.1: Comparison of raw speed test download speed distributions in a major US. city. The "Uncontextualized" line represents our starting point. The other lines represent the original data contextualized with subscription tier, access link speed or type, and/or device type.

prior work, a similar analysis, emphasizing the median value of the aggregated tests, was used to study the regional Internet quality of a congressional district in New York [131]. Based on these median performance results, the report recommended regions for Internet buildout and funding allocations to improve Internet quality in the constituency.

However, as this chapter will illustrate, the lack of context for these measurements prevents proper interpretation of such aggregate results. Figure 5.1 presents the distributions of the download speed in City-A disaggregated by subscription plan tiers, access speed or link type, and measurement device type. The "uncontextualized" line represents the original data without context applied. The figure shows that the median download speed of the lowest (slowest) subscription tier (Tier 1, with a maximum download speed of 25 Mbps) is 19.22 Mbps, almost six times as slow as the overall City-A median download speed. City-A's median download speed, on the other hand, is nearly four times less than the premium ISP subscription tier (Tier 6: 1.2 Gbps) and almost seven times less than that recorded by test takers on Tier 6 Ethernet connections (Tier 6: Ethernet).

Table 5.1: Number of measurements for datasets utilized in this work. Note that for Ookla and M-Lab, the data points are from each city, whereas for MBA, the data points are from the state that corresponds to each city.

City/State	ISP	Ookla	M-Lab	MBA
A	1	214 k	113 k	25.9 k
B	2	205 k	376 k	14.9 k
C	3	128 k	64 k	10.9 k
D	4	198 k	166 k	8.9 k

Similarly, for speed tests that do not experience local bottlenecks (tests whose performance is constrained by local WiFi factors such as WiFi band and RSSI), the median download speed of the highest subscription tier for this group of speed tests (Tier 6: Android) is almost four times more than the City-A median download speed. Still, the median for the group of tests not affected by local bottleneck factors is half the Tier 6 (Ethernet) median download speed rate.

In the remainder of this chapter, we describe the contributions that enable us to contextualize each measurement point with broadband plan subscription tier, local network characteristics, device context, test time, and speed test vendor. In so doing, we demonstrate that the ability to contextualize speed test measurements is critical for interpreting the quality of the Internet in a region.

5.3 Datasets

This section describes the three primary datasets we utilize for this work. Table 5.1 summarizes the number of data points of each type. We choose Ookla's Speedtest® (obtained from the Speedtest Intelligence® portal) as it is the largest Internet measurement vendor that is capable of measuring available bandwidth with high accuracy [223]. M-Lab's Speed Test, on the other hand, makes collected data publicly available. We utilize

the Measuring Broadband America (MBA) dataset because it provides the subscriber's purchased broadband plan information with the speed test measurements.

We use the Ookla and M-Lab data collected from January 1 – December 31, 2021. MBA data is also from this period but lacks data from September 1 – October 31 (this data is unavailable from the MBA website).

5.3.1 Ookla's Speedtest

Ookla's Speedtest[1] (data provided through Ookla's Speedtest Intelligence®) possesses over 16k measurement servers worldwide [99] and allows users to assess the quality of their Internet connection using either a web-based portal or native mobile application [93]. For each Speedtest, a nearby test server is selected and *multiple TCP connections* are used to calculate the throughput of the connection. Ookla's Speedtest Intelligence dataset contains individual Speedtest measurements that include QoS metrics (up/down throughput, latency, packet loss, jitter), as well as meta-features such as ISP, device type, and access type. Ookla provides performance data aggregated over time and space to the public [94].

A Data Usage Agreement (DUA) with Ookla provides us access to over 745k individual Speedtest measurements from four major metropolitan cities in the US, which we use for this study. Each of these cities has a population in the range of 400,000 – 700,000. For each city, we utilize the FCC Form 477 dataset [80] to identify the dominant ISP and conduct our analysis. Specifically, we use this dataset to compute the number of census blocks served by an ISP in a city and pick the one that covers the highest number of blocks.

The Ookla dataset tags the origin of each test, specifying whether the test was initiated through a web-based portal or a native application. The web-based tests do not provide device-related information. On the other hand, the native application dataset

[1]http://speedtest.net

indicates the type of device that started each measurement (Android, iOS, or desktop). 394k of the measurement points in our dataset originated from native applications. The dataset also contains critical metadata related to the wireless link for Android devices, such as frequency band, signal strength, maximum achievable theoretical downlink throughput, and available kernel memory. These metrics are essential in contextualizing the measurements, as we will show in section 5.6.

5.3.2 M-Lab's Speed Test

Section 4.2.1 in Chapter 4 provides an overview of the working mechanism of M-Lab's Speed Test[2] (note the different spelling and capitalization from Ookla's Speedtest) that reports network performance metrics using the Network Diagnostic Tool (NDT). We extracted 717k NDT measurements from the same four major US. cities in 2021 for the same major residential broadband ISPs as Ookla. Because NDT measurements do not associate an upload speed test with a download speed test initiated by the same client, we adopt a similar methodology to [219]. We compute a 120 second window for every download speed test and filter all upload speed tests issued from the same client and server IP address. If a single upload speed is captured during that window, we associate it with the download speed. In the event we observe more than one upload speed test started during this time frame that meets this criterion, we associate the earliest upload speed test with the download speed test. As a result, *our methodology enables us to compare Ookla and M-Lab measurements over the same period, in the same cities, for the same service provider.*

[2]https://speed.measurementlab.net/#/

5.3.3 Measuring Broadband America

Measuring Broadband America (MBA) [87] is an FCC-sponsored project that uses specialized hardware test units [101] to collect Internet measurement data from 4,000 US. households. These units measure and report upload and download speed multiple times per day [88]. Each device in the dataset also reports its location (at the granularity of census tract). Most critically, this dataset is generated from wired devices and contains the broadband plan subscription of the user hosting the device. Wired devices provide measurement data of the access link without confounding WiFi performance, while the broadband plan data provides ground-truth for our methodology to determine broadband subscription tier. We utilize the latest subscriber information, which was collected in 2020, for the measurements [100].

5.3.4 Ethics

While our work analyzes speed tests from users of two prominent speed test vendors, our work is not human subjects research. The private dataset shared by Ookla under DUA is fully anonymized, and we cannot identify the individual users of the platform. For the subset of measurements from devices with GPS geolocation enabled, Ookla provides GPS coordinates truncated after three decimal points. Such geolocation is accurate to 111 metres; therefore, we cannot associate it with any user/residence. The M-Lab dataset provides only public IP addresses that one can localize using IP geolocation tools. However, IP geolocation errors can exceed 30 KM, making it difficult to isolate specific users/homes. We also obtained the street address dataset from Zillow under a DUA. We do not have methods to identify residents, selected broadband subscription tiers, or the actual speed test performance at any address.

5.4 Determining Subscription Tiers

Our first step in contextualizing speed test data is to determine the home broadband subscription tier of the user from which the measurement originates. This step is critical because it provides context for the achieved download and upload speed; with information about the theoretical maximum speeds (the "plan" speeds), we can first determine whether a speed test measurement indicates the network is under-performing. Without this information, we may attribute a slow download speed to the under-performance of the access link instead of a lower ("slower") tier plan purchased by the user.

To determine the subscription tier, we must first obtain the residential broadband plans available at the location of the speed test so that we know the set of possible plans from which to select. As described in this section, we obtain this information by modifying a prior approach. Then, we apply our Broadband Subscription Tier (BST) methodology, a novel two-stage hierarchical unsupervised clustering technique that matches each ¡download speed, upload speed¿ measurement tuple to a specific subscription plan.[3] To evaluate the efficacy of BST, we utilize the MBA dataset as it provides both the speed test measurements and subscription tier information for more than 60k data points.

Challenges. There exist two significant challenges in associating crowdsourced measurements with subscription tier information. First, no dataset exists in the public domain that details all the broadband plan choices offered by ISPs to users at the granularity of street address, census block, or even census block group. Through its Form 477 [80], the FCC only provides the ISP-reported maximum download/upload speed in a census block. Unfortunately, it is impossible to associate measurements with subscription tiers without a complete picture of all the plans available from the ISPs. Second, crowdsourced mea-

[3]We use the terms "subscription tier" and "subscription plan" interchangeably.

surement results are inherently noisy, as they are vulnerable to environmental factors that range from poor WiFi router positioning to device memory, as shown in section 5.6.3. As such, it is crucial to understand the variability between different metrics reported through speed tests prior to assigning a measurement to a subscription tier.

5.4.1 Observations

To obtain the set of ISP-offered subscription plan choices, we modify the tool proposed in [179]. In particular, we augment the tool to collect available download/upload speed plans for major residential ISPs at specific US. street addresses. Our tool requires clean and well-formatted street addresses to obtain this information. Hence we utilize the residential property address dataset from Zillow [105] to create an address set for each of the four cities in our study. Then, we randomly select 100K residential addresses for each city and collect the ISP-offered plans. To prevent overloading ISP infrastructure, we carefully limit the number of queries we make per ISP. Our analysis of street-address level broadband plan choices in four cities reveals two significant trends.

The first trend we observe is that the plan choices remain unchanged across different street addresses within a city. For example, ISP-A offers six plans for all street addresses in City-A. Three of these plans have different download speeds (25 Mbps, 100 Mbps, and 200 Mbps) but the same upload speed (5 Mbps). The other three plans have different, faster download speeds (400 Mbps, 800 Mbps, and 1200 Mbps) with upload speeds of 10 Mbps, 15 Mbps, and 35 Mbps, respectively. We observe similar types of tiered offered plans that do not vary based on the specific address for the other three cities and major ISPs.

Second, although an ISP offers diverse plans for download speeds, varying in both number and speed range, the set of maximum available upload speeds is much smaller. Further, the upload speeds are much slower than available download speeds. This ob-

servation is noteworthy because, as discussed in [153], many factors, such as local home network conditions (e.g., WiFi interference or local congestion) and web-browser limitations, could prevent a speed test measurement from attaining high throughput. On the other hand, given the lower maximum upload speeds, fewer factors can limit the attainment of the maximum speeds [220].

As a result, crowdsourced measurements from individual users should exhibit less variation (and more consistency) in upload speed compared to download speed. Given this intuition about upload speed, we should expect to see that the recorded upload speeds during multiple measurements for a single user are more consistent than the set of download speeds for that user. To capture this per-user performance consistency, we calculate a consistency factor by taking the ratio of the mean and 95^{th} percentile for the sets of upload and download speeds recorded over multiple tests by the same user [214]. The closer the consistency factor is to 1, the greater the consistency for the evaluated metric over the set of tests from a single user.

Concretely, we select measurements from any Ookla user who conducted at least five tests using the native application while connected to the WiFi network [214]. In total, 23k (out of 85k) users issued more than five tests. These users contribute 80k measurements, about 70% of total measurements from native applications. For brevity, we present the results only for City-A. We base our analysis only on native app users because a public IP address identifies users of web-based tests. Given the prevalence of NAT employed by the ISPs, determining which group of tests belongs to an individual user based on the public IP address is highly challenging.

Figure 5.2 shows the CDF of the consistency factor of measurements from users who registered at least five tests using Ookla's native iOS application (we present only the iOS result for clarity and confirm that we observed similar trends for data from Android and desktop native applications). As shown in the figure, download speed variations

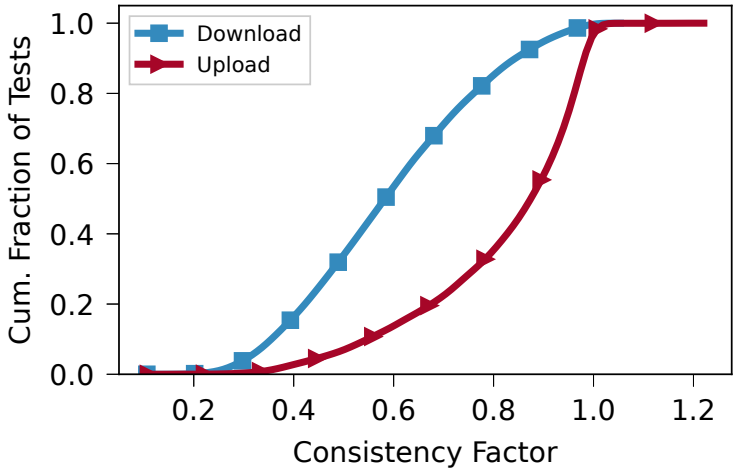


Figure 5.2: CDF of consistency factor for all iOS users who recorded at least five tests.

are much more significant than upload speed; upload speed is more consistent across all users. The median consistency factor for download speed is 0.58, compared to 0.87 for upload speed. The more consistent behavior of upload speed performance indicates the possibility of utilizing this metric to determine the subscription tier for each speed test. We confirm our observations of upload speed consistency for the other three cities. Note that while we report the mean value, we do observe that the consistency factor exceeds one for some users. The mean value of a (heavy-tailed) distribution can be skewed by larger items in the tail portion of the distribution.

Combining these two observations, *we hypothesize that we can utilize the measured upload speeds of the speed tests to identify the subset of possible subscription plans from which any given speed test originates.* In the next section, we describe our Broadband Subscription Tier methodology, which is our approach to matching speed test measurements to their corresponding subscription plan.

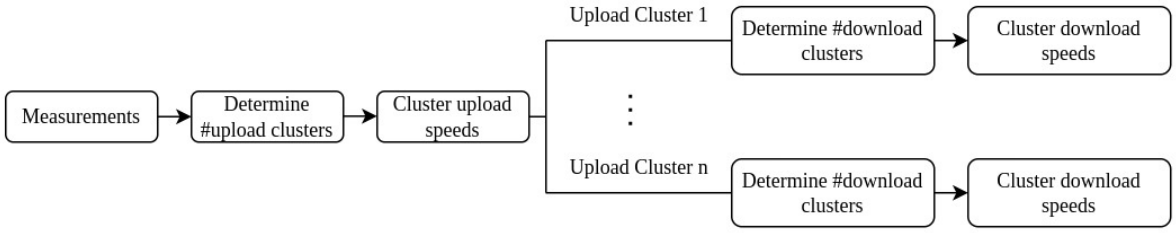


Figure 5.3: Broadband Subscription Tier methodology.

5.4.2 BST methodology

We propose a two-stage hierarchical unsupervised clustering methodology to match each $< download\ speed, upload\ speed >$ measurement tuple to a specific ISP subscription plan. In the first stage, our objective is to associate the recorded upload speed of a speed test to a cluster that corresponds to the correct ISP-offered upload speed. Because multiple plans might offer the same upload speed, in the second stage we use our first stage clustering to perform an inter-cluster analysis to identify the set of individual subscription tiers to which a recorded download speed can potentially match. Combining the two stages yields a probabilistic model that can map the results of speed test measurements to their respective subscription classes/tiers. Figure 5.3 gives an overview of our methodology.

For a given speed test dataset in a city, each of our two stages begins by first confirming the presence of clusters within the upload/download speed distribution. Taking the example of the first stage, we start by employing a Kernel Density Estimation (KDE) [86] method with multivariate Gaussian kernel functions to estimate the probability densities of the upload speeds recorded during the speed tests. Combining these multiple kernel functions results in a smooth function that produces clusters containing the upload speed densities. This stage checks whether the number of upload/download speeds offered by an ISP matches the number of clusters formed in the distribution of crowdsourced measurements.

Table 5.2: BST upload speed selection accuracy for the four states in the MBA dataset.

State	ISP	#Units	Accuracy
A	1	20	99.33%
B	2	17	98.19%
C	3	10	96.84%
D	4	11	99.10%

After determining the number of clusters using the KDE method, we cluster the upload speeds by employing the Gaussian Mixture Model (GMM) [90] to determine the upload speed of the subscription tier. Once a measurement is associated with a cluster of upload speed, we enter the second stage, where we re-apply GMM to determine the corresponding download speed cluster. Note here that we possess the information about the mapping between different offered download and upload speeds through the mechanism described in section 5.4.1.

We choose GMM because it is one of the most popular unsupervised clustering techniques employed on a distribution consisting of several components of Gaussian densities. In GMM, each cluster follows a Gaussian distribution, and the eventual goal is to assign measurements to different parts by estimating each cluster's parameters. The parameters associated with a GMM cluster/component include the mean, covariance matrix, and weight. As such, compared to other clustering methodologies such as K-Means, GMM is a probabilistic model that considers the clusters' variance in addition to the means. In each stage, we employ GMM in conjunction with the Expectation-Maximization (EM) [75] methodology (GMM-EM) to iteratively compute the maximum likelihood that each speed test data point belongs to its respective upload/download speed cluster.

5.4.3 Evaluation with MBA dataset

We leverage the MBA dataset to evaluate the efficacy of our BST methodology. This dataset contains not only active measurements collected hourly but also subscription information. We apply our BST methodology to 60k measurements in this dataset spanning the four states associated with the four cities in our study. We compare the result of BST with the ground truth subscription information available in the MBA dataset by calculating the $accuracy = (\frac{\#correctly\ associated\ measurements}{\#total\ measurements})$. Table 5.2 presents the total number of units and the corresponding accuracy achieved by the BST methodology for upload speeds. For all states, accuracy is above 96%; accuracy is above 99% for two states.

As a descriptive example, we provide a detailed explanation of the application of the BST methodology to the MBA dataset in State-A, where ISP-A is the dominant residential Internet service provider. Table 5.2 shows that 20 measurement units subscribe to ISP-A in this state. These units record a total of $25,927$ measurements during 2021. The plans recorded for the MBA subscribers in State-A are similar to the offered plans described previously for City-A. However, there are no records of the 25 Mbps download (5 Mbps upload) subscription plan in the MBA-State-A dataset. This observation is important when we match subscription plans to measurements in the following example.

Upload Speed Subscription Tiers. We begin by applying KDE on the set of upload speeds measured by the MBA nodes in State-A; Figure 5.4 presents the result (Figure 5.5 show the results for the other three cities). There are four significant clusters of upload speed densities in this dataset. The distinct peaks of upload speed densities in the regions of the offered upload speeds by ISP-A indicate the possibility of identifying the subscription plan of a given measurement.

After determining the number of clusters, we aim to assign each measurement point

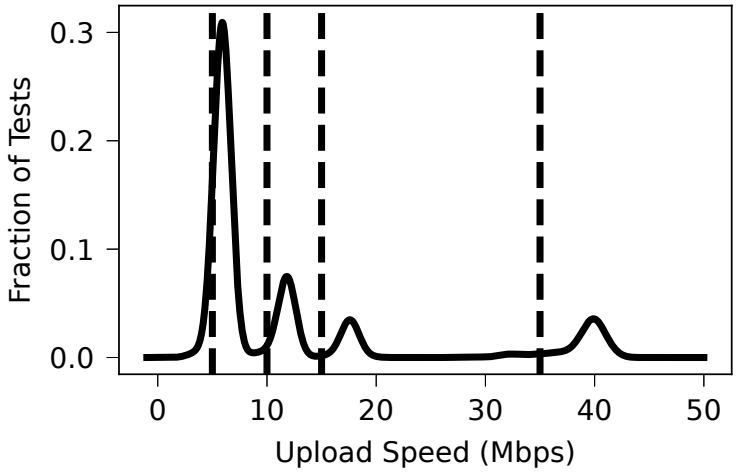
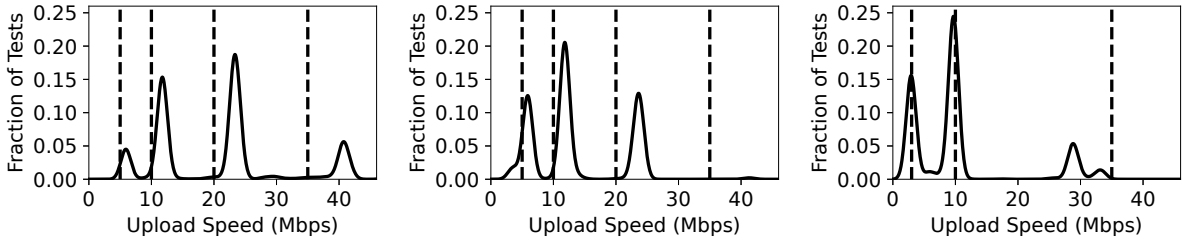


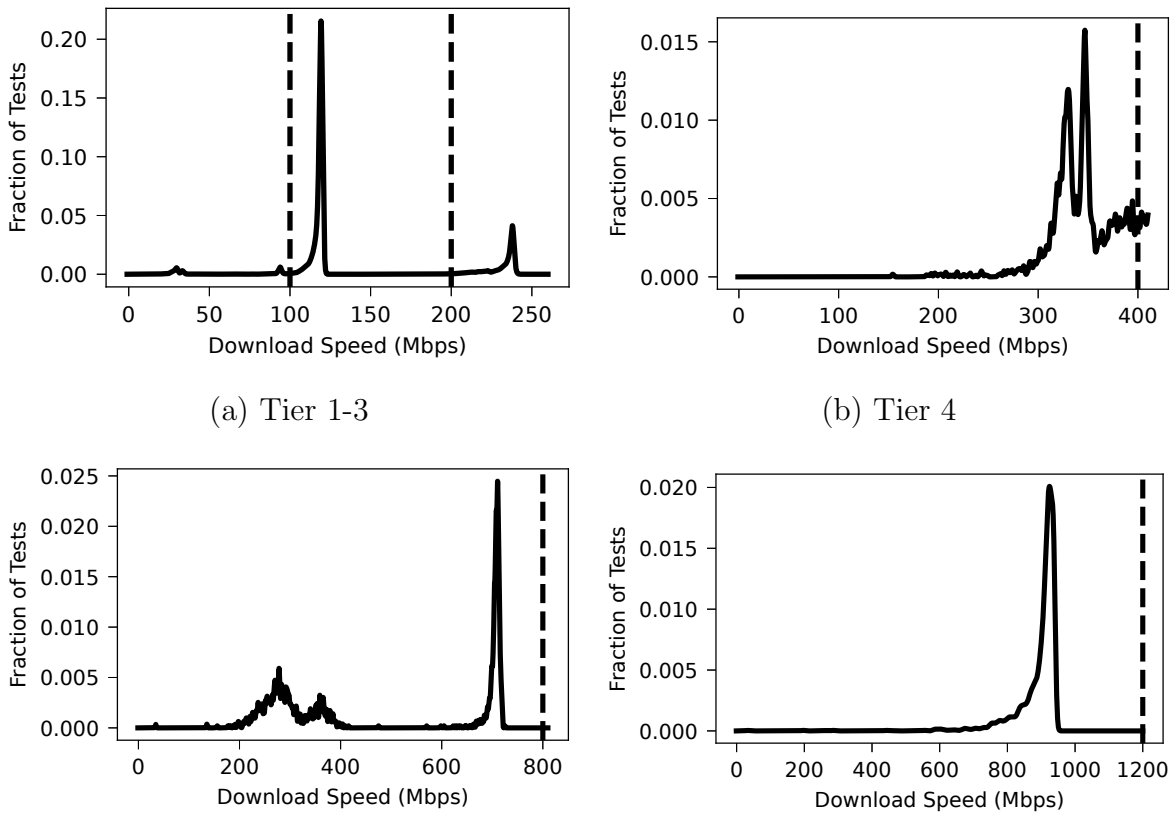
Figure 5.4: Upload speed density using KDE method on MBA State-A dataset. The vertical lines are the upload speed plans offered by ISP-A.



(a) State-B (b) State-C (c) State-D

Figure 5.5: Upload speed density using KDE method on MBA dataset for States B-D. The vertical lines are the upload speed plans offered by the dominant ISP in each state.

to the appropriate subscription tier by first using the recorded upload speed. To do so, we employ the BST methodology to detect the clusters of the upload speed recorded by the MBA units. The methodology converged after 20 iterations. The means of the four upload speed clusters were 5.87 Mbps, 11.55 Mbps, 17.57 Mbps, and 38.62 Mbps. We observe that the upload cluster means obtained through the BST methodology are close to the actual offered upload speeds by ISP-A. BST achieves an accuracy of 99.3% for this set of upload speed measurements. This result validates our hypothesis and demonstrates the ability to use upload speed to narrow down potential subscription plans from which



(a) Tier 1-3

(b) Tier 4

(c) Tier 5

(d) Tier 6

Figure 5.6: Download speed density using the KDE method within each cluster of upload speed. Black vertical lines represent the corresponding download speed plans for each upload speed.

a given speed test may originate.

Download Speed Subscription Tiers. After determining the upload speed cluster of the speed test measurements, we apply the BST methodology within each of the four clusters of upload speed. Figure 5.6 shows the clusters of download speeds present within the upload speed clusters identified in the previous step.

Tier 1-3: This cluster consists of measurements from users subscribed to the 5 Mbps upload speed. Within this tier, we label the three available download speed plans as Tier 1, Tier 2, and Tier 3 to refer to the offered 25 Mbps, 100 Mbps, and 200 Mbps

download speeds, respectively. Because the MBA dataset does not have the 25 Mbps download plan, our analysis consists of Tiers 2 and 3. There are 15,781 measurements total from Tiers 2 and 3 in the MBA-State-A dataset. From Figure 5.6(a), we see two major download speed peaks after applying the KDE method to the download speeds in this cluster.

After determining the number of clusters, we apply the BST methodology to attach each download speed measurement point to the appropriate subscription tier class. The means of the two clusters found by BST are 110.89 Mbps and 231.69 Mbps, which are greater than the advertised download speeds. This observation indicates that ISP-A provides performance that surpasses the subscribed download speed for these subscription tiers. Previous studies [220] observed similar ISP behavior in the past. In comparing our calculated download speed plan with the ground truth, we determine that our methodology can accurately identify 100% of the download speed measurements in this cluster.

Tier 4: There are 4,185 measurements in the MBA-State-A dataset that belong to this subscription cluster. The upload speed in this cluster is 10 Mbps; only one plan offers this upload speed, with a 400 Mbps download speed. Though our methodology achieves 100% accuracy in determining the subscription tier of these measurements, the KDE method reveals several download speed peaks within this cluster (see Figure 5.6(b)). We apply the BST methodology to detect four download speed clusters. The four means obtained through the process are 333.48 Mbps, 335.15 Mbps, 400.37 Mbps, and 463.31 Mbps. While it is unclear why four clusters are detected, it could be due to ISP throttling. It remains future work to diagnose the exact cause.

Tier 5: There are 2,453 measurements in this cluster, and the offered upload speed is 15 Mbps. This tier offers a download speed of 800 Mbps. Like Tier 4, BST achieves 100% accuracy in determining this subscription tier. Figure 5.6(c) shows a peak at around

700 Mbps, closer to offered speed, with the KDE method. We also observe multiple peaks around 300 Mbps and 400 Mbps. The BST methodology detects three clusters of download speed with means 269.98 Mbps, 358.06 Mbps, and 705.35 Mbps. We observe an overlap in download speed tier means between tiers 4 and 5. However, the proposed BST methodology isolates the download speeds into their respective subscription tiers.

Tier 6: ISP-A offers a plan with download speed 1200 Mbps and 35 Mbps upload speed; BST achieves 100% accuracy in inferring this subscription tier. In State-A, there are $3,508$ measurements in this subscription tier. Figure 5.6(d) shows a single major cluster of download speed after applying the KDE method. The BST methodology computes the mean of this download speed cluster to be 892.05 Mbps. This mean value is much lower than the offered download speed for this subscription class. This result shows the limitation of speed test-like measurements in saturating the available bandwidth in the higher end of the offered subscription plans. Previous work [153] made similar observations.

These promising results indicate the ability to infer subscription tier information for crowdsourced speed tests. In the following sections, we use the BST methodology to contextualize Ookla and M-Lab speed test measurements with subscription tier information.

5.5 Augmenting Ookla & M-Lab Data

Now that we have demonstrated the accuracy of our BST methodology, our next step is to apply our approach to contextualize crowdsourced speed test measurements. This step is critical to the interpretation of speed test data; by comparing speed test results to the subscribed broadband plan, we can gain insight into whether the network is underperforming. Our analysis in this section focuses on City-A.

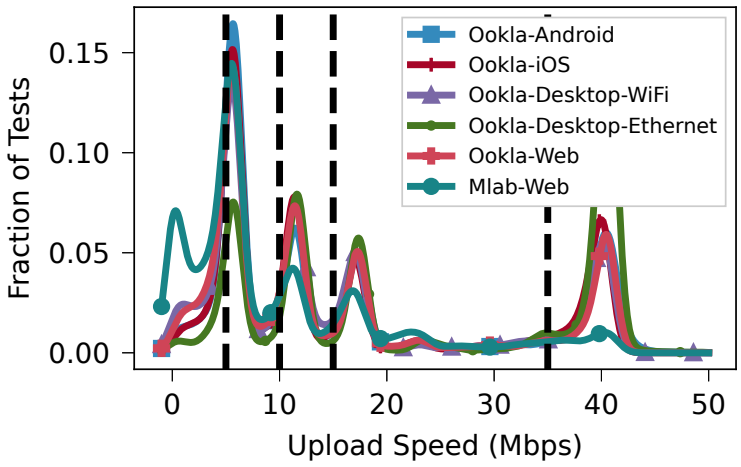
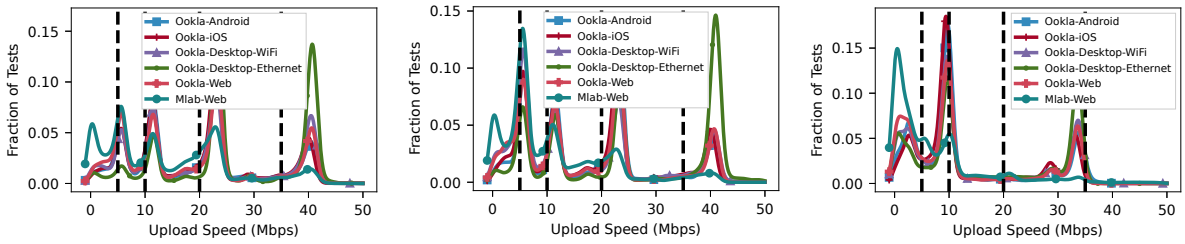


Figure 5.7: Upload speed density using the KDE method on City-A speed test measurements. The vertical lines represent the offered upload speed in each ISP-A plan.



(a) City-B (b) City-C (c) City-D

Figure 5.8: Upload speed density using the KDE method on Cities B-D speed test measurements. The vertical lines represent the offered upload speed of the dominant ISP in each city.

5.5.1 Contextualization with Subscription Plans

Upload Speed Subscription Tiers. The measurement nodes for the MBA project collect data directly from the cable modems, increasing the accuracy of capturing the access network performance. Unfortunately, a significant fraction of the speed test measurements in the Ookla and M-Lab datasets stem from end-user devices connected through a first hop WiFi link. The introduction of this single wireless link can significantly impact the speed test performance and can introduce additional skews [153, 214]. However, given

Table 5.3: Number of measurements and the means (Mbps) for upload speed clusters that form near the ISP-A offered upload speeds in City-A. For each dataset, the means are obtained using the BST methodology.

Platform	Type	Tier 1-3		Tier 4		Tier 5		Tier 6	
		#Measurements	Mean	#Measurements	Mean	#Measurements	Mean	#Measurements	Mean
Ookla	Android-App	8,890	5.25	3,088	11.29	2,810	17.04	5,152	40.23
	iOS-App	33,265	5.30	13,299	11.35	9,530	16.71	19,480	39.82
	Desktop WiFi-App	4,551	5.54	1,377	11.59	3,638	16.82	1,750	39.92
	Desktop Ethernet-App	1,031	5.69	746	11.65	1,400	16.95	2,098	40.13
	Net-Web	43,833	5.72	12,802	11.64	29,157	16.69	15,797	40.06
M-Lab	NDT-Web	70,789	5.32	17,014	10.74	16,417	16.71	9,490	39.94

Table 5.4: Number of measurements and the means (Mbps) for upload speed clusters that form near the ISP B offered upload speeds in City B. For each dataset, the means are obtained using the BST methodology.

Platform	Type	Tier 1-2		Tier 3		Tier 4-5		Tier 6	
		#Measurements	Mean	#Measurements	Mean	#Measurements	Mean	#Measurements	Mean
Ookla	Android-App	4965	5.73	2483	11.54	6794	22.42	2819	39.21
	iOS-App	18940	5.81	11358	11.48	29960	21.95	15042	38.08
	Desktop WiFi-App	2012	5.1	1281	11.48	3009	21.97	2093	39.01
	Desktop Ethernet-App	492	5.63	811	11.39	2048	23.32	2904	36.87
	Net-Web	30132	5.38	11925	11.56	37553	22.37	17504	39.62
Mlab	NDT-Web	144345	5.44	63805	11.16	135897	22.04	25553	39.23

the small range of possible maximum upload speeds, we hypothesize that it should still be possible to cluster these crowdsourced active measurements based on the recorded upload speed.

Figure 5.7 shows the upload speed densities for speed test takers who accessed Ookla as well as M-Lab tests run through the web-based portal (the rest of the cities are depicted in Figure 5.8). Similar to the peaks in the MBA data shown in Figure 5.4, we observe densities of upload speed in the crowdsourced measurements that peak near the ISP-A offered upload speeds for all datasets. In addition to the four major peaks, there is an additional upload speed cluster in the 1 Mbps region in the M-Lab data.

We apply the BST methodology to associate the upload speed measurements to the four peaks around the ISP-A-provided upload speeds. Table 5.3 presents the number of measurements and means for the upload speed clusters (corresponding to an ISP subscription upload speed tier) detected by the BST methodology, broken down by device

Table 5.5: Number of measurements and the means (Mbps) for upload speed clusters that form near the ISP C offered upload speeds in City C. For each dataset, the means are obtained using the BST methodology.

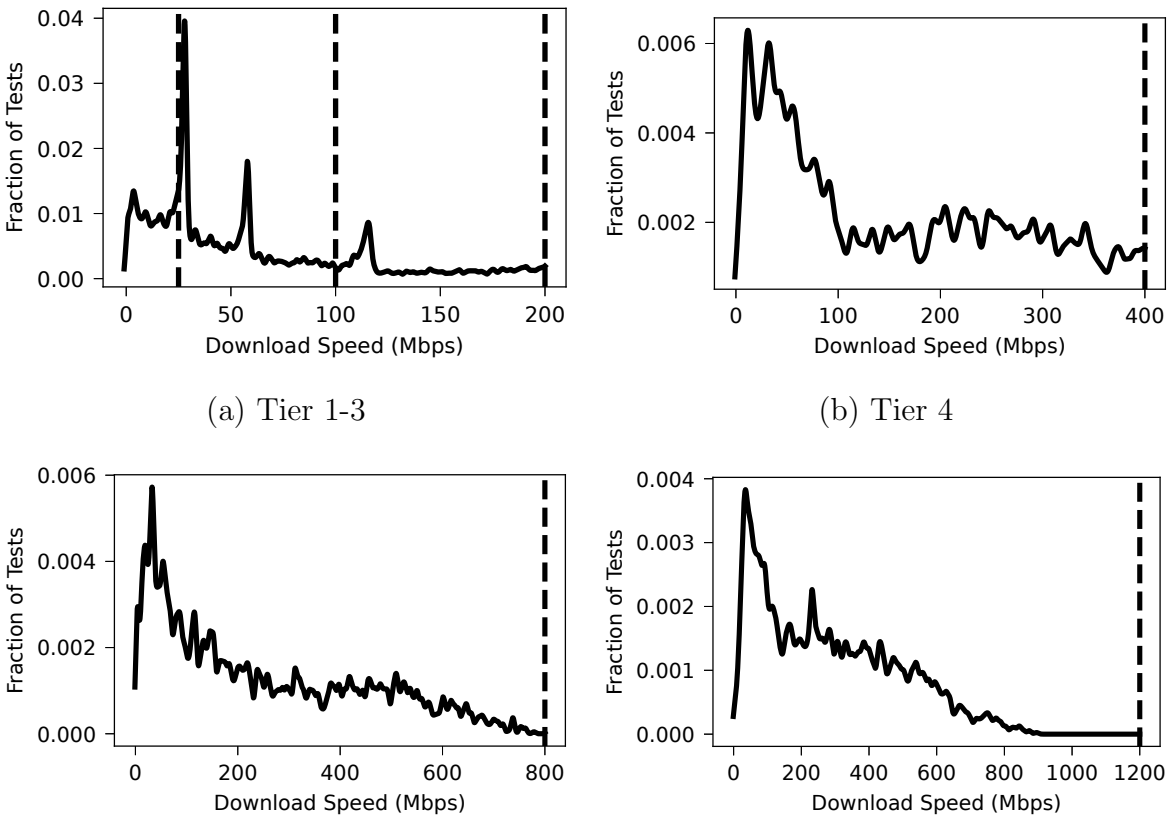
Platform	Type	Tier 1-3		Tier 4-5		Tier 6-7		Tier 8	
		#Measurements	Mean	#Measurements	Mean	#Measurements	Mean	#Measurements	Mean
Ookla	Android-App	6766	5.28	3168	11.53	8307	22.28	3030	39.49
	iOS-App	11725	5.18	4711	11.45	12322	21.96	5579	38.84
	Desktop WiFi-App	2015	4.86	606	11.47	1094	21.61	854	38.21
	Desktop Ethernet-App	1020	4.92	628	11.48	868	23.36	2416	37.71
	Net-Web	24148	4.89	7982	11.54	21478	22.02	9697	39.53
Mlab	NDT-Web	34523	4.76	12789	10.72	13041	19.82	4416	35.47

Table 5.6: Number of measurements and the means (Mbps) for upload speed clusters that form near the ISP D offered upload speeds in City D. For each dataset, the means are obtained using the BST methodology.

Platform	Type	Tier 1-2		Tier 3-4		Tier 5	
		#Measurements	Mean	#Measurements	Mean	#Measurements	Mean
Ookla	Android-App	7244	3.51	8142	9.73	6462	28.69
	iOS-App	18598	3.72	26699	9.39	19177	28.03
	Desktop WiFi-App	2525	3.04	2233	9.59	2348	28.72
	Desktop Ethernet-App	1845	3.6	1716	9.68	3096	28.99
	Net-Web	40452	3.05	29642	9.7	27517	28.51
Mlab	NDT-Web	71833	2.95	61435	7.6	24541	24.94

type when possible (Tables 5.4 – 5.6 in the appendix present the same breakdown for Cities B-D). We observe the means of each cluster to be similar across all datasets. These means are also consistent with the means detected in the State-A dataset in section 5.4.3 for ISP-A offered plans. Given this similarity, we can associate the crowdsourced measurements to their subscription tier.

Download Speed Subscription Tiers. The much larger download speed plans offered by ISP-A and the performance variability caused by the end user's home wireless link create considerable challenges to clustering the measured download speeds. Figure 5.9 shows the densities of download speeds recorded by Ookla tests conducted on Android devices within each cluster of upload speed (the remaining three locations are depicted in Figures 5.10 - Figure 5.12).



(a) Tier 1-3

(b) Tier 4

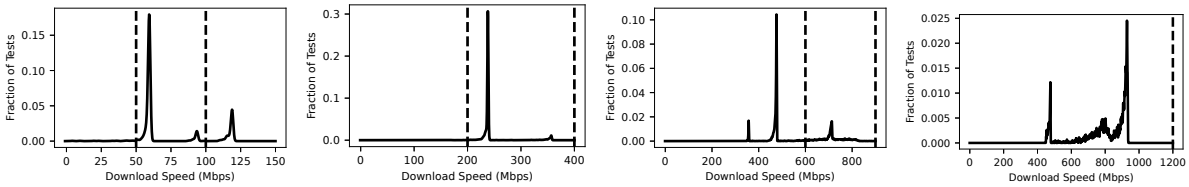
(c) Tier 5

(d) Tier 6

Figure 5.9: Download speed density using the KDE method within each upload speed cluster of Ookla Android device measurements.

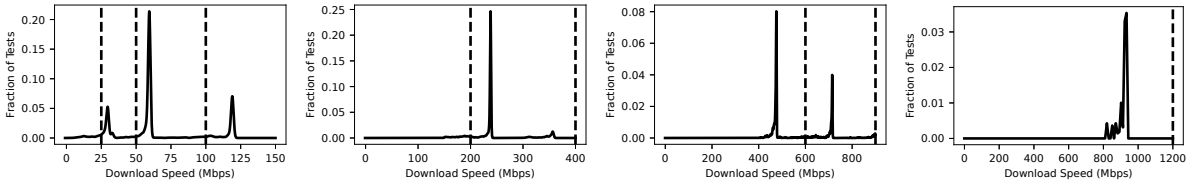
There are five major download speed clusters in Tiers 1-3 of the Android dataset[4]. This number is three more than the number detected in the MBA State-A dataset for the same cluster and two more than what is offered by ISP-A for this subscription tier. After applying the BST methodology, we associate the download speed measurements to five clusters of download speed with means 8.04 Mbps, 27.14 Mbps, 57.85 Mbps, 115.65 Mbps and 214.01 Mbps. We associate the measurement points that belong to the components with mean values of 8.04 Mbps and 27.14 Mbps to Tier 1 as these measurements are close to the offered download speed. Similarly, we assign the measurements associated with

[4]All Android measurements occur over WiFi.



(a) Tier 1-2 (b) Tier 3 (c) Tier 4-5 (d) Tier 6

Figure 5.10: Download speed density using KDE method within each cluster of upload speed in State-B. Black vertical lines represent the corresponding download speed plans offered for each upload speed.



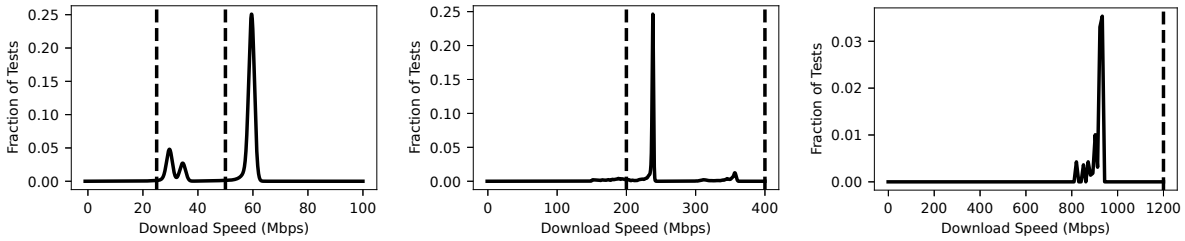
(a) Tier 1-3 (b) Tier 4-5 (c) Tier 6-7 (d) Tier 8

Figure 5.11: Download speed density using KDE method within each cluster of upload speed in State-C. Black vertical lines represent the corresponding download speed plans offered for each upload speed.

clusters of mean values 57.85 Mbps and 115.65 Mbps to Tier 2. Finally, we associate measurements in the cluster of mean 214.01 Mbps to Tier 3.

Compared to the clusters formed by tests conducted over WiFi access links, the measurements in Tier 1-3 run by desktop devices connected with wired links (presented in Table 5.7) produce three download speed clusters with means of 16.04 Mbps, 93.76 Mbps and 231.44 Mbps. These three means are closer to the three offered download speeds provided by ISP-A for this subscription tier.

We know that ISP-A offers a single download speed for each of the other upload speed tiers. However, Figure 5.9 indicates a large number of download speed clusters at various magnitudes. We apply the BST methodology and associate measurements with 10 clusters of download speed for each of tiers 4-6. Table 5.7 presents the download speed



(a) Tier 1-2 (b) Tier 3-4 (c) Tier 5

Figure 5.12: Download speed density using KDE method within each cluster of upload speed in State-D. Black vertical lines represent the corresponding download speed plans offered for each upload speed.

Table 5.7: Download speed means (Mbps) for each subscription tier in City-A. For each dataset, the means are obtained using the BST methodology.

Platform	Type	Tier 1	Tier 2	Tier 3	Tier 4	Tier 5	Tier 6
Ookla	Android-App	8, 27	58, 116	214	21, 53, 93, 152, 212, 268, 327, 390, 445, 599	27, 73, 139, 219, 309, 403, 489, 574, 672, 879	40, 91, 160, 232, 304, 381, 461, 550, 636, 763
	iOS-App	9, 28	55, 84, 113	155, 197, 226	25, 57, 95, 144, 196, 244, 289, 337, 389, 442	28, 73, 121, 193, 264, 339, 421, 502, 589, 693	37, 88, 152, 223, 295, 367, 447, 535, 624, 737
	Desktop WiFi-App	15, 27	53, 86, 113	154, 202, 227	34, 77, 117, 155, 193, 251, 302, 340, 408, 453	22, 59, 105, 156, 211, 268, 345, 444, 540, 714	71, 177, 251, 345, 436, 540, 644, 735, 889, 1328
	Desktop Ethernet-App	16	94	231	68, 288, 461	147, 596	104, 907
	Net-Web	7, 28	55, 85, 114	170, 225	23, 55, 92, 146, 204, 265, 336, 405, 458, 637	19, 54, 97, 166, 239, 333, 437, 543, 692, 884	66, 162, 251, 350, 458, 568, 692, 820, 913, 1299
M-Lab	NDT-Web	6, 25, 47	100, 164, 221	18, 53, 84, 135, 196, 258, 337, 422, 569, 852	18, 53, 84, 135, 196, 258, 337, 422, 569, 852	22, 60, 105, 165, 229, 325, 413, 501, 652, 868	31, 93, 183, 260, 342, 429, 507, 610, 732, 892

cluster mean values that belong to each upload speed cluster. The number of components detected for wired measurements in each of these tiers is less than in wireless ones.

For Tier 4, we observe three clusters with mean values of 67.77 Mbps, 288.29 Mbps, and 461.18 Mbps. For Tier 5, we identify two groups with mean values of 146.46 Mbps and 595.59 Mbps. We also observe two clusters for Tier 6, with mean values of 103.96 Mbps and 906.87 Mbps. The wide range of values represented by these download speed clusters means, as well as for WiFi tests, indicates a significant variance in the results of the speed tests. This result further justifies our approach of first clustering these measurements using the less noisy, slower upload speeds before associating the measurements with complete subscription tier information.

WiFi-connected devices contribute to almost 97% of the native application tests in the Ookla dataset. Roughly half of these tests originate from the lowest subscription tier. As a result, if we take any aggregate (such as the median) of speed test data in a locality, we would, at best, get a representation of the Internet quality obtained by

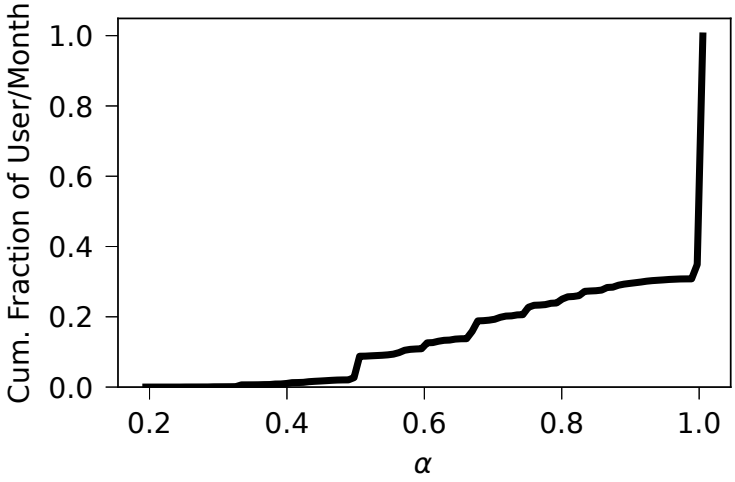


Figure 5.13: CDF of α values per user per month.

the lower subscription tiers, as opposed to the complete picture. Contextualizing these measurements with subscription tier information is crucial before making any general assessment of the Internet quality in a region.

5.5.2 Investigation of Consistency

Because we lack ground truth for the Ookla and M-Lab measurements, we turn to other approaches to evaluate the accuracy of our BST methodology in these noisy environments. In this section, we analyze the consistency of BST in its association of speed test measurements with subscription tiers. To do so, we focus on users who conducted more than five speed tests in a month, and we examine whether each measurement from a single user is assigned to the same subscription plan, or whether there is variability in the assignment.

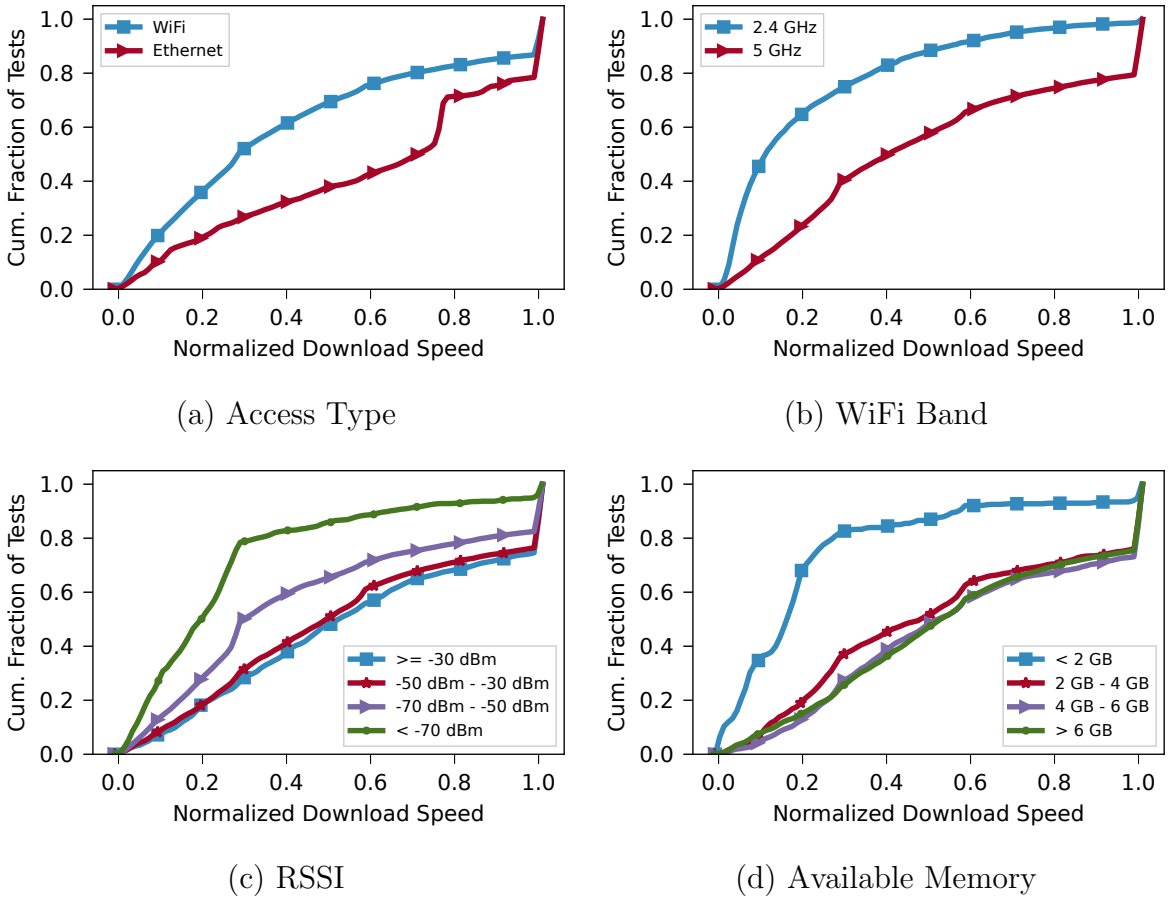
For every user u in month m, we determine the ratio r of tests that were associated with each of the six subscription tiers. For the i^{th} cluster, this can be denoted as:

$$r_{ium} = \frac{N_i}{\sum_{k=1}^{6} N_k} \tag{5.1}$$

where N_i is the number of tests associated with tier i. We denote α as the maximum of these four ratios to represent the tier that had the highest portion of tests associated for a given user in some month i.e $\alpha_{um} = \max_{i \in \{1,..,6\}} r_{ium}$. A higher α value indicates that our BST methodology is consistent for an user across multiple tests in a month. Conversely, if multiple tiers are associated for a user in a month, α will be lower. Figure 5.13 shows the distribution of the α values recorded for users during the 12 months in 2021. The skew of α values towards 1 indicates that, for most users in a month, our BST methodology associates the user to a single tier the majority of the time (the median is 1).

5.6 Diagnosing Speed Test Performance

The association of subscription tier to speed test measurement provides the context needed to determine whether a measurement indicates under-performance relative to the purchased plan data rate. Armed with this information, our objective is now to determine the potential causes of speed test measurements failing to achieve performance close to their subscription plan upload and download speed maximums. For ease of presentation, we present the analysis in this section on measurements from City-A; we verify separately that our findings are consistent with the other three cities. Additionally, because tiers 1-3 for ISP-A in City-A all share the same upload speed, we combine these measurements into one group for the analysis in this section. Finally, we focus the majority of our presentation on download speed due to its greater variability and susceptibility to performance degradation.



(a) Access Type

(b) WiFi Band

(c) RSSI

(d) Available Memory

Figure 5.14: Impact of WiFi characteristics and available memory on speed test performance.

5.6.1 Effect of Home Network and Device

Previous work [220, 153] has documented that the home WiFi link can act as a significant barrier to saturating available bandwidth in the access network. Therefore, our first objective is to understand whether and how characteristics of the client's home network configuration lead to speed test under-performance with respect to the maximum bandwidth of the subscription plan. Our analysis in this section is possible because we can contextualize the measurements with their respective subscription tier information using our BST methodology. To capture any performance impacts, for every measurement, we

normalize the recorded download speed by the offered download speed for the subscription tier. In the following, we quantify the number of speed tests that we identify as affected by different characteristics of the home WiFi link. We then study the effect of kernel memory limitations in the user device on speed test performance.

Access Link. Given the challenges and complexities of WiFi communication, our first step is to compare the speed test results that were conducted over WiFi with those from desktop computers connected to the home network via Ethernet. For this study, we include speed test measurements from all subscription tiers. We examine all WiFi speed tests conducted via Android, iOS and desktop devices (the Ookla-web and M-Lab datasets do not contain metadata about device/access type, and so these are not included in this analysis). Where relevant, we compare the WiFi performance of these devices with that of desktop devices connected through Ethernet.

As can be observed from Figure 5.14a, the difference in the normalized download speed distributions of WiFi and Ethernet access links is significant. For speed tests conducted over a WiFi network, the median normalized download speed is 0.28. This value is almost three times less than the median normalized download speed of 0.71 for Ethernet speed tests. We observe similar results for other cities. Without proper contextualization, the lower download speeds from tests conducted over WiFi could be misconstrued to be under-performance of the provider network.

WiFi Band. Next, we more deeply examine WiFi speed test performance and investigate the impact of the WiFi spectrum band on download speed. Modern routers are equipped to operate in both the 2.4 GHz and 5 GHz WiFi bands [217, 135]. The 5 GHz band supports greater bandwidth while more susceptible to attenuation compared to the 2.4 GHz band [130]. Amongst our datasets, only the Ookla Android measurements contain information about the WiFi band a device used during the speed test. About

23% (15k) of all Android measurements were conducted over the 2.4 GHz WiFi band; the remaining were on the 5 GHz band.

We normalize the reported download speed by the respective ISP offered download speed within a subscription tier. Figure 5.14b shows the distribution of the normalized download speed for all Android measurements separated by the WiFi band. The figure shows a striking difference between the performance of tests in the two bands. While the median normalized download speed is just 0.11 for 2.4 GHz speed tests, it is 0.4 for 5 GHz tests. This median difference in performance between these two bands is amplified for higher subscription tiers. For Tier 6, the median normalized download speed for 5 GHz speed tests (0.25) is over six times more than that of 2.4 GHz measurements (0.04). This finding demonstrates that the WiFi spectrum utilization has an outsized impact on speed test performance, and again, without proper contextualization, the cause of the lower performance on 2.4 GHz devices could be misconstrued.

WiFi RSSI. We next analyse the impact of WiFi RSSI on speed test performance. As our analysis previously demonstrated, 2.4 GHz tests under-perform compared to 5 GHz tests. Hence, for this analysis we only consider the tests conducted in the 5 GHz WiFi band in the Ookla Android dataset. We bin the tests into four categories of WiFi RSSI values. Similar to the access type and WiFi band analysis, for each RSSI bin, we calculate the distance between the measured and subscribed performance for each test. Figure 5.14c shows the distribution of the normalized download speed achieved by speed tests for each RSSI bin.

9% of the 5 GHz Android tests have RSSI values lower than −70 dBm; these tests record the lowest median normalized download speed of 0.2. The median normalized download speed increases to 0.3 for the speed tests conducted in the WiFi RSSI region −70 dBm - −50 dBm; these tests account for 49% of 5 GHz Android speed tests. The

next RSSI bin (−50 dBm - −30 dBm) contains 37% of the total 5 GHz Android speed tests; these tests recorded a median normalized download speed of 0.49. Finally, 5% of all 5 GHz Android speed tests had an RSSI better than −30 dBm; the median normalized download speed for these tests was 0.52. As shown in figure 5.14c, the performance difference varies by over a factor of two between the lowest and highest RSSI bins for all subscription tiers. This difference increases to more than five when considering speed tests in Tier 6. It is therefore critical to contextualize WiFi speed test measurements with signal strength as poorer RSSI can significantly affect the measured performance.

Kernel Memory. We next study the memory available to the Android device kernel during the speed tests to understand its role in achieved performance. For Android measurements, Ookla reports the amount of memory (in megabytes) available to the kernel. To minimize the impact of other factors, we only consider Android measurements in the 5 GHz WiFi band with an RSSI better than −50 dBm (9k measurements).

We bin the available kernel memory into four groups: less than 2 GB, 2 GB – 4 GB, 4 GB – 6 GB and more than 6 GB. Figure 5.14d presents the CDFs of the distance between subscribed and achieved speed test performance grouped by available kernel memory. The distance increases as less memory is available to the kernel during the speed test. 7% of measurements have less than 2 GB of available kernel memory. This group of measurements also recorded the smallest median normalized download speed of 0.16. The next two bins each contribute 17% of the speed tests. The median normalized download speed is 0.48 and 0.52 for 2 GB – 4 GB and 4 GB – 6 GB of available kernel memory, respectively. The majority of speed tests (59%) are issued from devices with over 6 GB of available memory; these tests record the highest median normalized download speed of 0.53, three times more than the 2 GB tests. This difference increases further for higher subscription tiers with Tier 6 tests recording a difference of five times in median

normalized download speed between these two groups. This result shows that speed test performance can be greatly impacted by available memory and is therefore another important piece of context for speed test measurements.

Combination of Local Effects. In our final analysis of the impact of local characteristics on speed tests, we divide the entire Android dataset, across all subscription tiers, into two groups. The first group contains measurements that were conducted on 5 GHz WiFi band, with better than -50 dBm RSSI, and with more the 2 GB of available kernel memory. Based on our results in figure 5.14, this group of tests should experience the lowest impact of the home network and device characteristics on achieved speed test performance. We, therefore, term this group "Best".[5] Conversely, the measurements that do not belong to this group are placed in the "Local-bottleneck" group, as they are more likely to experience constraints from the home network or device memory. It is worth mentioning that the Ookla Android dataset does not provide metadata about other potential local impacts, such as WiFi interference and WiFi channel occupancy. In the absence of this information, we are restricted to the subset of local characteristics presented in this analysis.

In total, 61% ($\sim$12k) of all Android measurements belong to the Local-bottleneck category. This indicates that the performance of the majority of speed tests is likely negatively impacted by home network or device characteristics. Figure 5.15 presents the normalized (with respect to the respective subscription tier) median download speed recorded by both groups. The difference in performance is captured by the median normalized download speed of 0.22 for Local-bottleneck tests, over twice as low as the 0.52 achieved by "Best" tests.

[5]We do not claim that this group of tests does not have other bandwidth constraints, such as a poorly performing cable modem, or a faulty access link, etc. The labeling of "Best" reflects the fact that, amongst the context we investigate, this group of measurements is least likely to experience performance limitations.

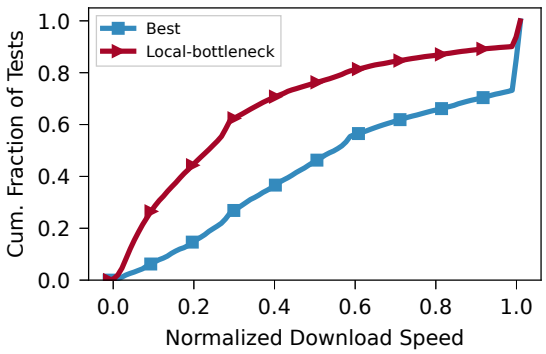


Figure 5.15: Comparison of normalized download speed with and without local bottlenecks.

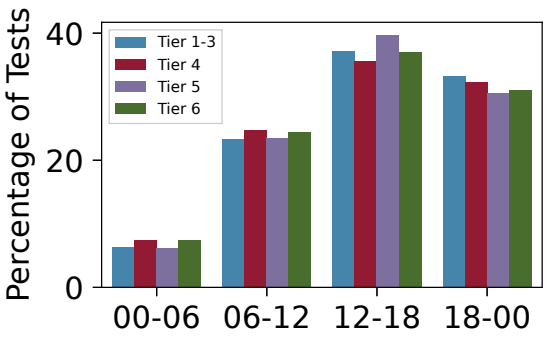
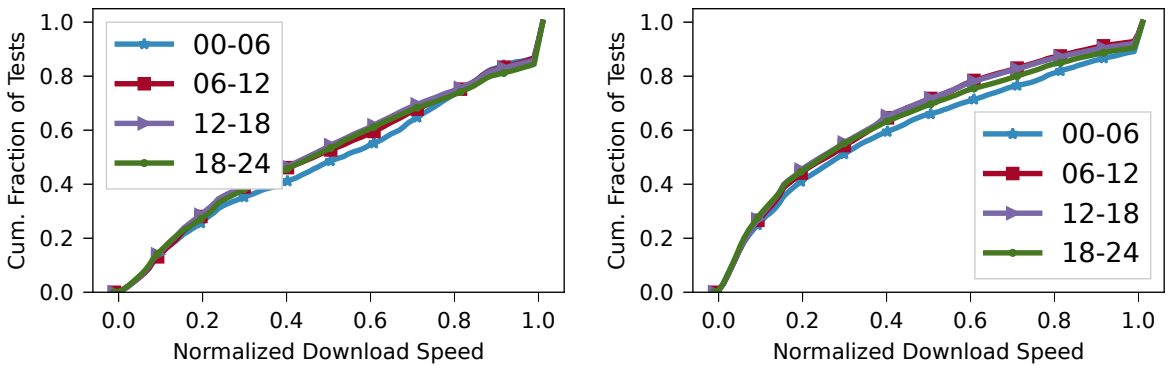


Figure 5.16: Percentage of speed tests in each six hour time bin.

5.6.2 Time of Day Effect

To study the effect of the time of day of a speed test, our first step is to determine the percentage of speed tests that originate at each time of day, for each subscription tier. With this data, we can then analyze the download speed performance, per tier, to determine whether there are measurable differences based on time of day. To explore this time of day effect, we bin the tests into four 6-hour periods: 12am-6am (00-06), 6am-12pm (06-12), 12pm-6pm (06-18) and 6pm-12am (18-00), all with respect to local time of the user. For each time bin, we calculate the percentage of speed tests issued by each subscription tier across all devices in the Ookla dataset; Figure 5.16 shows the result. We observe that there is not a significant difference in the percentage of speed tests in each time bin by subscription tier. We observe a similar trend across all subscription tiers in the M-Lab dataset, but omit these results for brevity. The smallest percentage of tests occur during the night and early morning hours, while the majority of tests, across all subscription tiers, occurs in the afternoon and evening/early night hours. This finding is contrary to the observation made in [219], where it was reported that speed tests are primarily issued during the day.



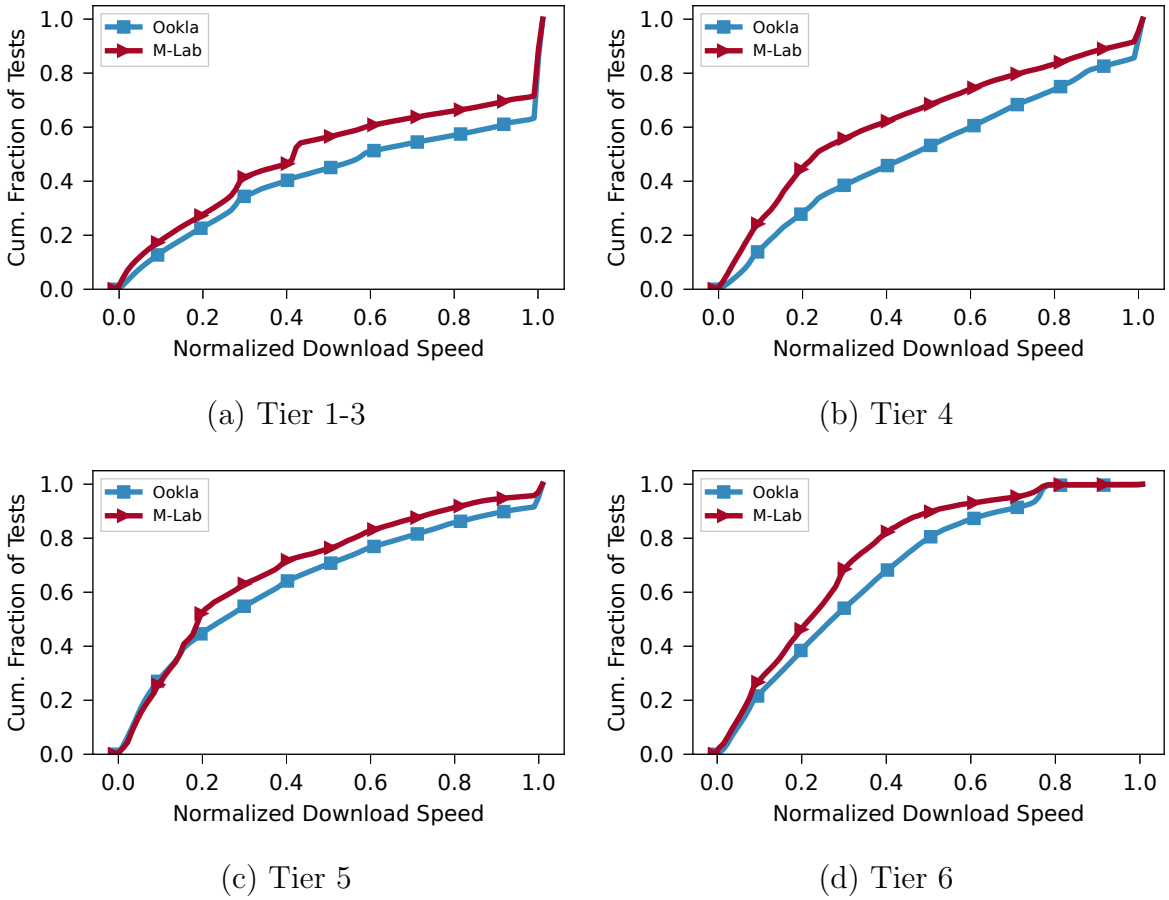
(a) Tier 4 (b) Tier 5

Figure 5.17: Normalized download speed between measured and offered values for Ookla tests based on time of day.

We next explore whether the performance measured by each speed test differs based on the time it is executed. In particular, our objective is to evaluate how much further (or closer) measurement download speeds are compared to the subscription plan maximums based on the time of day. With this approach, we will be able to quantify whether performance drops are more likely to occur during specific time periods.

5.6.3 Effect of Home Network and Device

Figure 5.17 shows the CDFs of the normalized download speed for two subscription tiers across all device types. Our results demonstrate that the speed test performance with respect to the subscribed performance remains similar across all time bins within the day, with slightly better performance recorded for tests conducted during 00-06 hours. For example, the median normalized download sped for iOS tests for Tier 4 are 0.53, 0.46, 0.45 and 0.46 during the 00-06, 06-12, 12-18 and 18-24 time periods, respectively. Similarly, when we analyse the results in the higher subscription tiers, we observe slightly better median normalized download speeds during the off-peak time periods (e.g. 00-06). The median distances for Tier 5 tests are 0.21, 0.19, 0.18 and 0.19 during the 00-06,



(a) Tier 1-3

(b) Tier 4

(c) Tier 5

(d) Tier 6

Figure 5.18: Comparison of Ookla and M-Lab speed test normalized download speed per subscription tier.

06-12, 12-18 and 18-24 time periods, respectively. Based on these results (and similar results for M-Lab data), we conclude that the time of the test does not play a meaningful role in the achieved performance.

5.6.4 Effect of Speed Test Vendor

As stated earlier, Ookla and M-Lab are two of the most popular speed test vendors, and hence the datasets on which we base our study. However, there are some key methodological differences between their speed test measurements. Critically, M-Lab's NDT

conducts its speed test measurements with a single TCP thread, while Ookla speed tests utilize multiple threads [57, 153]. Prior work has found that, as a result, the M-Lab speed test suffers from an underestimation of the available bandwidth [153, 134, 140]. In this section, our objective is to quantify the amount by which the performance reported by Ookla and M-Lab measurements differs. Because we have been able to associate speed test measurements with their subscription tiers, we have the ability to closely compare speed tests that, in theory, should achieve similar performance. Hence, in this study, we are able to compare Ookla and M-Lab measurements within the same subscription tier, for the same city, and the same ISP.

Figure 5.18 shows the distributions of the distances between subscribed and achieved performance for each subscription tier for Ookla and M-Lab measurements in City-A and ISP-A. Across all tiers, M-Lab measurements record greater distance from the subscribed performance than Ookla tests. For tiers 1-3, the median normalized download speed of M-Lab (0.83) is roughly 1.2 times worse than that of Ookla (1). Similarly, the factors by which M-Lab's median normalized download speed lags Ookla's are 2, 1.4 and 1.2 for tiers 4-6, respectively. As a result of these differences, it is critical for users of each test to understand what each test measures before drawing any specific conclusions, or making policy recommendations, based on performance results.

5.7 Related Work

Multiple prior studies have characterized crowdsourced speed test measurements to better understand their utility and usability. In [214], the performance of three million Ookla measurements from 15 cities was analysed. The results demonstrated the high variability that exists in speed test measurement, particularly for wireless tests. However, the work did not analyse the impact of any factors in impeding speed tests from achieving

subscribed performance. The authors in [134] benchmarked Internet performance across multiple metro areas using Ookla speed tests. Their analysis reveals the presence of a large number of low performing speed tests in all cities. Recently, [140] studied the M-Lab dataset and highlighted the need for proper contextualization of measurements prior to drawing generalizable conclusions. The authors of [219] demonstrated the shortcomings of crowdsourced measurements in detecting overall Internet congestion. In [195], the location and income group biases of speed test origin are analysed. The work in [157] illustrates the shortcomings of speed tests in terms of not reaching subscribed speed through a sample of 50 tests from a single home. Similar to our finding, their result shows that upload speed has a small variance compared to download speed. In [153, 127], a detailed analysis of factors that can impact speed test performance is presented. In comparison to these and other similar studies, our work goes significantly further, in part by adding ISP subscription tier context to quantify how close (or far) current speed test results are from actual subscribed performance.

Other prior work has analysed how local network factors can create performance bottlenecks. Local factors are demonstrated to create a bottleneck to achieving download speeds greater than 20 Mbps in [220]. The negative impact of suboptimal WiFi parameters was studied in [144, 130]. The work in [217] demonstrated that factors such as RSSI significantly affect the overall measured latency. In [173], the Secure Digital Input Output bus sleep in smartphone was identified as a large contributor to overall latency. Our study finds that the vast majority of measurements experience bottlenecks by home network and device characteristics, resulting in significant performance underachievement compared to the theoretical maximum of the subscribed broadband plan.

5.8 Conclusion

In this work, we develop a novel BST methodology to augment crowdsourced speed test datasets with ISP subscription tier information. This critical context enables us to analyze and quantify the impact of a variety factors that can degrade speed test performance. The extensive impacts we uncover, which at times differentiate performance more than seven-fold, underlines the need for meaningful contextualization of crowdsourced speed test measurements prior to drawing generalizable conclusions about regional broadband access and quality. This is particularly important for policymakers prior to basing funding and investment decisions on this data. We also highlight the need for speed test platforms used to challenge provider coverage claims to ensure their test methodologies maximize link throughput. We believe that the need for accurate broadband mapping has never been greater, and that crowdsourced speed test measurement platforms will provide an invaluable part of the data needed to generate these maps. We hope that our work contributes to the advancement of this critical mapping effort.

Part III

Internet Affordability

Chapter 6

Decoding the Divide: Analyzing Disparities in Broadband Plans Offered by Major US ISPs

6.1 Introduction

The National Digital Inclusion Alliance (NDIA) in the US defines digital equity as "a condition in which all individuals and communities have the information technology capacity needed for full participation in our society, democracy, and economy" [186]. As modern life has moved increasingly online, high-quality Internet access has become a key component of digital equity. The Covid-19 pandemic, and the post-pandemic "new normal" of remote interaction, have drastically changed the need for home Internet access; work-from-home, online/remote schooling, telemedicine, and other networked applications have become increasingly indispensable. As a result, individuals without home access to highly reliable, high-speed broadband are severely disadvantaged [37].

Policymakers cannot take effective corrective actions, such as offering subsidies [154], regulating rates [40], and funding access infrastructure [187], without understanding the true characteristics of digital inequity. Digital equity, especially in the context of Internet access, is often measured along three axes: availability, affordability, and adoption [215]. Many past efforts [197, 196, 132], including ones in our research community, have focused on measuring availability. Researchers have disaggregated availability into coverage and quality. Here, coverage answers whether broadband access is available in a geographical region, while quality answers questions related to access type (e.g., cable, fiber, DSL), and upload/download speed. Researchers and policymakers use publicly-available datasets, such as the FCC's Form 477 [116], Measuring Broadband America (MBA) [155], and Measurement Lab (M-Lab) speed test [56], as well as proprietary ones, such as Ookla's speed test [93], to characterize Internet connectivity. More recently, as part of the Broadband Equity, Access, and Deployment (BEAD) program, the US Congress directed the FCC to develop an accurate map of fixed broadband availability across the US. Though it is still a work in progress, when completed, the FCC National Broadband Map [78] will provide information regarding broadband availability (i.e., provider, access type, maximum upload/download speed) at the granularity of street addresses.

Whereas the existing datasets in the US broadband sector, including the most recent FCC National Broadband Map, measure availability, affordability has received less attention. To answer any question related to broadband affordability, extracting the "cost of broadband connectivity", i.e., the nature of the "deal" a user is getting, at fine-grained geographical granularity, is important. Using cost data, one can answer policy questions such as (1) what pricing policies do ISPs employ to users in different regions (i.e., neighborhoods, cities, states)?; (2) where, within a region, are different types of deals offered by ISPs?; (3) how does the (lack of) competition among ISPs affect broadband prices in a region?; and (4) how do socioeconomic and demographic factors correlate with

broadband prices?

Most previous studies have either focused on manually querying ISP websites [185, 172] or self-reporting from ISPs [156], and, at best, they scratch the surface of questions (1) and (2). A more recent study by a team of investigative journalists curated broadband availability and cost data at street-level granularity for four major ISPs across 43 cities.[1] However, among other limitations, this study did not analyze the broadband plans for major cable-based ISPs (e.g., Cox), and thus, it could not fully answer questions (3) and (4).

Our goal is to curate a new dataset that enables a better understanding of broadband affordability in the US, addressing the limitations of prior related efforts. To this end, we present the design and implementation of a new *broadband plan querying tool (BQT)*. BQT takes a street-level address as input and returns the available broadband plans offered by major ISPs at that address. Here the plans entail the maximum upload speeds, download speeds, and corresponding prices in US dollars; typically, multiple plans are available to each residential address. BQT automates mimicking the behavior of a real user interacting with an ISP's website to query available broadband plans for a given street address. It addresses various challenges to offer a high hit rate, i.e., the number of street addresses it can successfully query for an ISP and the number of major ISPs it can query.

We use BQT to curate our broadband plans dataset while ensuring our data collection effort does not overwhelm ISP websites. Specifically, we collect and analyze plan data in thirty US cities with diverse populations, population density, and median income. We identify seven major ISPs that reach 89% of the total census blocks in the US [78]. For each (ISP, city) pair, we sample a subset of residential addresses extracted from a dataset provided by Zillow [105]. We feed these addresses to BQT to curate the desired

[1]Our team provided technical assistance for this investigative reporting.

broadband affordability dataset.

We use this dataset to answer multiple policy questions about broadband affordability in the US. Specifically, we use the metric *carriage value* to characterize broadband plans.[2] This metric quantifies the amount of user Internet traffic (in megabits) that an ISP can carry per second, per dollar spent on a monthly broadband plan. For example, the carriage value for a broadband plan with a download speed of 100 Mbps at $50/month is 2 Mbps/$. Intuitively, the higher the carriage value, the better the deal the user receives for their broadband subscription, and vice versa. We use this metric to study the quality of "deals" ISPs offer within and between different cities. From an end user's perspective, we explore how this metric varies across different ISPs active in a region, how the nature of the deal correlates with various demographic and socioeconomic factors, and the state of competition among ISPs locally. By using this metric, this chapter and its findings can contribute directly to the ongoing discussion currently active in the US on broadband pricing, ensuring consistency and relevance.

In summary, our work offers three major contributions:

Broadband plan querying tool (Section 6.3). We present the design and implementation of a broadband plan querying tool that reliably queries the websites of seven major ISPs, mimicking a real user, to extract the available broadband plans for a given street address.

Broadband plans dataset (Section 6.4). We present our methodology to curate a broadband plans dataset by querying 837 k unique addresses (1.2 M plans) across 30 cities (18 k census block groups) and seven major ISPs in the US. Our emphasis is metropolitan/urban areas across the US. However, our work can be expanded to include small towns and rural areas.

[2]A paper recently proposed this metric in the legal literature [185] that the White House referred to in announcing a new Executive Order [47] citing a call to arms to address the lack of competition among broadband service providers in the US.

Characterization of broadband plans (Section 6.5). We conduct a multi-dimensional analysis to study the intra- and inter-city distribution of broadband plans (i.e., carriage value) for each ISP and how these plans are affected by competition among ISPs and various demographic and socioeconomic factors. Our analysis offers the following key insights: (1) ISP plans vary by city, i.e., the fraction of census block groups that receive high (and low) carriage value plans are variable across cities.[3] (2) ISP plans within a city are spatially clustered, and the carriage value can vary as much as 600% within a city. (3) Cable-based ISPs deliver up to 30% greater carriage value to users when in competition with fiber-based ISPs within a block group, as opposed to when they operate independently or alongside a DSL-based ISP. (4) Block groups with higher average income tend to be associated with higher fiber deployments, which offer superior carriage values. However, racial composition and population density, when considered independently of average income, do not correlate with differences in fiber deployment.

We view this work as an important step towards understanding broadband affordability in the US at scale. We note that broadband affordability is multifaceted, with numerous factors to consider. While our analysis provides valuable insight, it only scratches the surface of what policymakers must address when assessing broadband affordability. The evaluation of broadband affordability in a specific region or for a particular population may require consideration of additional factors beyond the scope of this chapter. To enable other researchers and policymakers to advance our understanding of this critical topic, we will make our tool and a privacy-preserving version of our dataset publicly available. We conclude this study with recommendations for different stakeholders to further improve the understanding of broadband affordability.

Ethical concerns. Please refer to Section 6.4.2 for a discussion of how we address

[3]Xfinity emerges as an exception as its plans are invariant across the specific cities we study in this work.

ethical concerns regarding our data-collection tool and methodology.

6.2 Background & Motivation

Broadband providers in the US. Thousands of US ISPs offer broadband connectivity, reaching approximately a hundred million residences. Most of these ISPs operate locally and have a fairly small footprint [123, 114, 122]. This paper considers seven major ISPs, each serving at least one million residences. Together they reach 89% of the total census block groups in the US. We can divide these ISPs into two broad categories: DSL/fiber-based[4] and cable-based providers. Our work, like others [31], confirms that these ISPs either operate as a monopoly or duopoly, i.e., at max, only two major ISPs compete with each other in a census block group. Also, ISPs of the same type do not compete with each other: DSL/fiber-based ISPs do not compete with each other, and cable-based ISPs do not compete [31]. Moreover, in major cities, cable-based ISPs dominate in terms of coverage, i.e., they serve almost all the block groups [78]. In contrast, DSL/fiber providers serve a smaller fraction of block groups. Finally, in part because fiber deployments are relatively new and more expensive to deploy, DSL is often (though not always) offered in more block groups than fiber. Given these trends, cable-based ISPs operate in three distinct modes: *cable monopoly*, *cable-DSL duopoly*, and *cable-fiber duopoly*.

Existing broadband availability datasets. The FCC recently launched a street address-level map of broadband availability [78]. This is an improvement over the previous iteration, based on provider input through Form 477 [116], which offered this information at census block-level granularity. This new map reports the maximum upload and download speeds and the access technology (e.g., fiber, cable) at street-level granularity and relies on self-reporting from ISPs. Previous efforts curated similar data by

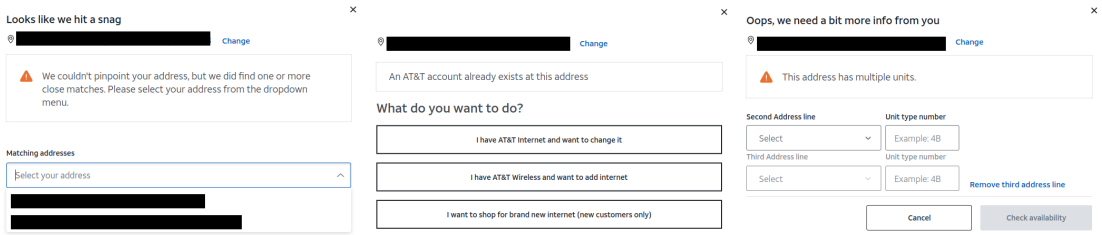
[4]We categorize DSL and fiber providers together as, if an ISP offers a DSL-based service, it typically also offers a fiber-based service, and vice versa.

manually [110] or automatically [179] querying ISP web interfaces, also referred to as a broadband availability tool (BAT). Such third-party efforts enable auditing self-reported data from different ISPs [27, 179, 121].

These datasets improve our understanding of broadband availability, both in terms of coverage and quality. However, without any pricing information, it is not possible to characterize broadband affordability.

Existing broadband plan datasets. Prior efforts have typically curated broadband plan datasets by manually querying ISP BATs. For example, the California Community Foundation and Digital Equity Los Angeles queried Spectrum's website to curate a list of broadband plans for 165 street addresses in Los Angeles County (California) [172]. One study [185] manually compiled a dataset of 126 street addresses across seven states to obtain available plan information. While these studies highlight the disparity in broadband plans, small-scale datasets are, at best, suggestive of broader and more general trends.

More recently, an online investigative platform, The Markup [98], extended the BAT client [179] approach to automate the extraction of broadband plans for four major ISPs in 43 US cities. Their study [73], which is the most closely related prior work to ours, finds significant variability in the download speed offered by major ISPs at different price points. For instance, the authors found that, for $55/month, AT&T offers 1000 times greater maximum download speed to some addresses in the same city; this phenomenon is referred to as "tier-flattening" [7]. The Markup's study also finds that some major ISPs, such as AT&T and CenturyLink, provide lower speeds to more vulnerable populations, e.g., low-income and high-minority communities, than others. Based on this analysis, the authors highlight the importance of analyzing the cost of Internet service and download speed instead of download speed in isolation. A limitation of the Markup's study, however, is that it does not include cable-based ISPs, which serve most of the US population [71]. Consequently, their dataset is not suited to explore the



(a) Wrong address (b) Existing account (c) Multi-dwelling units

Figure 6.1: Illustration of different steps that BQT handles while querying ISP broadband plans through their BATs.

dynamics between cable and DSL/fiber providers nor to the study of how competition between the two changes the nature of broadband plans in a region. In addition to that, as discussed in Section 6.3.2, extending BAT clients to collect data for all major ISPs is non-trivial.

Our approach. In this work, we address the key gaps of previous efforts by curating a comprehensive broadband plan dataset in terms of location and type of ISPs. First, we develop BQT to obtain plan information across 837 k street addresses for three major cable providers and four major DSL/fiber providers. Our dataset provides insight into the ISP plan structure in 30 cities around the US. Using this dataset, we can characterize how ISP plans change between cities, within a city, and in the presence of another ISP.

6.3 The Broadband-plan Querying Tool

Our goal is to develop a *robust* measurement tool that can *accurately* report the broadband plans offered by major ISPs for a given set of street-level addresses at *scale*. Rather than relying on user surveys [172] or self-reporting [156] from ISPs, we focus on directly querying ISP BATs. Minimizing disruption to end users using BAT is an important priority while developing this tool. In essence, for a given list of input addresses, we want this tool to achieve a high hit rate, i.e., successfully extract broadband plans for as many

input street addresses as possible, promptly, yet without disrupting the normal service offered by the ISP to end users.

6.3.1 Challenges

In theory, obtaining broadband plan information from an internet service provider's BAT should be straightforward. However, in practice, it is often complicated due to the quality of street address datasets. Most street address datasets are crowdsourced [124, 118], which can result in incomplete, incorrect, or ambiguous information. As a result, the querying process is a dynamic, multi-step process, where the information displayed at each step is based on the internal logic and state of each BAT, as well as the input provided by the user in the previous step. For instance, after the user enters a street address, the next web page may either show available broadband plans, indicate an incorrect input address, or inform the user that they are already a subscriber at that address. Additionally, ensuring that the tool can query all major ISPs is challenging because different ISPs use different formats and interfaces, such as drop-down menus or click buttons, to present this information and allow users to respond.

To illustrate, Figure 6.1 shows different steps that our tool needs to follow to extract the broadband plans. Here we use AT&T as an example, but we confirm that all other ISP BATs also follow these steps. In the first case, as illustrated in Figure 6.1a, AT&T could not identify the input street address.[5] When faced with this scenario, the expected response for the end user is to access the drop-down menu that the BAT provides and then select an address from the offered address set. As a next step, AT&T could indicate that an active customer already exists in this specific street address. In this scenario, the BAT offers three distinct choices, as shown in Figure 6.1b. If a user is already an AT&T subscriber residing in that address, the first two options given them the ability

[5]Note for privacy reasons, we have blurred the specific street address in this example.

to change their plan or add a new plan. This would prompt the BAT to render an authentication form to ensure the user is an active subscriber. The third option applies to a new customer who is interested in viewing the set of AT&T plans at that address. This step does not require any authentication. Finally, a particular address could be a multi-dwelling unit, i.e. with an apartment/unit number that was not input during the initial stage. For that scenario, as demonstrated in Figure 6.1c, the BAT provides an option to select one of the possible apartments/units at that address.

6.3.2 Strawman: Extend Existing BAT Client

A potential solution to obtain broadband plan information is to enhance the BAT client approach proposed in previous research [179]. This approach was designed to query the binary availability of broadband service (i.e., service/no service) for a specific street address. For every ISP, a BAT client was designed, which involved reverse-engineering each ISP BAT by observing how it uses different RESTful APIs to extract the desired information, such as broadband availability. For example, the BAT client can observe that when a browser sends a request with a street address, it receives a response with an ID, and subsequent requests in the next step use this ID and, in some cases, a session cookie from the previous step. The BAT client then uses the Python `requests` library to directly send a series of requests to the ISP's RESTful APIs. Directly querying the APIs is scalable; thousands of street addresses can be handled in parallel. In 2020, the authors in [179] used this approach to query approximately 35 million street addresses. Their data analysis revealed the limitations of the information provided by the FCC's Form 477 [116], reinforcing the need for such information to be made available at street-level granularity as previously suggested by other research [196, 143].

Limitations. Since the BAT client approach has been successfully used to query millions

of street addresses for all major ISPs, extending it to extract offered broadband plans seems like a natural choice. However, we observed that the proposed approach has several limitations that make it difficult to adapt to satisfy our goals. Specifically, since the publication of the previous work [179], ISPs have safeguarded their RESTful APIs from such direct querying.[6] For example, some ISPs have now started using dynamic cookies that append unique server-side parameters to each user session. Some BATs have started blocking queries from an IP address that uses the same cookie across multiple API requests. Dynamically generating a new cookie for each API request is non-trivial and is not supported by the original BAT client.

6.3.3 BQT Approach

To decouple the querying process from ISP safeguarding strategies, our approach avoids directly querying their RESTful APIs. Instead, we use a popular web automation tool, Selenium, to mimic different end-user interactions for extracting the desired broadband plan information.

As a first step, we manually inspect the workflow for different ISP BATs. Each BAT employs a specific template to display the information for each step in the workflow. As part of this manual bootstrapping step, we enumerate all possible templates and identify unique patterns in their HTML content using regular expressions to help detect them at runtime.

The second challenge is to identify how to mimic a user's behavior using Selenium to advance successfully to the next step. This step is critical for ensuring a high hit rate for BQT. Specifically, we handle different templates as follows.

Incorrect address. As mentioned earlier, street addresses are noisy due to inherent am-

[6]We do not assert that ISPs have changed their safeguarding strategies in response to previous data-collection efforts.

biguity between different identifiers. For example, for the same street address, some databases might use "Ave" instead of Avenue and "CT" or "Ct" instead of Court. Whenever there is a mismatch between the input street address and the one in the ISP's database, it shows an "incorrect address" web page and often provides a list of one or more street addresses as suggestions. Given the prevalence of this occurrence, addressing it is critical to ensure a high hit rate for BQT. We address this issue by storing the list of suggested street addresses for offline analysis. We then apply string-matching over each suggested address in this list to find the one that best matches the input street address. As a sanity check, we ensure that the selected street addresses have the same zip code as our initially queried address. We then query the ISP's BAT to extract the broadband plan information.

Multi-dwelling units. For addresses where a specific street address has multiple dwelling units (e.g., two or more apartments), the ISP BAT typically shows a "multi-dwelling unit" web page and suggests more refined street addresses (e.g., specific apartment numbers). Similar to previous work [179], we replace the input street address with a randomly selected address from this list. We then use this new address to query the ISP's BAT to extract the broadband plan information.

Existing customers. If the resident of an input street address is already a subscriber, the ISP BAT displays an "existing customer" web page and offers two options. The first option directs the user to their account, while the second allows a new user to query offered plans. Given our interest in extracting the available broadband plans, we select the second option.

To avoid failures, we must ensure that all the Document Object Model elements for a step are successfully downloaded before applying any user action. The download times can vary across different templates and ISPs. For example, the step that displays available broadband plans after inputting the street address takes less than 30 seconds

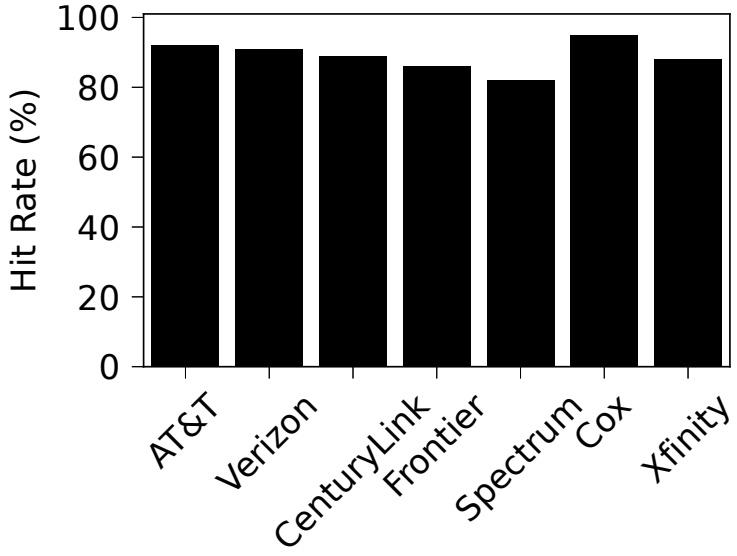


Figure 6.2: BQT hit rate per ISP.

for AT&T but 60 seconds for Spectrum. Thus, we measure the download times for all possible templates and pause for this period (i.e., max observed download time) before applying the user action.

Microbenchmarks. The two crucial performance metrics of BQT are hit rate and query resolution time. The hit rate informs the fraction of total queried addresses for which we are able to obtain a response from a particular ISP BAT successfully. As shown in Figure 6.2, our hit rate for all ISPs exceeds 80%; we achieve the highest hit rate of 96% for Cox and the lowest for Spectrum (82%). Such high hit rates across all ISPs ensure that BQT is able to extract plan information for the majority of the addresses. Our investigation into the instances where BQT encounters failures reveals that the primary cause is the denial of connectivity by the IP proxy service. Furthermore, some ISPs classify certain requests as originating from data centers due to IP addresses, resulting in service denial and subsequent failures. If we re-run the addresses that previously failed, there is an increase in the BQT's hit rate for each ISP. The query resolution

time for a given street address is the amount of time it takes BQT to obtain a response from an ISP BAT. Figure 6.3 presents the distribution of query resolution time for each ISP. The median time for Frontier query resolution is lowest, at 27 seconds, while it is highest, at 100 seconds, for Spectrum, despite no significant difference in the number of intermediate webpages rendered. Given that this latency can be significant, we describe the methodology we adopt to make BQT more scalable in Section 6.4.1.

Limitations. BQT has been specifically designed to work with the BATs offered by seven major ISPs. However, any changes made to the interfaces of these BATs by the ISPs, such as the addition of new drop-down forms, will require BQT to be updated. To ensure that BQT continues to function properly over time, we must monitor the BATs for all the supported ISPs and upgrade BQT as necessary to accommodate any changes. In the future, we plan to make BQT more modular, which will help minimize the effort required to adapt it to these changes.

6.4 Broadband Plan Dataset Curation

In this section, we describe the dataset we aggregate through BQT. We first describe our methodology to query a subset of street addresses and ISPs to curate the broadband affordability dataset. We then describe how we selected the ISPs, cities, and street addresses for data collection (Section 6.4.1). Next, we discuss how we addressed different ethical concerns regarding our data-collection methodology (Section 6.4.2). Finally, we discuss the limitations of our dataset (Section 6.4.3).

6.4.1 Data Collection Methodology

In the US, seven ISPs serve approximately 90 million street addresses (87% of the total US census blocks) [78]. Through our data usage agreement with Zillow [124], we have access

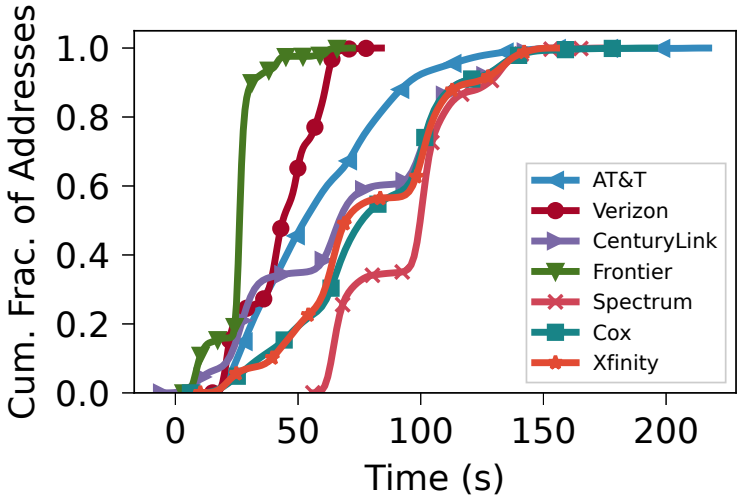


Figure 6.3: BQT query time resolution distribution per ISP.

to about 104 million "residential" US street addresses. Note that while this database does not represent every US address (it is comprised of addresses that had a transaction during a specific period), it encompasses a very large subset. Further, compared to alternative address datasets, such as the National Address Database (NAD) [118] offered by the US Department of Transportation, the Zillow dataset offers more complete coverage and is less noisy. Specifically, it includes nearly every county in the US, and USPS has validated the addresses as suitable for postal delivery [74]. Note that validation for postal delivery from USPS does not guarantee a perfect match with an ISP's BAT; addresses can still be flagged as incorrect, incomplete, or ambiguous. However, it offers an excellent starting point.

In theory, we can use BQT to extract the available broadband plans for all ISPs that serve each street address in the Zillow dataset. However, we realized that curating such an extensive dataset has diminishing returns. Our initial exploration of the collected data revealed that broadband plans are spatially clustered, so plans for street addresses in the same neighborhood (i.e., a census block group) are similar. Additionally, our primary

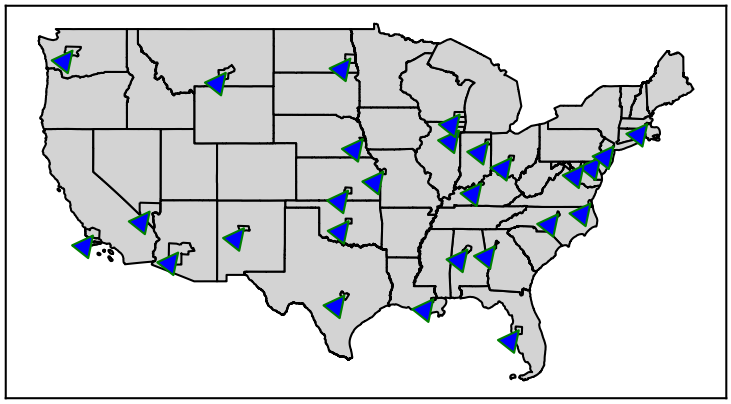


Figure 6.4: Geographical location of the 30 cities in our study.

focus for this study is metropolitan/urban areas around the US. Given the coverage of Zillow's address database, we can extend the scope of the study to micropolitan and rural areas in future work.

With this context in mind, we now describe our selection methodology for the ISPs, cities, and street addresses for our study.

ISP selection. We focus on fixed, terrestrial broadband providers that offer queriable BATs and serve at least a million street addresses in Zillow's dataset. After applying this filter, there are seven major ISPs: AT&T, Verizon, CenturyLink, Frontier, (Comcast) XFinity, (Charter) Spectrum, and Cox. Among them, Xfinity, Spectrum, and Cox are cable-based providers, and AT&T, Verizon, CenturyLink, and Frontier offer DSL and fiber-based plans. Previous work reported that Comcast Xfinity's offerings are invariant to location [73]. Our analysis using the data collected using BQT from six major US cities confirmed these observations, and so we omit collecting data for this provider.

City selection. With a goal of wide geographic distribution, we examined cities with a range of population densities as well as diverse socioeconomic attributes (e.g., average income) that are well represented in the Zillow dataset. After applying this filter, we selected 30 major cities in 27 states (see Figure 6.4). As shown in Table 6.1, these cities

	Block Groups	Street Addresses (k)	Population Density (k)	Median Income (k)	Major ISPs 1	2	3	4	5	6	7
Albuquerque, NM	387	14	1.8	53			•				
Atlanta, GA	389	12	1.2	65	•						•
Austin, TX	487	25	1.7	74	•				•		
Baltimore, MD	1188	42	1.7	81		•					•
Billings, MT	98	3	1.1	61			•		•		
Birmingham, AL	354	24	716	47	•				•		
Boston, MA	37 3	17	8.4	72		•					•
Charlotte, NC	472	21	2	73	•				•		
Chicago, IL	1933	86	3.8	64	•						•
Cleveland, OH	754	35	4.8	31	•				•		
Columbus, OH	662	20	1.9	58	•				•		
Durham, NC	138	5	1	59				•	•		
Fargo, ND	67	5	1.5	62			•				
Fort Wayne, IN	209	11	0.9	54				•			•
Kansas City, MO	305	15	1.2	51	•				•		
Los Angeles, CA	1787	90	8.5	67	•				•		
Las Vegas, NV	881	38	1	65			•			•	
Louisville, KY	505	41	1.6	56	•				•		
Milwaukee, WI	560	27	2.9	50	•				•		
New Orleans, LA	439	67	2.9	41	•					•	
New York City, NY	1567	51	41.7	96		•			•		
Oklahoma City, OH	493	20	1.3	50	•					•	
Omaha, NE	455	28	1.7	62			•			•	
Philadelphia, PA	981	32	8	46		•					•
Phoenix, AZ	802	32	1.9	64			•			•	
Santa Barbara, CA	211	6	2	79				•		•	
Seattle, WA	634	28	2.1	101			•				
Tampa, FL	536	25	1.5	57				•	•		
Virginia Beach City, VA	112	4	1.8	80		•				•	
Wichita, KS	304	13	1.3	50	•					•	
Total	**18k**	**837**			**14**	**5**	**7**	**4**	**13**	**8**	**6**

Table 6.1: Dataset coverage. The major ISPs are listed in the following order: (1) ATT, (2) Verizon, (3) CenturyLink, (4) Frontier, (5) Spectrum, (6) Cox, and (7) Xfinity. Note that Xfinity also provides service in Albuquerque, but we did not include this service in our study.

represent a broad spectrum of demographic and socioeconomic attributes. For example, the range of population densities varies from 1 k to 42 k per sq. mile [2], and the median yearly household income varies from $31 k to $101 k. We focus on cities that are served by at least two of the seven ISPs considered in our work to ensure that we capture any trends that emerge as a result of competition between ISPs in a region.

Street address selection. Each city in the US is divided into census blocks, which are aggregated into census block groups. The US Census Bureau defines a census block group (CBG) as representing approximately 600–3000 people that are considered to be homogeneous in terms of their demographic and socioeconomic characteristics. For the

cities considered in this work, Zillow's database includes addresses in all the census block groups for each city, ensuring comprehensiveness. However, as querying every address in a city would impose a significant load on the ISPs' infrastructure, we opt for a sampling strategy. To ensure that our sampling strategy mimics the socioeconomic composition of the city, we uniformly sample street addresses at the census block group level. Specifically, for each (ISP, city) pair, we identify the set of block groups covered by the ISP in a city. We randomly sample 10% of street addresses for each such block group. If we are unable to obtain the BQT data for any of those addresses, we continue sampling until we have a successful sample of 10% of street addresses in each CBG.

Scaling data collection. To gather the needed samples for our study, BQT needs to query 837 k street addresses, the total number of addresses resulting from sampling 10% of every census block group. We run multiple instances of BQT in parallel to scale the data collection. We use Docker containers to run these instances concurrently on a single local data-collection server. We can theoretically use as many containers as street addresses for different ISPs to expedite data collection. However, such an approach will overwhelm ISP BATs and degrade the user experience for actual customers.

Though we cannot directly measure the experience for real users, we conducted an experiment where we measured ISP response time for 1, 50, 100, and 200 Docker instances. We hypothesize that if running multiple Dockers is affecting user experience, we should expect a statistically significant difference in ISP response time for different settings. However, we observed that the response time for any ISP did not change as we increased the number of Docker instances. Based on this experiment, we are confident that using up to 200 Docker instances does not overwhelm ISP servers enough to disrupt the user experience. Nevertheless, we scale back and utilize 50-100 distinct containers for our data collection. Note that our choice of 200 instances is based on the intuition that an ISP should not get overwhelmed by such a small number. By no means is it an

upper bound on how many Docker containers we can run in parallel.

To ensure that all our queries do not originate from a single non-residential IP address, we utilize a pool of residential IP addresses provided by Bright Initiative, the non-profit branch of Bright Data [109] (formerly known as Luminati). This organization offers free access to data scraping tools for nonprofits and academic organizations. Previous efforts [179, 73] have also used this service.

We conduct our data collection campaign from December 2022 to February 2023.

Public release. We will make a version of this dataset publicly available to empower other researchers and policymakers to improve our communal understanding of broadband affordability in the US. Due to the proprietary nature of Zillow's data, we cannot include specific street addresses in our dataset. Instead, for each queried street address, we will only reveal its block group identifier along with the corresponding ISP plans. Considering the limited variability in broadband plans within a block group (see Section 6.5.1), we believe the released dataset will still hold value for various stakeholders.

6.4.2 Ethical Considerations

We query ISP plans at the street address level and do not collect or analyze Personally Identifiable Information (PII). Our work does not involve human subjects research, and the private dataset provided by Zillow under the data use agreement does not reveal any individual's identity. Furthermore, the data gathered from the website does not include any PII. We do not have the means to identify residents, the selected broadband subscription tiers, or the actual performance received at any address. Our methodology involves obtaining ISP plan information from their websites, which is consistent with legal requirements and research community norms [65, 83, 66].

6.4.3 Limitations

We now discuss a few limitations of our dataset and how to address them in the future.

Staleness issues. Our dataset provides a single snapshot of broadband plans, which may change over time as ISPs update their infrastructure and pricing structures. We observe that many ISPs are actively deploying new fiber, and we expect their offered plans to change in the near future. Also, ISPs occasionally offer discounts (i.e., higher carriage value (cv) plans), especially in areas where they compete with other major ISPs. Our dataset does not discriminate between normal and discounted offers and, thus, might not best reflect the most recent carriage for a subset of street addresses.

Limited coverage. Although our dataset includes addresses from every census block group in the 30 cities examined in this study, it represents only about 7.5% of all block groups in the US. We currently use Zillow's data, which is biased toward high-density urban areas. We need a better representation of street addresses in semi-urban and rural areas. Though curating such datasets is challenging, recent efforts from the FCC to develop broadband availability maps at street address granularity demonstrate such an approach's feasibility. In future work, we will complement Zillow's dataset as needed with other sources, such as the NAD, to cover other areas where Zillow's data alone lacks sufficient representation.

Veracity of reported plans. There is no system or database to confirm the accuracy of the download speed and price data provided by the ISPs when querying a street address. However, as mentioned in [179], it is not in the interest of ISPs to report false or misleading information to potential customers, including poor performance or low-valued plans. We note that the total cost incurred by subscribers for ISP services often exceeds the initially advertised prices. This includes subscriber-specific discounts, undisclosed fees, taxes, and additional charges [104]. However, our focus in this study is the initial

advertised price offered by ISPs; the analysis of the final amount paid by subscribers is beyond the scope of the current work.

6.5 Broadband Plan Characterization

In this section, we will demonstrate how our broadband affordability dataset provides the means for various stakeholders to address crucial policy questions that previously were difficult to answer. To do so, we will first present an overview of the BQT dataset. We will then answer the following critical questions: ❶ Do the broadband plans, characterized by their carriage value, change by city for different ISPs? ❷ Does the carriage value change within a city? If yes, which neighborhoods (identified by their census block groups) receive good and bad deals (high and low carriage values)? ❸ Does competition among ISPs impact the carriage value offered to the end users? If yes, is there a trend in which neighborhoods experience competition? ❹ Is the quality of available deals correlated with demographic and socioeconomic factors? If yes, which population groups receive better or worse deals from the ISPs?

6.5.1 Dataset and Metrics

Dataset overview. Table 6.1 summarizes the number of street addresses and block groups we cover for each of the thirty cities. It also shows which of the seven major ISPs are active in each city and hence in our dataset. Overall, our dataset covers 837 k distinct street addresses, representing 18 k block groups (around 7.5% of total block groups in the US). None of the thirty cities are served by more than two major ISPs. This trend indicates the presence of monopolies and duopolies in these cities [31].

Table 6.2 summarizes the available broadband plans from the seven major ISPs. The range of plans is more diverse for fiber/DSL-based providers than cable-based providers.

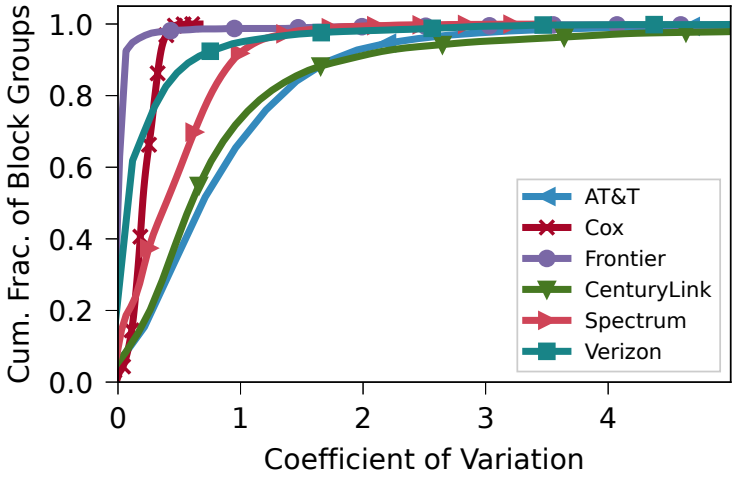


Figure 6.5: Distribution of coefficient of variation of carriage values in a block group for each ISP.

The extremely low upload/ download speeds (and related carriage values) are attributable to broadband plans via DSL.

Calculating carriage values. We use the carriage value to characterize a broadband plan offered by an ISP, and we curate this metric for all input street addresses. Since the entropy of available download speeds is greater than the upload speeds, we focus on download speed to calculate carriage value. While not shown, we verified that our results are consistent if we use upload speed to determine carriage value.

Each ISP offers a fixed number of plans across all cities. For example, AT&T offers 11 different plans across the 14 cities it serves in our study. However, an ISP only offers a subset of these plans at any given street address. For example, for a specific street address in New Orleans, AT&T offers three different plans: (1000 Mbps, $80/month), (500 Mbps, $65/month), and (300 Mbps, $55/month), which translates to carriage values of 12.5, 7.7, and 5.5, respectively. To represent the value provided by an ISP through a set of plans to a street address, for every address, we consider the best carriage value (cv), i.e., 12.5 in the case of the above address. We note that the cv metric has inherent

	Unique Plans	**Download (Mbps)**	**Upload (Mbps)**	**Monthly Price ($)**	*cv*
AT&T	11	0.768–1000	0.768–1000	55–80	0.01–12.5
Verizon	4	3.1–1000	1–1000	50–100	0.4–11.1
CenturyLink	8	1.5–940	0.5–940	50–65	0.03–14.5
Frontier	2	0.2–2000	0.2–2000	50–100	0.0004–20.0
Spectrum	5	30–1000	5–35	20–70	11.1–14.3
Cox	6	100–1000	5–35	20–120	10.0–28.6
Xfinity	3	25–1200	5–35	20–80	3.8–15.0

Table 6.2: Overview of broadband plans offered by the seven major ISPs. The dashed line separates DSL/fiber-based providers from cable-based ones.

limitations due to the nature of broadband pricing. Since speed tends to vary more than price—e.g., at the address mentioned above, 1.5x cost gets 3.3x more bandwidth—the highest carriage value (*cv*) plan available for an address is also the highest-*speed* plan. Users may not require the highest speed available or want to pay for it, so *cv* is not necessarily a reliable proxy for the subjective value of a plan to its customers. For this and other reasons, policy decisions should not optimize around *cv* alone.

In some of the analysis that follows, we compare block groups by carriage value. The *cv* of a block group provided by an ISP is computed as the median of the maximum carriage values of the plans sampled from the addresses in that block group. Using an aggregate metric at block group granularity simplifies spatial analysis, and ensures that our analysis is not biased by block groups with more street addresses in the Zillow dataset. However, it also hides variability within block groups. To characterize this variability, Figure 6.5 shows a distribution of the coefficient of variation (CoV), i.e., the ratio of the standard deviation to mean, for the *cv*s available per ISP for every block group in our data set. Most ISPs show low CoV across all block groups, meaning the aggregate *cv* metric hides little information. However, there is a long tail for AT&T and CenturyLink, which sometimes offer both DSL (very low *cv*) and fiber (very high *cv*)

plans within the same block group. We checked the robustness of our per-block-group findings by performing an analysis where block group *cv* was computed as the median of the *minimum cv*s of the sampled plans; our conclusions (e.g., Section 6.5.2, Figure 6.8) were consistent regardless of this choice.

Comparing plans. To compare an ISP's plans across different cities or the plans of two competing ISPs within a city, we need to quantify the differences in the plans. To this end, we represent the available plans from an ISP in a city using a plans vector of 30 dimensions, each representing a discrete carriage value.[7] We then quantify the differences using the L1 norm between the two vectors. The weight for each dimension is determined by the fraction of block groups in the city that receive that specific carriage value, and the `ceil` operator is used to discretize the carriage values. For example, Cox offers a carriage value of around 10.5 and 11.3 in 35% and 12% of block groups in New Orleans, 12% and 6% of block groups in Oklahoma City, and 4% and 21% of block groups in Wichita. The L1 norm between New Orleans and Oklahoma City plans is 1.78 (different). Between New Orleans and Wichita, is 1.57 (different), and between Oklahoma City and Wichita is 0.36 (relatively similar).

6.5.2 Inter-City Broadband Plans

To answer ❶ (do the broadband plans, characterized by their carriage value, change by city for different ISPs?), we analyze the distribution of plans at block group granularity. We only visualize one major provider from each DSL/fiber (AT&T) and cable (Cox) category for brevity. To simplify the exposition, Figure 6.7 shows the distribution of carriage value for only five cities (out of 14 and 6, respectively) for each ISP.

For AT&T, we observe two sets of peaks in broadband plans. The higher carriage value peak is attributable to fiber-based plans and the lower to DSL-based plans. The

[7]Note that the maximum carriage value we observed across all ISPs and cities is 28.6 (Table 6.2).

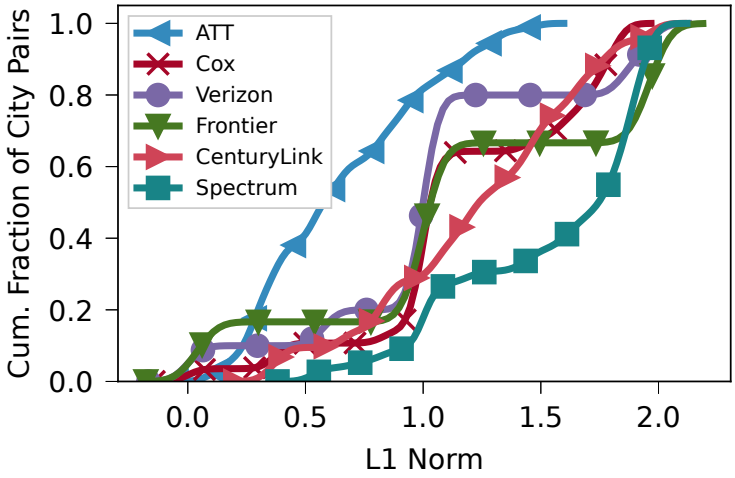
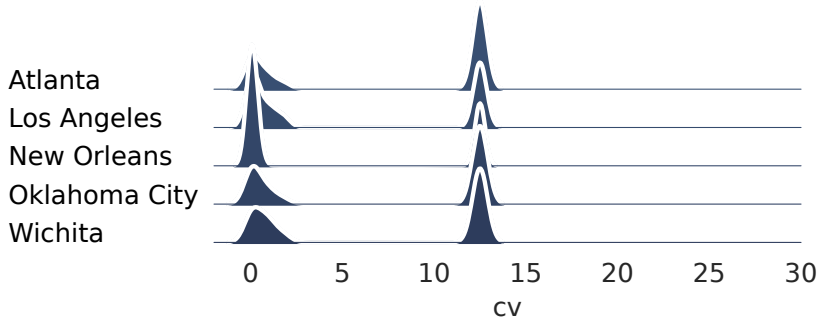


Figure 6.6: Distribution of difference in ISP plans across different city pairs. A higher L1 norm indicates more diverse offerings.

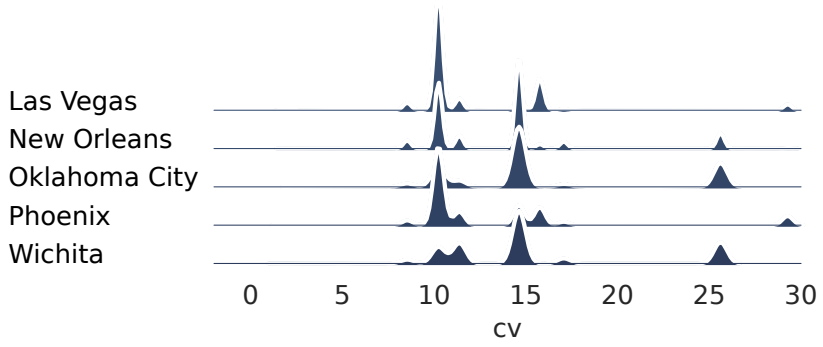
fraction of block groups that receive fiber plans differs in each of the cities. For example, in New Orleans, 32% of block groups receive fiber-based access, which is significantly smaller than the 54% and 57% of block groups in Wichita and Oklahoma City.

For Cox, we observe six different peaks, and the distribution of the carriage values across block groups varies significantly by city. For example, Cox offers *cv* of about 28 Mbps/$ to 7% of block groups in New Orleans. In contrast, Cox offers similar plans to 21% and 18% of block groups in Oklahoma City and Wichita, respectively. On the other hand, 44%, 46%, and 50% of block groups in Wichita, New Orleans, and Oklahoma City receive *cv* of 14.6 Mbps/$.

To illustrate how this trend generalizes for other cities and ISPs, Figure 6.6 shows the distribution of L1 norm, i.e., the difference in available plans between all pairs of served cities for each ISP. A low L1 norm indicates similarities in broadband plans and vice versa. We observe that DSL/fiber-based provider plans are less diverse across different cities than cable-based providers, with AT&T (most similar) and Spectrum (most diverse) at the extremes. This result demonstrates that some ISPs alter their plans between cities



(a) AT&T (DSL/fiber)



(b) Cox (cable)

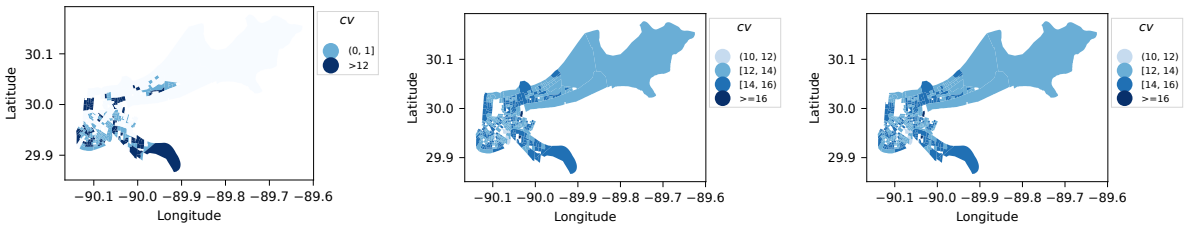
Figure 6.7: Distribution of broadband plans in different cities for two major ISPs.

while others maintain consistent offerings throughout their service areas.

6.5.3 Intra-City Broadband Plans

To answer ❷ (does the carriage value change within a city? If yes, which neighborhoods (identified by their census block groups) receive good and bad deals (high and low carriage values)?), we analyze broadband plans within each city. At a high level, Figure 6.7 shows that ISPs offer disparate plans to users within a city. These differences in *cv* can be as high as **600%** for DSL/fiber and **92%** for cable-based providers.

Individual and composite plans. To better understand broadband plans within



(a) AT&T (b) Cox (c) Best of Cox and AT&T

Figure 6.8: Spatial distribution of broadband plans in New Orleans. All three scenarios are spatially clustered. Darker shades indicate block groups with higher *cv*.

a city, we zoom in on Cox and AT&T in New Orleans, individually and as a pair (see Figure 6.8c). Comparing Figures 6.8a and 6.8b, we observe that Cox offers better coverage and higher carriage values than AT&T in most block groups.

Given its lower proliferation of high *cv* fiber plans, if we look at the plans only from AT&T in this city, which was the case in one of the previous studies [73], we might get an impression that the nature of broadband plans is problematic for all New Orleans residents. Specifically, the broadband plans are sparse and highly variable (DSL vs. fiber), and most residents get the "worst" deal, i.e., low carriage values. However, the competing cable-based provider is the dominant ISP in the city, and its plans are not as extreme nor sparse. Figure 6.8c shows that if we consider the AT&T and Cox plans together, i.e., when we report the highest carriage value from either of the two providers, the best carriage value is similar to that of the dominant cable-based ISP, i.e., Cox in this case. We make similar observations for other cities as well. In our dataset, we do not find a case where the DSL/fiber-based providers offer better coverage or higher average carriage values than the cable-based providers.

Spatial clustering. We visually observe that broadband plans are clustered, i.e., the likelihood that two contiguous block groups have similar available plans is high. To validate this visual understanding, we compute the spatial autocorrelation metric using

Individual ISPs						
1	2	3	4	5	6	7
0.34	0.52	0.33	0.45	0.23	0.35	0

ISP Pairs									
1-5	1-6	3-5	3-6	4-5	2-5	2-6	1-7	2-7	3-7
0.23	0.35	0.23	0.35	0.23	0.23	0.35	0	0	0

Table 6.3: Statistical evidence for spatial clustering. We report the median of Moran I statistics across all cities.

Moran's I method [91] to characterize the extent of correlation in carriage values among nearby block groups. This metric has been widely used in previous studies [226, 176] to understand the spatial distribution of a variable of interest (i.e., carriage value) within a geographic region (i.e., city). A positive value of Moran's I statistic means that similar carriage values tend to be found near each other, while a negative value means dissimilar values are found near each other, with zero indicating a complete lack of association of carriage values with locations.

We computed the Moran's I statistic for all (ISP, city) pairs to measure the spatial autocorrelation of broadband plans. The results show that, with the exception of Xfinity, the median value ranges between 0.3–0.5, indicating a high level of spatial clustering in broadband plans across ISPs within a city. Given that AT&T is a DSL/fiber-based provider, such clustering of its carriage value can be attributed to its fiber infrastructure deployment around the city. Table 6.3 reports the median value across all cities for each ISP.

Our results show that both DSL/fiber and cable ISPs offer similar *cv* plans to neighboring census block groups within a city. Similar to the case for AT&T, the spatial clustering of plans for DSL/fiber providers is related to the nature of access technology. Neighborhoods with fiber deployments receive better carriage value and vice versa. However, since cable-based ISPs use the same technology across the city, spatial clustering in their plans is intriguing. In the next section, we explore whether this behavior is

attributable to competition among ISPs.

6.5.4 Impact of Competition

To answer ❸ (does competition among ISPs impact the carriage value offered to the end users? If yes, is there a trend in which neighborhoods experience competition?), we explore whether the cable-based ISP's plans change when they operate as a monopoly vs. when they compete as a duopoly. We did not analyze DSL/fiber-based providers alone from the perspective of operating as both a monopoly and a duopoly because we did not observe this pattern in any of the thirty cities. We employ a statistical test to discern whether competition (or lack thereof) leads to a change in cable providers' carriage value. For every city with competition between cable and DSL/fiber providers, we run two one-tailed 2-sample Kolmogorov–Smirnov (KS) tests [117].

Our null hypothesis ($H0$) is that there is no difference in the *cv* offered by a cable provider in locations where they operate as a cable monopoly compared to locations where they operate as a cable-DSL duopoly or cable-fiber duopoly. To test this hypothesis, we run one test for each of the following alternate hypotheses (H).

In the first one-tailed test, we propose $H1$, which states that the *cv* provided by the cable provider is greater for block groups in duopoly locations than those in cable monopoly locations. In the second test, we reverse the hypothesis from the previous test and propose $H2$, which states that cable providers provide better *cv* for block groups in cable monopoly locations than those in each duopoly category. By conducting two tests per category, we can detect either scenario and provide robust statistical evidence of the impact of competition on cable offerings for different types of DSL/fiber-based offerings.

If we achieve a p-value of less than 0.05, we reject the null hypothesis ($H0$) for the corresponding test and record the corresponding KS test statistic, denoted by the D

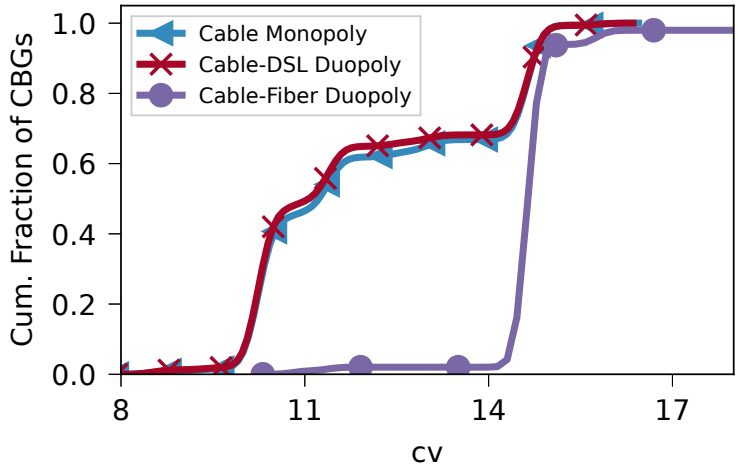


Figure 6.9: Distribution of carriage value for Cox in its three operational modes in New Orleans. To simplify exposition, we prune the long tail that is attributable to block groups that receive subsidized broadband access through the ACP plan [154].

value. We conduct this analysis for all combinations of cable and DSL/fiber providers in other cities. In the remainder of the section, we use New Orleans as a case study to explain our findings.

Cable-DSL Duopoly: In the first test, our $H1$ is Cox's *cv* in cable monopoly block groups is lower than the cable-DSL duopoly block groups. Conversely, our $H2$ is Cox's *cv* in cable monopoly block groups is higher than cable-DSL duopoly block groups. Figure 6.9 shows that Cox's offered *cv* in the DSL duopoly block groups is similar to its *cv* in monopoly block groups. This is further confirmed through the K-S test, where we fail to reject $H0$, which signifies there is no statistical difference in Cox's *cv* distribution in block groups where it serves alone and block groups where it competes with AT&T's DSL offerings. The median *cv* for both cases is 11.38 Mbps/$. We observe the same trend for other pairs of Cable-DSL duopolies within cities in our dataset.

Cable-Fiber Duopoly: We posit a similar hypothesis for cable-fiber duopolies. Figure 6.9 shows the difference in Cox's *cv* distributions between these block group types, which is further reinforced by the K-S test where we reject $H0$ with statistical signifi-

cance in favor of $H1$. Contrarily, $H2$ cannot be accepted as the p-values exceed 0.05. This result points towards Cox increasing the *cv* provided through its plans by lowering the price for the same download speed in block groups where it faces competition from AT&T's higher *cv* fiber offerings. The median *cv* from Cox in such addresses is 14.63 Mbps/$, 30% more than the monopoly and DSL block groups' median *cv*. For the remaining combinations of cable and DSL/fiber providers in other cities, we capture the same trend, indicating differential pricing structures from cable providers in the presence of high *cv* competition.

Our analysis in this section has demonstrated that cable providers tend to improve the carriage value offered through their plans in locations where fiber-based plans are present. This places fiber plans in a critical position because they tend to yield better broadband deals.

6.5.5 Influencing Socioeconomic Factors

In the prior sections, we established that low *cv* is associated with DSL plans. In this section, we investigate whether there is a trend in which sociodemographic groups predominantly receive DSL plans and, therefore, worse *cv*. This analysis will enable us to answer ❹ (is the quality of available deals affected by demographic and socioeconomic factors? If yes, which population groups receive better or worse deals from the ISPs?). To do so, we compute the percentages of block groups within every city that receive DSL or fiber plans disaggregated by the block group level median household income. The American Community Survey (ACS) [68] publicly releases this information through a 5-year dataset. Although the demographic information for the 2020 census survey is available, it is known to have a significantly lower number of responses due to the COVID-19 pandemic [120]; hence we utilize the 2019 dataset. We merge our dataset with the ACS data

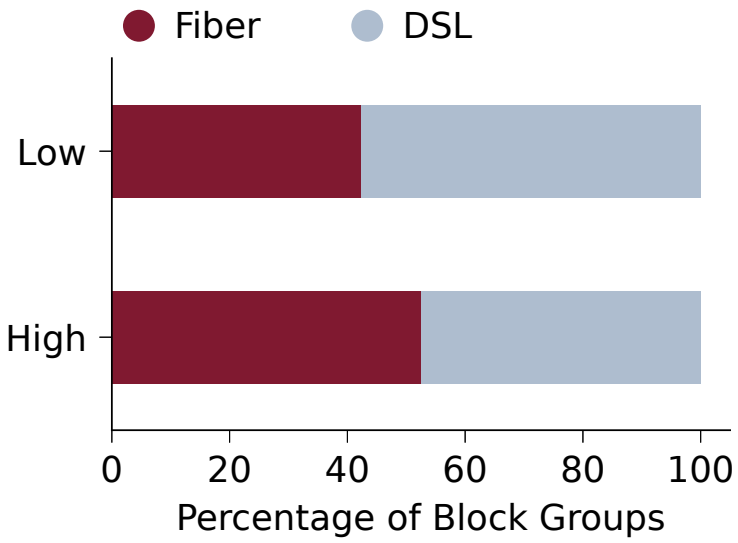


Figure 6.10: The percentage of AT&T's DSL/fiber deployment in terms of addresses served by the two technology types, disaggregated by income level in New Orleans.

to obtain the median household income of every census block group.

Concretely, we adopt a methodology similar to [132, 73] to group each city's census block group-level income into two distinct categories: low (below the city's median household income) and high (exceeding the city's median household income). For each income group class within a city, we calculate the percentage of block groups that have access to fiber-based plans. Subsequently, we compute the percentage difference in fiber deployment between the high and low-income groups of the block group.

Figure 6.10 presents the breakdown of the percentage of block groups that receive AT&T's DSL and fiber plans in the two income categories of block groups in New Orleans. 41% of the low-income census block groups receive AT&T's fiber plans while 57% of the high-income block groups in have fiber plans available.

In the 14 cities where we collected AT&T plan data, the fiber deployment gap between the high-income and low-income block groups exceeds 10% in seven cities, while in four cities, it is below 10%. No difference is observed in Austin, TX; however, in Wichita, KS,

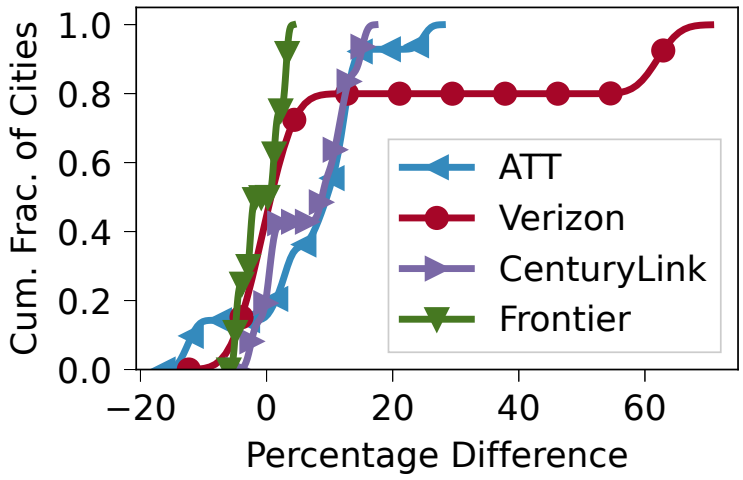


Figure 6.11: The overall distribution of the percentage difference in fiber deployment between high-income and low-income block groups across all cities and ISPs.

and Atlanta, GA, a higher proportion of low-income census block groups receive fiber from AT&T compared to high-income groups. Figure 6.11 shows that CenturyLink and Verizon exhibit a comparable pattern, where a larger proportion of high-income block groups across cities receive fiber compared to lower-income groups. Frontier emerges as an outlier in this analysis.[8] Given that the lack of fiber also leads to lower *cv* from cable providers, internet users in block groups that lack fiber connectivity tend to get more bad deals overall compared to others.

We conducted a similar analysis for the demographic attributes of race and population density. The results for these variables did not produce comparable trends.

[8]It is worth noting that in 2020, Frontier declared bankruptcy and received financial assistance from the US. Federal Communications Commission to enhance its fiber connectivity for millions of households [48]. Despite claiming to utilize these funds for the stated purpose, Frontier was found guilty by the Federal Trade Commission of deceiving and overcharging its customers [81]. This highlights the importance of extending the scope of our study and investigating the actual price subscribers pay for ISP service.

6.6 Related Work

In [31], the authors analyzed FCC Form 477 data and reported that close to 50 million people in the US live in locations served by a single ISP, i.e. in an ISP monopoly. While not considering the price/ cost associated with internet access, several studies have sought to understand how internet quality itself varies between different locations and demographic variables. The Census Bureau produces an annual list of US cities with the lowest Internet connectivity using data from the American Community Survey (ACS) One Year estimates [107]. However, this estimate does not take into account the cost of access. The work conducted in [179] demonstrated that the FCC National Broadband Report significantly overestimates coverage and examined the digital divide in terms of the lack of coverage in rural and marginalized communities. Similar inaccuracies of the FCC map were found for mobile networks in [181]. In [22], the authors analyzed the relationship between income and download speed at the geographic granularity of US zip codes. 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. The authors of [132] analyzed publicly available data from Ookla [93], a popular speed test vendor, and found significant differences in key internet quality metrics between communities with different income levels. In [196], the authors utilized M-Lab [56] speed test data in California and found higher internet quality in urban and high-income areas.

Several studies have also examined how the cost of electricity varies across locations and demographic variables. The authors in [211] discovered that minority groups in various US cities pay a disproportionate amount for electricity compared to other communities. Similar findings are reported in [92].

6.7 Conclusion

In this work, we explore broadband affordability in the US. Specifically, we analyze the nature of broadband plans offered by seven major ISPs across thirty different US cities. To aid this effort, we developed BQT, a new scalable tool that extracts broadband plans offered by the seven major US ISPs at any street address. We use this tool to curate a dataset that reports broadband plans offered to 837 k street addresses, spanning 18 k census block groups in the thirty cities. To the best of our knowledge, this is the largest such broadband plan pricing dataset in existence. Our analysis sheds light on pricing strategies adopted by different ISPs, which have previously been opaque. Our results highlight the importance of competition, and specifically on how fiber deployments benefit end users. It also identifies the population groups reaping the benefits of competition and fiber deployments. We believe this effort is a step towards improving public understanding of US broadband affordability. We will make our tool and dataset publicly available to facilitate further research.

Chapter 7

Conclusion, Future Directions, and Recommendations

7.1 Conclusion

Internet access is crucial for success in today's world, but many population groups continue to remain excluded from it, resulting in digital inequity or a digital divide. Simply having Internet services available does not ensure usability; it must also be both high-quality and affordable. In 2022, the US authorized the investment of billions of taxpayers' money through the BEAD [187] program to improve Internet infrastructure in underserved areas, aiming to reduce the digital divide. The success of the BEAD program in achieving its objectives hinges upon identifying the areas experiencing digital inequity in Internet access. As Internet access encompasses availability, quality, and affordability, the funding must be directed to regions experiencing digital inequities in one or more of these dimensions across the country. Yet, the lack of high-quality, fine-grained, accurate, and comprehensive datasets on Internet access currently hinders a complete understanding and assessment of the extent and prevalence of digital inequity, making it challenging

to identify areas that would benefit most from funding efforts.

This dissertation makes significant strides in advancing the collective comprehension of digital inequity in Internet access. It focuses on the characterization and analysis of current Internet availability and quality datasets, uncovering critical patterns in regions facing digital inequity. Moreover, the dissertation highlights the limitations of existing Internet access datasets, such as inaccuracies, quality issues, or insufficient quantity. Additionally, it reveals the potential risks of using these datasets as-is to identify areas affected by digital inequity. Critically, the dissertation proposes new methodology and tools that are able to i) improve the usability of existing Internet quality datasets and ii) curate novel datasets of Internet availability and affordability. The contributions in this dissertation make important steps towards analyzing and understanding the current state of digital inequity in the US.

Drawing on the dissertation's findings, we conclude with a brief discussion on future research directions. We also put forth some recommendations for relevant stakeholders for further exploring digital inequity and implementing effective actions to eventually mitigate it.

7.2 Future Directions

Passively Infer Contextual Information from Speed Tests. The work in Chapter 5 demonstrated the critical role played by various contextual information such as signal strength and device memory on the eventual Internet quality reported by speed tests. While our study focused on speed tests carried out on Android devices that provide this contextual information, speed tests conducted on different devices like iOS phones and desktops do not offer the ability to gather such details. New methodologies capable of extracting such information passively, when a user conducts a speed test, are therefore

needed. A potential approach could be to conduct controlled lab experiments where contextual information is known and the packets transferred during speed tests are captured. As speed test vendors can capture packets from the server end, various statistics from the packet captures such as packet inter-arrival times and number of reordered packets can be extracted. These statistics can then be used as features to train machine learning models to ultimately predict contextual information such as whether a speed test was run over a wired or wireless connection.

Expanding BQT's Functionalities and Footprint. BQT, proposed in Chapter 6, enabled the curation of the most comprehensive dataset of the cost of Internet access in the US. However, the functionalities of the BQT can be enhanced to make the system more dynamic and scalable. At present, BQT is run with 100 Docker containers to ensure scalability. This number could be adjusted at runtime by observing the query response times when interacting with ISP BATs. Additionally, as ISPs arm their BATs with upgraded features to prevent scraping, an interesting research question emerges: how can we integrate additional approaches like simulating realistic user mouse movements on websites into the current BQT to bypass the prevention mechanisms? Another potential challenge stems from the task of identifying ISPs that offer queryable BATs. For BQT to increase its current footprint and support hundreds of ISPs, it is crucial to implement an automated method for pinpointing these ISPs.

Network performance analysis. Poor network performance can impede a user's ability to make the best use of Internet connectivity, and poor performance may also discourage adoption. Understanding ISP performance is especially important in census blocks with only a single high-speed ISP (a significant proportion of census blocks in the case of Chicago as mentioned in Chapter 2), where incentives to maintain and upgrade infrastructure may be less. Unfortunately, crowdsourced performance datasets, such as Ookla and Measurement Lab's speed test datasets, are biased in a variety of ways, in-

cluding lacking data from under-connected communities [132, 182]. Future research could focus on measuring performance with more extensive targeted sampling from neighborhoods of interest [178, 213]. The actual performance can then be compared with the ISP-advertised performance to understand if the promised services are being delivered by the ISPs. Beyond speed, future research can be conducted to develop scalable platforms capable of measuring the quality of experience (QoE) of users while interacting with different applications such as YouTube and Zoom.

7.3 Recommendations

Augmenting FCC map with Pricing Information. The FCC should consolidate the broadband availability maps [78] and urban rate survey [156] to ensure that the public has access to both availability and pricing information at the street address level. Based on our findings in Chapter 6, it is evident that the speed offered by an ISP is a crucial factor to take into account. However, there is significant variability in the prices at which these speeds are offered to customers. Moreover, it is essential to assess the complete, actual costs incurred by subscribers for these ISP services. Previous work [104] has documented that ISPs frequently include extra fees and charges in their pricing structure. If such comprehensive information is collected by the FCC and subsequently made public, these pricing strategies can be better studied, decreasing the lack of transparency that currently exists within the ISP service provider sector.

Collecting Internet Quality Data with Proper Contexts. Beyond the availability and cost of access, actual performance data about fixed broadband service is critical for fully characterizing digital inequality, yet it remains elusive. If ISPs are mandated to provide information about actual performance experienced by their subscribers, we can complement the research presented in this dissertation to understand not just what

service ISPs promise to deliver, but what service they actually do deliver. Additionally, as part of the Broadband DATA Act [32], the FCC has outlined and continues to refine a process for consumers to challenge fixed and mobile provider coverage claims. In this challenge process, consumers can submit speed test measurements taken from specified tools. This dissertation has identified critical metadata that we believe must accompany each measurement. It is possible to collect some of these metrics, such as access link, WiFi RSSI, etc., without user-level intervention. However, extracting all the recommended metadata for all end hosts might not be possible depending on their operating systems and browsers. Nevertheless, the measurement platforms should collect as much contextual information as possible to better understand the speed test measurements. Though it is possible to infer the subscription plan, we recommend collecting this information from as many users as possible. Our recommendation is motivated by the observation made in Chapter 5 that subscription plans play a critical role in assessing Internet quality in a region. Importantly, *we believe the context we recommend must be coupled to (i.e., publicly accessible with) measurement results as meta-data so that such measurements can be properly analyzed and contextualized.* Note that we do not claim our work to be an all-inclusive list of needed context. Other factors, such as the make and model of the cable modem or additional relevant home router information, are likely also essential. Finally,*we encourage all speed test vendors who wish to create platforms for such coverage challenges to ensure that the speed test is constructed so that it maximizes the throughput of the measured path.* Designing such test methodologies, especially for high-speed access links, is non-trivial and requires further exploration [153].

Curating Accurate Street Addresses. Even if the FCC provides information regarding Internet access, third-party audits are essential to verify the accuracy of self-reported information from ISPs. However, existing US street address datasets are private, sparse, and noisy, posing a challenge to such third-party efforts. Therefore, local governments

(e.g., county) should put more effort into improving the quality and availability of street address datasets in their areas. This will enable and encourage additional research within this field, consequently leading to a more comprehensive understanding of facets related to ISP service provisioning.

Policy Intervention. Finally, policymakers should consider subsidizing fiber deployment efforts [63] or enforcing rate regulations [40], even in urban areas, to help improve the carriage value for broadband plans in low-income block groups that can be ignored or deprioritized by major ISPs. This would improve competition and carriage value, as our work in Chapter 6 has demonstrated that fiber deployments play a critical role in providing subscribers with the option of high carriage value plans from different types of ISPs.

Bibliography

[1] Understanding the urban digital divide. URL: https://bipartisanpolicy.org/blog/urban-broadband-blog/.

[2] Land Area and Persons Per Square Mile, 2010. URL: https://www.census.gov/quickfacts/fact/note/US/LND110210.

[3] TIGER Line Shapefile, 2010. URL: https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2010&layergroup=Blocks.

[4] In Sandy's wake, here's why millions of Americans have cell service but no power, 2012. [Online; accessed 23-September-2019]. URL: https://qz.com/21909/hurricane-sandy-and-cell-phone-network-service.

[5] A-CAM Census Block Groups PN, 2015. URL: https://www.fcc.gov/document/cam-census-block-groups-pn.

[6] Residential Fixed 25 Mbps/3 Mbps Broadband Deployment, 2016. URL: https://www.fcc.gov/reports-research/maps/bpr-2016-fixed-25mbps-3mbps-deployment/.

[7] Tier flattening: At&t and verizon home customers pay a high price for slow internet, 2018. URL: https://www.digitalinclusion.org/wp-content/uploads/2018/07/NDIA-Tier-Flattening-July-2018.pdf.

[8] When you can't trust the data, flaws in the federal communications commission's broadband forms, 2018. URL: https://ilsr.org/when-you-cant-trust-the-data-flaws-in-the-federal-communications-commissions-broadband-forms/.

[9] America's Digital Divide, 2019. URL: https://www.pewtrusts.org/en/trust/archive/summer-2019/americas-digital-divide.

[10] Crimson Hexagon. https://www.brandwatch.com/#from-ch, 2019.

[11] Hurricane Maria — Wikipedia, the free encyclopedia, 2019. [Online; accessed 23-September-2019]. URL: https://en.wikipedia.org/wiki/Hurricane_Maria.

[12] IBM Watson. https://www.ibm.com/watson/services/natural-language-understanding, 2019.

[13] Keras. https://github.com/keras-team/keras, 2019.

[14] Labelbox. https://labelbox.com, 2019.

[15] List of United States hurricanes — Wikipedia, the free encyclopedia, 2019. [Online; accessed 17-September-2019]. URL: https://en.wikipedia.org/wiki/List_of_United_States_hurricanes.

[16] Newspaper3k. https://newspaper.readthedocs.io/en/latest/, 2019.

[17] Tensorflow code and pre-trained models for BERT. https://github.com/google-research/bert, 2019.

[18] Tuning the hyper-parameters of an estimator, 2019. [Online; accessed 23-September-2019]. URL: https://scikit-learn.org/stable/modules/grid_search.html.

[19] America's Racial Gap & Big Tech's Closing Window, 2020. URL: https://www.db.com/newsroom_news/.

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

[21] COVID-19 and the rise of Telemedicine, 2020. URL: https://medicalfuturist.com/covid-19-was-needed-for-telemedicine-to-finally-go-mainstream/.

[22] Decoding the digital divide, 2020. URL: https://www.fastly.com/blog/digital-divide.

[23] During coronavirus, high-speed internet is a lifesaver - that millions lack, 2020. URL: https://www.nbcnews.com/think/opinion/during-coronavirus-high-speed-internet-lifesaver-millions-lack-ncna1165321.

[24] Expanding Internet Access Improves Health Outcomes, 2020. URL: https://www.govtech.com/network/Expanding-Internet-Access-Improves-Health-Outcomes.html.

[25] FCC Improves Broadband Data and Maps to Bridge the Digital Divide, 2020. URL: https://www.fcc.gov/document/fcc-improves-broadband-data-and-maps-bridge-digital-divide-0.

[26] FCC Proposes the 5G Fund for Rural America, 2020. URL: https://www.fcc.gov/document/fcc-proposes-5g-fund-rural-america.

[27] FCC Reports Broadband Unavailable to 21.3 Million Americans, BroadbandNow Study Indicates 42 Million Do Not Have Access, 2020. URL: https://broadbandnow.com/research/fcc-underestimates-unserved-by-50-percent.

[28] FCC underestimates americans unserved by broadband internet by 50%, 2020. URL: https://broadbandnow.com/research/fcc-underestimates-unserved-by-50-percent.

[29] Governor Gavin Newsom Issues Stay at Home Order, 2020. URL: https://www.gov.ca.gov/2020/03/19/governor-gavin-newsom-issues-stay-at-home-order/.

[30] Guidelines for Broadband Data Submission, 2020. URL: https://www.cpuc.ca.gov/industries-and-topics/internet-and-phone/broadband-mapping-program/guidelines-for-broadband-data-submission.

[31] Profiles of monopoly: Big cable and telecom, 2020. URL: https://cdn.ilsr.org/wp-content/uploads/2020/08/2020_08_Profiles-of-Monopoly.pdf.

[32] S.1822 - Broadband DATA Act. https://www.congress.gov/bill/116th-congress/senate-bill/1822, 2020.

[33] Social Distancing, Internet Access and Inequality, April 2020. URL: https://www.nber.org/papers/w26982.pdf.

[34] The Results Are In for Remote Learning: It Didn't Work, 2020. URL: https://www.wsj.com/articles/schools-coronavirus-remote-learning-lockdown-tech-11591375078.

[35] U.S. Broadband Usage Percentages, 2020. URL: https://github.com/microsoft/USBroadbandUsage\Percentages.

[36] U.S. Schools Trying to Teach Online Highlight a Digital Divide, 2020. URL: https://www.bloomberg.com/news/articles/2020-03-26/covid-19-school-closures-reveal-disparity-in-access-to-internet.

[37] US's Digital Divide 'is going to kill people' as COVID-19 exposes Inequalities, 2020. URL: https://www.theguardian.com/world/2020/apr/13/coronavirus-covid-19-exposes-cracks-us-digital-divide.

[38] ArcGis, 2021. URL: https://www.arcgis.com/index.html.

[39] archive-measurement-lab, 2021. URL: https://console.cloud.google.com/storage/browser/archive-measurement-lab/ndt/ndt7/?forceOnBucketsSortingFiltering=false&project\=measurement-lab.

[40] Assembly Bill A6259A, 2021. URL: https://www.nysenate.gov/legislation/bills/2021/A6259#:~:text=Requires%20broadband%20providers%20to%20offer,if%20proper%20notice%20is%20given.

[41] Census block, 2021. URL: https://en.wikipedia.org/wiki/Census_block.

[42] Census blocks and block groups, 2021. URL: https://www2.census.gov/geo/pdfs/reference/GARM/Ch11GARM.pdf.

[43] Check your connection., 2021. URL: https://projectstream.google.com/speedtest.

[44] Computer and Internet Use, 2021. URL: https://www.census.gov/content/census/en/programs-surveys/acs/library/keywords/computer-and-internet-use.html/.

[45] ESRI, 2021. URL: https://en.wikipedia.org/wiki/Esri.

[46] ESRI's Demographics: #1 for Accuracy, 2021. URL: https://www.esri.com/library/fliers/pdfs/esris-demographics-accuracy.pdf.

[47] Fact sheet: Executive order on promoting competition in the american economy, 2021. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2021/07/09/fact-sheet-executive-order-on-promoting_competition_in-the_american-economy/.

[48] Frontier exits bankruptcy, claims it will double fiber-to-the-home footprint, 2021. URL: https://arstechnica.com/information-technology/2021/05/frontier-exits-bankruptcy-claims-it-will-double-fiber-to-the-home-footprint/.

[49] Haversine formula, 2021. URL: https://en.wikipedia.org/wiki/Haversine_formula.

[50] Input sought on mobile challenge, verification technical requirements, 2021. URL: https://www.fcc.gov/document/input-sought-mobile-challenge-verification-technical-requirements.

[51] Internet/Broadband Fact Sheet, 2021. URL: https://www.pewresearch.org/internet/fact-sheet/internet-broadband/.

[52] IPinfo.io, 2021. URL: ipinfo.io.

[53] Locate API Usage, 2021. URL: https://github.com/m-lab/locate/blob/master/USAGE.md.

[54] Locate API v1, 2021. URL: https://www.measurementlab.net/develop/locate-v1/.

[55] Locate API v2, 2021. URL: https://www.measurementlab.net/develop/locate-v2/.

[56] Measurement Lab, 2021. URL: https://www.measurementlab.net/.

[57] Mlab test your speed, 2021. URL: https://speed.measurementlab.net/#/.

[58] NDT-Network diagnostic tool, 2021. URL: https://www.measurementlab.net/tests/ndt/ndt7/.

[59] Pearson Correlation Coefficient, 2021. URL: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

[60] Point-biserial Correlation, 2021. URL: 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.

[61] Relating block groups to zip code areas, 2021. URL: http://proximityone.com/bg-zip.htm.

[62] Research 101: Census tracts vs. census block groups, 2021. URL: https://current360.com/research-101-census-tracts-vs-census-block-groups/.

[63] Senate Bill No. 156, 2021. URL: https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202120220SB156.

[64] Updated demographics, 2021. URL: https://doc.arcgis.com/en/esri-demographics/data/updated-demographics.html.

[65] VAN BUREN v. UNITED STATES, 2021. URL: https://www.supremecourt.gov/opinions/20pdf/19-783_k53l.pdf.

[66] 9-48.000 - COMPUTER FRAUD AND ABUSE ACT, 2022. URL: https://www.justice.gov/opa/press-release/file/1507126/download.

[67] Absent New FCC Broadband Maps, Local Govs Plot Coverage. https://www.govtech.com/network/absent-new-fcc-broadband-maps-local-govs-plot-coverage, 2022.

[68] American community survey 5-year data (2009-2021), 2022. URL: https://www.census.gov/data/developers/data-sets/acs-5year.html.

[69] Broadband availability and access. https://www.rural.pa.gov/publications/broadband, 2022.

[70] Broadband Speed: FCC Map Vs. Experience on the Ground. https://dailyyonder.com/broadband-speed-fcc-map-vs-experience-ground/2018/07/25/, 2022.

[71] Cable internet in the usa, 2022. URL: https://broadbandnow.com/Cable.

[72] Calspeed a home broadband study. https://www.calspeed.net/index.html, 2022.

[73] Dollars to megabits, you may be paying 400 times as much as your neighbor for internet service, 2022. URL: https://themarkup.org/still-loading/2022/10/19/dollars-to-megabits-you-may-be-paying-400-times-as-much-as-your-neighbor-for-internet-service.

[74] DPV — PostalPro, 2022. URL: https://postalpro.usps.com/address-quality/dpv.

[75] Expectation–maximization algorithm, 2022. URL: https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm.

[76] FAST. https://fast.com/, 2022.

[77] FCC Form 477 Local Telephone Competition and Broadband Reporting, 2022. URL: https://us-fcc.app.box.com/v/Form477Instructions.

[78] FCC National Broadband Map, 2022. URL: https://broadbandmap.fcc.gov/home.

[79] Federal Communications Commission NOTICE OF PROPOSED RULEMAKING MB Docket No. 22-13. https://docs.fcc.gov/public/attachments/FCC-21-20A1.pdf, 2022.

[80] Fixed Broadband Deployment Data from FCC Form 477. https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477, 2022.

[81] FTC Takes Action Against Frontier for Lying about Internet Speeds and Ripping Off Customers Who Paid High-Speed Prices for Slow Service, 2022. URL: https://www.ftc.gov/news-events/news/press-releases/2022/05/ftc-takes-action-against-frontier-lying-about-internet-speeds-ripping-customers-who-paid-high-speed.

[82] Grow north encourages people to take internet speed test to help improve broadband infrastructure in the region. https://www.wxpr.org/business-economics/2022-03-29/grow-north-encourages-people-to-take-internet-speed-test-to-help-improve-broadband-infrastructure-in-the-region, 2022.

[83] HiQ Labs, Inc. v. LinkedIn Corp., 2022. URL: https://casetext.com/case/hiq-labs-inc-v-linkedin-corp-5.

[84] H.R.3684 - Infrastructure Investment and Jobs Act. https://www.congress.gov/bill/117th-congress/house-bill/3684/text, 2022.

[85] Ingham county asks residents and businesses to participate in survey on broadband internet access and speed. https://www.prnewswire.com/news-releases/ingham-county-asks-residents-and-businesses-to-participate-in-survey-on-broadband-internet-access-and-speed-301518936.html, 2022.

[86] Kernel density estimation, 2022. URL: https://en.wikipedia.org/wiki/Kernel_density_estimation.

[87] Measuring Broadband America. https://www.fcc.gov/general/measuring-broadband-america, 2022.

[88] Measuring Broadband Raw Data Releases - Fixed. https://www.fcc.gov/oet/mba/raw-data-releases, 2022.

[89] Millions of Americans can't get broadband because of a faulty FCC map. There's a fix. https://www.cnet.com/home/internet/features/millions-of-americans-cant-get-broadband-because-of-a-faulty-fcc-map-theres-a-fix/, 2022.

[90] Mixture model, 2022. URL: https://en.wikipedia.org/wiki/Mixture_mode.

[91] Moran's i, 2022. URL: https://en.wikipedia.org/wiki/Moran%27s_I.

[92] Race and energy poverty: Evidence from african-american households. *Energy Economics*, 108:105908, 2022.

[93] Speedtest. https://www.speedtest.net/, 2022.

[94] Speedtest by ookla global fixed and mobile network performance map tiles. https://github.com/teamookla/ookla-open-data, 2022.

[95] Speedtest is the world's #1 internet utility. https://www.ookla.com/analysis, 2022.

[96] The FCC's broadband map won't be ready for a year. This data company has already built one. https://www.cnet.com/home/internet/the-fccs-broadband-map-wont-be-ready-for-a-year-this-data-company-has-already-built-one/, 2022.

[97] The internet as a human right, 2022. URL: https://www.un.org/en/about-us/universal-declaration-of-human-rights.

[98] The Markup, 2022. URL: https://themarkup.org/.

[99] The Speedtest Server Network™, 2022. URL: https://www.ookla.com/network.

[100] Unit Profile. http://data.fcc.gov/download/measuring-broadband-america/2021/unit-profile-sept2020.xlsx, 2022.

[101] USA, actively participate in FCC policy-making! https://samknows.com/blog/usa-actively-participate-in-FCC-policy-making, 2022.

[102] Welcome to speedsurvey: presented by the state of alabama. https://www.google.com/search?q=ctc+alabama+speed+test&oq=ctc+alabama+speed+test&aqs=chrome..69i57j33i160.6405j0j4&sourceid=chrome&ie=UTF-8, 2022.

[103] Xfinity Speed Test. https://speedtest.xfinity.com/, 2022.

[104] You May Be Paying Too Much for Your Internet, 2022. URL: https://www.consumerreports.org/electronics-computers/telecom-services/you-may-be-paying-too-much-for-your-internet-a7157329937/.

[105] Zillow's Transaction and Assessment Database (ZTRAX), 2022. URL: https://www.zillow.com/research/ztrax/.

[106] After defending false data, comcast admits another fcc broadband map mistake, 2023. URL: https://arstechnica.com/tech-policy/2023/02/comcast-could-have-avoided-giving-false-map-data-to-fcc-by-checking-its-own-website/.

[107] American community survey 1-year data (2005-2021), 2023. URL: https://www.census.gov/data/developers/data-sets/acs-1year.html.

[108] Bridging the Divide: Answering Internet Policy Questions with Cutting-Edge Network Measurement Algorithms, Datasets, and Platforms, 2023. URL: https://www.law.berkeley.edu/research/bclt/bcltevents/bridging-the-divide-answering-internet-policy-questions-with-cutting-edge-network-measurement-algorithms-datasets-and-platforms/.

[109] Bright Data, 2023. URL: https://brightdata.com/.

[110] BroadbandNow, 2023. URL: https://broadbandnow.com/.

[111] California Community Foundation, 2023. URL: https://www.calfund.org/.

[112] California Public Utilities Commission, 2023. URL: https://www.cpuc.ca.gov/.

[113] Congress passes bill to improve broadband mapping data, 2023. URL: https://www.electric.coop/congress-passes-bill-to-improve-broadband-mapping-data.

[114] Consolidated communications, 2023. URL: https://www.consolidated.com/.

[115] Digital divide, 2023. URL: https://en.wikipedia.org/wiki/Digital_divide.

[116] Fixed broadband deployment data from fcc form 477, 2023. URL: https://www.fcc.gov/general/broadband-deployment-data-fcc-form-477.

[117] Kolmogorov–smirnov test, 2023. URL: https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test.

[118] National Address Database, 2023. URL: https://www.transportation.gov/gis/national-address-database.

[119] OaklandUndivided, 2023. URL: https://www.oaklandundivided.org/.

[120] Sample size, 2023. URL: https://www.census.gov/acs/www/methodology/sample-size-and-data-quality/sample-size/index.php.

[121] Senators fear 'deeply flawed' fcc broadband map could screw them out of millions in federal funds, 2023. URL: https://gizmodo.com/senators-fcc-broadband-map-deeply-flawed-federal-fund-1849975157.

[122] Utah telecommunication open infrastructure agency, 2023. URL: https://en.wikipedia.org/wiki/Utah_Telecommunication_Open_Infrastructure_Agency.

[123] Wow!, 2023. URL: https://www.wowway.com/.

[124] Zillow, 2023. URL: https://www.zillow.com/.

[125] Firoj Alam, Ferda Ofli, Muhammad Imran, and Michaël Aupetit. A Twitter Tale of Three Hurricanes: Harvey, Irma, and Maria. *CoRR*, abs/1805.05144, 2018.

[126] Moumita Basu, Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. Automatic matching of resource needs and availabilities in microblogs for post-disaster relief. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, 2018.

[127] Steven Bauer, David D. Clark, and William Lehr. Understanding broadband speed measurements. *Annual Research Conference on Communications, Information and Internet Policy*, aug 2010.

[128] Konstantin Bauman, Alexander Tuzhilin, and Ryan Zaczynski. Virtual power outage detection using social sensors. In *NYU Working Paper*, Sept 2015.

[129] Konstantin Bauman, Alexander Tuzhilin, and Ryan Zaczynski. Using social sensors for detecting emergency events: A case of power outages in the electrical utility industry. *ACM Trans. Manage. Inf. Syst.*, 8(2–3), 2017.

[130] Sanjit Biswas, John Bicket, Edmund Wong, Raluca Musaloiu-E, Apurv Bhartia, and Dan Aguayo. Large-scale measurements of wireless network behavior. *SIGCOMM Comput. Commun. Rev.*, 45(4), aug 2015.

[131] Anthony Brindisi. Report on the state of broadband access in new york's 22nd congressional district, 2020.

[132] Francesco Bronzino, Nick Feamster, Shinan Liu, James Saxon, and Paul Schmitt. Mapping the digital divide: Before, during, and after covid-19. In *Conference on Communications, Information, and Internet Policy (TPRC)*, pages 1–11, 2021.

[133] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011. `arXiv:https://doi.org/10.1177/1745691610393980`.

[134] Igor Canadi, Paul Barford, and Joel Sommers. Revisiting broadband performance. In *Proceedings of the 2012 Internet Measurement Conference*, IMC '12, 2012.

[135] Germán Capdehourat, Federico Larroca, and Gastón Morales. A nation-wide wi-fi rssi dataset: Statistical analysis and resulting insights. In *Proceedings of the 2020 IFIP Networking Conference*, IFIP '20, 2020.

[136] ACS Census. American Community Survey Variance Replicate Estimate Tables. URL: `https://www.census.gov/programs-surveys/acs/data/variance-tables.html`.

[137] ACS Census. Calculating Measures of Error for Derived Estimates. URL: `https://www.census.gov/content/dam/Census/library/publications/2018/acs/acs_general_handbook_2018_ch08.pdf`.

[138] Reddick CG, Enriquez R, Harris RJ, and Sharma Bl. Determinants of broadband access and affordability: An analysis of a community survey on the digital divide. In *PubMed*, 2020.

[139] Nitesh V. Chawla, Bowyer Kevin W., Hall Lawrence O., and Kegelmeyer W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi:https://doi.org/10.1613/jair.953.

[140] David D. Clark and Sara Wedeman. Measurement, meaning and purpose: Exploring the m-lab ndt dataset. *Annual Research Conference on Communications, Information and Internet Policy*, aug 2021.

[141] Federal Communications Commission. Measuring Broadband America Program. URL: https://www.fcc.gov/general/measuring-broadband-america.

[142] Chicago Connected. Providing Fast and Free High-Speed Internet for Students who Need it the Most. URL: https://www.cps.edu/strategic-initiatives/chicago-connected/.

[143] Cooperative Network Services. Rdof and flawed 477 reporting, 2023. URL: https://www.cooperative-networks.com/rdof-477-reporting/.

[144] Diego da Hora, Karel van Doorselaer, Koen van Oost, and Renata Teixeira. Predicting the effect of home wi-fi quality on qoe. In *Proceedings of the IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*, INFOCOM '18, 2018.

[145] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL: http://arxiv.org/abs/1810.04805, arXiv:1810.04805.

[146] Marcel Dischinger, Andreas Haeberlen, Krishna P. Gummadi, and Stefan Saroiu. Characterizing residential broadband networks. IMC '07, page 43–56, New York, NY, USA, 2007. Association for Computing Machinery.

[147] Ramakrishnan Durairajan and Paul Barford. A techno-economic approach for broadband deployment in underserved areas. *Computer Communication Review*, 47:13–18, 04 2017.

[148] Luigi Ermini and David F Hendry. Log income vs. linear income: An application of the encompassing principle. *Oxford Bulletin of Economics and Statistics*, 70:807–827, 2008.

[149] FCC. Broadband Adoption and Use in America. URL: https://transition.fcc.gov/national-broadband-plan/broadband-adoption-in-america-paper.pdf.

[150] FCC. Form 477 Instructions. URL: https://us-fcc.app.box.com/v/Form477Instructions.

[151] FCC. Proposed new standards for Broadband speeds. URL: https://www.fcc.gov/document/chairwoman-rosenworcel-proposes-increase-minimum-broadband-speedsp.

[152] FCC. In *Communications Status Report for Areas Impacted by Hurricane Maria*. FCC, 2017. URL: https://www.fcc.gov/document/hurricane-maria-communications-status-report-oct-2.

[153] Nick Feamster and Jason Livingood. Measuring internet speed: Current challenges and future recommendations. *Commun. ACM*, 63(12):72–80, nov 2020.

[154] Federal Communication Commission. Affordable connectivity program, 2023. URL: https://www.fcc.gov/acp.

[155] Federal Communication Commission. Measuring broadband america, 2023. URL: https://www.fcc.gov/general/measuring-broadband-america.

[156] Federal Communication Commission. Urban rate survey data & resources, 2023. URL: https://www.fcc.gov/economics-analytics/industry-analysis-division/urban-rate-survey-data-resources.

[157] George S. Ford. Form 477, speed-tests, and the american broadband user's experience. *SSRN*, mar 2021.

[158] Hernan Galperin, Thai V. Le, and Kurt Wyatt. Who gets access to fast broadband? evidence from los angeles county. *Government Information Quarterly*, 38(3):101594, 2021.

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

[160] Devanandham Henry and Jose Emmanuel Ramirez-Marquez. On the impacts of power outages during hurricane sandy—a resilience-based analysis. *Systems Engineering*, 19(1):59–75, 2016.

[161] Martin Hilbert. 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):567–581, 2016.

[162] Ahmad Hany Hossny and Lewis Mitchell. Event detection in twitter: A keyword volume approach. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2018.

[163] Amanda Lee Hughes and Leysia Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6:248–260, February 2009.

[164] Carolynne Hultquist, Mark Simpson, Guido Cervone, and Qunying Huang. Using nightlight remote sensing imagery and twitter data to study power outages. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on the Use of GIS in Emergency Management*, 2015.

[165] Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM Comput. Surv.*, 47(4), 2015.

[166] Muhammad Imran, Carlos Castillo, Ji Lucas, Patrick Meier, and Sarah Vieweg. AIDR: Artificial Intelligence for Disaster Response. WWW '14 Companion, 2014.

[167] Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a Lifeline: Human-annotated Twitter Corpora for NLP of Crisis-related Messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016.

[168] Hamed Jelodar, Yongli Wang, Chi Yuan Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent Dirichlet Allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

[169] Sifat Shahriar Khan and Jin Wei. Real-time power outage detection system using social sensing and neural networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 927–931, 2018.

[170] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014.

[171] Yury Kryvasheyeu, Haohui Chen, Nick Obradovich, Esteban Moro, Pascal Van Hentenryck, James Fowler, and Manuel Cebrian. Rapid assessment of disaster damage using social media activity. *Science Advances*, 2(3), 2016.

[172] Broadband internet isn't equally available to L.A. County's low-income residents, report says, 2022. URL: `https://www.latimes.com/business/story/2022-10-13/broadband-internet-not-equally-available-to-la\-county-low-income-residents-report-says`.

[173] Weichao Li, Daoyuan Wu, Rocky K.C. Chang, and Ricky K.P. Mok. Demystifying and puncturing the inflated delay in smartphone-based wifi network measurement. In *Proceedings of the 12th International on Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '16, 2016.

[174] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[175] Yu-Hsin Liu, Jeffrey Prince, and Scott Wallsten. Distinguishing bandwidth and latency in households' willingness-to-pay for broadband internet speed. *Information Economics and Policy*, 45:1–15, 2018.

[176] Ossola Alessandro Locke, Dexter Henry, Emily Minor, and Brenda B. Lin. Spatial contagion structures urban vegetation from parcel to landscape. *People and Nature.*, 2022.

[177] Andra Lutu, Diego Perino, Marcelo Bagnulo, Enrique Frias-Martinez, and Javad Khangosstar. A characterization of the covid-19 pandemic impact on a mobile network operator traffic, 2020. `arXiv:2010.02781`.

[178] Kyle MacMillan, Tarun Mangla, James Saxon, Nicole P Marwell, and Nick Feamster. A comparative analysis of ookla speedtest and measurement labs network diagnostic test (ndt7). *arXiv preprint arXiv:2205.12376*, 2022.

[179] David Major, Ross Teixeira, and Jonathan Mayer. 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)*, page 393–419, 2020.

[180] Tarun Mangla, Udit Paul, Arpit Gupta, Nicole P. Marwell, and Nick Feamster. Internet inequity in chicago: Adoption, affordability, and availability. *The Research Conference on Communications, Information and Internet Policy*, 2022.

[181] Tarun Mangla, Esther Showalter, Vivek Adarsh, Kipp Jones, Morgan Vigil-Hayes, Elizabeth Belding, and Ellen Zegura. A tale of three datasets: Characterizing mobile broadband access in the u.s. *Commun. ACM*, 65(3):67–74, 2022.

[182] Tarun Mangla, Esther Showalter, Vivek Adarsh, Kipp Jones, Morgan Vigil-Hayes, Elizabeth Belding, and Ellen Zegura. A tale of three datasets: characterizing mobile broadband access in the us. *Communications of the ACM*, 65(3):67–74, 2022.

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

[184] Matt Mathis and Mark Allman. A Framework for Defining Empirical Bulk Transfer Capacity Metrics. RFC, July 2001. URL: `https://tools.ietf.org/html/rfc3148`.

[185] Tejas N. Narechania. Convergence and a case for broadband rate regulation. *Berkeley Technology Law Journal*, 2021.

[186] National Digital Inclusion Alliance. Definitions – National Digital Inclusion Alliance, 2021. URL: https://www.digitalinclusion.org/definitions/.

[187] National Telecommunications and Information Administration. Broadband equity access and deployment (bead) program, 2022. URL: https://grants.ntia.gov/grantsPortal/s/funding-program/a0g3d00000018ObAAI/broadband-equity-access-and-deployment-bead-program.

[188] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. Applications of online deep learning for crisis response using social media information. *CoRR*, abs/1610.01030, 2016.

[189] NTIA. Broadband Equity Access and Deployment (BEAD) Program. URL: https://grants.ntia.gov/grantsPortal/s/funding-program/a0g3d00000018ObAAI/broadband-equity-access-and-deployment-bead-program.

[190] University of Chicago. The "Sides" of Chicago. URL: https://chicagostudies.uchicago.edu/sides.

[191] Ookla. Speedtest by ookla - the global broadband speed test. URL: https://www.speedtest.net/.

[192] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135, 2008.

[193] Udit Paul, Alexander Ermakov, Michael Nekrasov, Vivek Adarsh, and Elizabeth Belding. outage: Detecting power and communication outages from social networks. WWW '20, 2020.

[194] Udit Paul, Vinothini Gunasekaran, Jiamo Liu, Tejas N. Narechania, Arpit Gupta, and Elizabeth Belding. Decoding the divide: Analyzing disparities in broadband plans offered by major us isps. SIGCOMM '23, 2023.

[195] Udit Paul, Jiamo Liu, Vivek Adarsh, Mengyang Gu, Arpit Gupta, and Elizabeth Belding. Characterizing performance inequity across u.s. ookla speedtest users, 2021. URL: https://arxiv.org/abs/2110.12038.

[196] Udit Paul, Jiamo Liu, David Farias-llerenas, Vivek Adarsh, Arpit Gupta, and Elizabeth Belding. Characterizing internet access and quality inequities in california m-lab measurements. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, 2022.

[197] Udit Paul, Jiamo Liu, Mengyang Gu, Arpit Gupta, and Elizabeth Belding. The importance of contextualization of crowdsourced active speed test measurements. IMC '22, 2022.

[198] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

[199] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, October 2014. Association for Computational Linguistics. URL: `https://www.aclweb.org/anthology/D14-1162`, `doi:10.3115/v1/D14-1162`.

[200] Linda Poon. To Bridge the Digital Divide — and Close the Homework Gap — Cities Are Tapping Their Own Infrastructure. URL: `https://www.the74million.org/article/to-bridge-the-digital-divide-and-close-the-homework-gap-cities-are-tapping-their-own-infrastructure/`.

[201] James E. Prieger. 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, January 2003. URL: `https://ideas.repec.org/p/cda/wpaper/50.html`.

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

[203] Ruoxi Qin, Kai Qiao, Linyuan Wang, Lei Zeng, Jian Chen, and Bin Yan. Weighted focal loss: An effective loss function to overcome unbalance problem of chest X-ray. *IOP Conference Series: Materials Science and Engineering*, 428:012022, 2018.

[204] Christian Reuter, Amanda Lee Hughes, and Marc-André Kaufhold. Social media in crisis management: An evaluation and analysis of crisis informatics research. *International Journal of Human–Computer Interaction*, 34(4):280–294, 2018.

[205] Colin Rhinesmith, Bianca Reisdorf, and Madison Bishop. The ability to pay for broadband. *Communication Research and Practice*, 5(2):121–138, 2019.

[206] Koustav Rudra, Pawan Goyal, Niloy Ganguly, Prasenjit Mitra, and Muhammad Imran. Identifying sub-events and summarizing disaster-related information from microblogs. In *The 41st International ACM SIGIR Conference on Research; Development in Information Retrieval*, SIGIR '18, 2018.

[207] Koustav Rudra, Ashish Sharma, Niloy Ganguly, and Muhammad Imran. Classifying information from microblogs during epidemics. In *Proceedings of the 2017 International Conference on Digital Health*, 2017.

[208] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Tweet analysis for real-time event detection and earthquake reporting system development. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):919–931, 2013.

[209] Haji Mohammad Saleem, Faiyaz Al Zamal, and Derek Ruths. Tackling the challenges of situational awareness extraction in twitter with an adaptive approach. *Procedia Engineering*, 107:301–311, 2015.

[210] Robert J Sampson. Great american city. In *Great American City*. University of Chicago Press, 2021.

[211] Eric Scheier and Noah Kittner. A measurement strategy to address disparities across household energy burdens. *Nat Commun 13*, 288, 2022.

[212] Axel Schulz, Eneldo Loza Mencía, Thanh Dang, and Benedikt Schmidt. Evaluating multi-label classification of incident-related tweets. In *Proceedings, 4th Workshop on Making Sense of Microposts (#Microposts2014) at WWW: Big things come in small packages*, April 2014.

[213] Ranya Sharma, Tarun Mangla, James Saxon, Marc Richardson, Nick Feamster, and Nicole P. Marwell. Benchmarks or Equity? A New Approach to Measuring Internet Performance, 2022. URL: `https://ssrn.com/abstract=4179787`.

[214] Joel Sommers and Paul Barford. Cell vs. wifi: On the performance of metro area mobile connections. In *Proceedings of the 2012 Internet Measurement Conference*, IMC '12, 2012.

[215] Geoffrey Starks. Availability, adoption, and access: The three pillars of broadband equity, 2022. URL: `http://soba.iamempowered.com/availability-adoption-and-access-three-pillars-broadband-equity`.

[216] Kevin Stowe, Michael J. Paul, Martha Palmer, Leysia Palen, and Kenneth Anderson. Identifying and Categorizing Disaster-Related Tweets. In *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, 2016.

[217] Kaixin Sui, Mengyu Zhou, Dapeng Liu, Minghua Ma, Dan Pei, Youjian Zhao, Zimu Li, and Thomas Moscibroda. Characterizing and improving wifi latency in large-scale operational networks. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, 2016.

[218] Haifeng Sun, Zhaoyu Wang, Jianhui Wang, Zhen Huang, NichelleLe Carrington, and Jianxin Liao. Data-driven power outage detection by social sensors. *IEEE Transactions on Smart Grid*, 7(5):2516–2524, 2016.

[219] Srikanth Sundaresan, Xiaohong Deng, Yun Feng, Danny Lee, and Amogh Dhamdhere. Challenges in inferring internet congestion using throughput measurements. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, 2017.

[220] Srikanth Sundaresan, Nick Feamster, and Renata Teixeira. Home Network or Access Link? Locating Last-Mile Downstream Throughput Bottlenecks. In *Proceedings of the 2016 Passive and Active Measurement Conference*, PAM '16, 2016.

[221] Hien To, Sumeet Agrawal, Seon Ho Kim, and Cyrus Shahabi. On identifying disaster-related tweets: Matching-based or learning-based? In *2017 IEEE Third International Conference on Multimedia Big Data (BigMM)*, 2017.

[222] Brian Whitacre, Roberto Gallardo, and Sharon Strover. Does rural broadband impact jobs and income? evidence from spatial and first-differenced regressions. *The Annals of Regional Science*, 53:649–670, November 2014.

[223] Xinlei Yang, Xianlong Wang, Zhenhua Li, Yunhao Liu, Feng Qian, Liangyi Gong, Rui Miao, and Tianyin Xu. Fast and light bandwidth testing for internet users. In *Proceedings of the 18th USENIX Symposium on Networked Systems Design and Implementation*, NSDI '21, 2021.

[224] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016.

[225] Himanshu Zade, Kushal Shah, Vaibhavi Rangarajan, Priyanka Kshirsagar, Muhammad Imran, and Kate Starbird. From situational awareness to actionability: Towards improving the utility of social media data for crisis response. *Proc. ACM Hum.-Comput. Interact.*, 2:1–18, 2018.

[226] Bell Nathaniel Zahnd, Whitney E. and Annie E. Larson. Geographic, racial/ethnic, and socioeconomic inequities in broadband access. *The Journal of Rural Health*, 2021.