DOI:10.1145/3462672

Needed improvements to mobile broadband deployment require more accurate mapping of mobile coverage, especially in rural and tribal areas.

BY 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.

AFFORDABLE, QUALITY INTERNET access is critical for full participation in the 21st century economy, educational system, and government.[23] Mobile broadband can be achieved through commercial Long-Term Evolution (LTE) cellular networks, which are a proven means of expanding access[11] but are often concentrated in urban areas—leaving economically marginalized and sparsely populated areas underserved.[1] The U.S. Federal Communications Commission (FCC) incentivizes LTE operators that serve rural areas[3,22] and maintains transparency by releasing maps from each operator showing geographic areas of coverage.[9] Recently, third parties have challenged the veracity of these maps, claiming they over-represent true coverage and can discourage much-needed investment.

However, most of these claims are either mainly qualitative in nature or are focused on limited areas, where a few dedicated researchers can collect controlled coverage measurements (through wardriving, for instance).[12,24,25] As dependence on mobile broadband connectivity increases, especially in the face of the COVID-19 pandemic, mechanisms that quantitatively validate FCC coverage datasets at scale are becom-

» key insights

- **We compare LTE coverage data from the FCC with a crowdsourced dataset from Skyhook for New Mexico. While the two coverage datasets tend to agree in urban areas, there is significant disparity, up to 15%, in rural and tribal census blocks.**
- **On-ground LTE coverage measurements collected across 120 miles of rural and tribal New Mexico indicate that even the crowdsourced data exhibits over-reporting, although to a lesser degree than the FCC data.**
- **The findings make a case for including mechanisms to validate ISP-reported FCC coverage data. While crowdsourcing is a good alternative, targeted active measurement campaigns are needed in areas where existing crowdsource datasets are sparse.**

ing acutely necessary to evaluate and direct resources toward Internet access deployment efforts.[15,18] This is a technology policy issue which carries equity and fairness implications for society as a whole.

An increasingly widespread approach to measuring coverage at scale is through crowdsourcing, wherein LTE network users contribute to coverage measurements. The FCC has recently advocated for the use of crowdsourcing to validate coverage data reported by operators.[6] In this context, our work employs a data-driven, empirical approach, comparing coverage from a representative crowdsourced dataset with the FCC data. More specifically, our analysis is guided by the following questions:

- How consistent are existing LTE coverage datasets?
- Where and how do their coverage estimations differ?

We specifically consider a crowdsourced coverage estimate from Skyhook, a commercial location service provider that uses a variety of positioning tools to offer precise geolocation. We selected Skyhook because it crowdsources cellular coverage measurements from end-user applications that subscribe to its location services. Such incidental crowdsourcing can potentially provide richer coverage data compared to a voluntary form of crowdsourcing, where users must explicitly commit to contributing coverage data. Our research examines this by comparing the Skyhook measurements with those of OpenCellID, an open but voluntary crowdsourced dataset.[21] As our findings show, the density of the crowdsourced datasets varies significantly by the methodology of data collection, especially in rural areas. In the regions we studied, incidental crowdsourcing (Skyhook) gathered up to 11.1x more cell IDs than voluntary crowdsourcing (OpenCellID).

Using Skyhook as an extensive crowdsourced dataset, we can quantify how widely and where the crowdsourced coverage data differs from the FCC data. We specifically selected the state of New Mexico[a] for its mix of demographics, diverse geographic landscape, and our partnership with community stakeholders within the state. In our research, we compare coverage at the level of census blocks,[b] which are further grouped into urban, rural, and tribal[c] categories. We found that the FCC and Skyhook LTE datasets have a discrepancy as great as 15% in rural census blocks, with the FCC data claiming higher coverage than Skyhook. A major concern in interpreting this comparison is accounting for coverage discrepancies due to a lack of data points in the crowdsourced dataset. To confirm the availability of users to provide data points, we checked for the presence of alternate cellular technologies—for example, 2G or 3G—within these census blocks and observed a significant number (up to 9% in tribal rural areas) where such alternates are present, evidence that users do visit those blocks but cannot access LTE. These results, like a recent study on fixed broadband,[16] suggest a need to incorporate mechanisms to validate operator-submitted data into the FCC's LTE access-measurement methodology, especially in rural and tribal areas.

Finally, this article compares both FCC and Skyhook coverage maps to our own controlled coverage measurements collected from a northern section of New Mexico. Interestingly, both FCC and Skyhook datasets report higher coverage relative to our controlled measurements, with the former showing a higher degree (by up to 26.7%) of over-reporting than the latter. Understanding the causes of these inconsistencies is important for effectively using crowdsourced data to measure LTE coverage, especially as crowdsourcing is increasingly viewed as preferable to provider reports. We conclude with recommendations for improving LTE coverage measurements, whose importance has only increased in the COVID-19 era of remote working and learning.

a Our methodology is not specific to New Mexico and can be easily extended to other regions in the U.S.

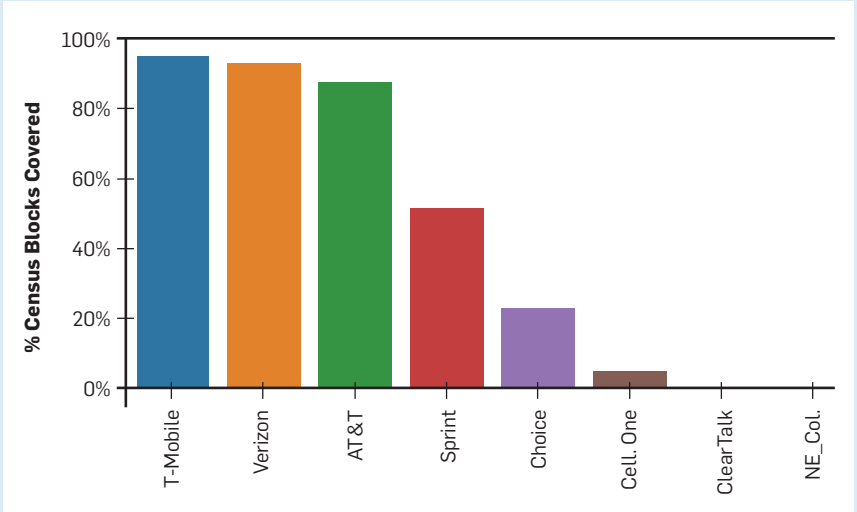
b We use the FCC methodology, wherein a census block is considered covered if the centroid is covered.[7]

c Tribal areas have consistently experienced the lowest broadband coverage rates in the U.S. for the past decade.[1]

Table 1. Summary of coverage datasets.

Dataset	Points of Collection	Format	Methodology
FCC	Polygon overlay	Shapefile	Operator-reported with Form 477
Skyhook	Cell signal point	CSV	Incidental crowdsourcing
Author-controlled measurements	Cell signal point	CSV	Wardriving

Figure 1. LTE operators by census-block coverage based on FCC data.



Background and Datasets

This section offers an overview of the LTE network architecture, followed by a description of the LTE coverage datasets compared in our analysis. These datasets are summarized in Table 1. Limitations associated with each data collection methodology are also noted.

LTE network architecture. Internet access in an LTE network is available through base stations, known as eNodeBs, which the network provider operates. User equipment (UE), such as smartphones, tablets, or LTE modems, connects to the eNodeB over the radio link. The eNodeB connects to a centralized cellular core, known as the evolved packet core (EPC), typically via a wired link forming a middle-mile connection. The EPC consists of several network elements, including a packet data network gateway (PGW), which is the connecting node between an end-user device and the public Internet. Thus, LTE broadband access depends on multiple factors, including radio coverage, middle-mile capacity, and interconnection links with other networks—transit providers and content providers, for instance—in the public Internet. However, the focus of this article is to understand last-mile LTE connectivity characterized by the radio coverage of the eNodeB.

An eNodeB controls a single cell site and consists of several radio transceivers or cells mounted on a raised structure, such as a mast or a tower. The radio cells use directional antennas, with each antenna providing coverage in a smaller geographical area using one frequency band. The radio cells can be identified through a globally unique number called a cell identifier (or cell ID), which is also visible to an end-user device in range of the cell. The cell ID enables aggregation of connectivity and signal-strength information from multiple UEs connected to the same cell, which can then be used to estimate the geolocation of a cell along with its coverage.

FCC dataset. The FCC LTE broadband dataset consists of coverage maps in shapefile format that depict geospatial LTE network deployment for each cellular operator in the U.S. The FCC uses Form 477 to compile this dataset semi-annually from operators, and every operator that owns cellular network facilities must participate in this data collection. Operators submit shapefiles containing detailed network information in the form of geo-polygons along with the frequency band used in the polygon and the minimum advertised upload and download speeds. The methodology used for obtaining these polygons is proprietary to each operator. Ultimately, the FCC publishes only a coverage map that represents coverage as a binary indicator: in any location, cellular service is either available through an operator or it is not.

Our research uses binary coverage shapefiles, available on the FCC's website, from June 2019.[d] Figure 1 shows New Mexico's eight LTE network operators and the percentage of the state's total census blocks covered by each operator. Note: we use one of the FCC methodologies to report mobile broadband access, wherein a census block is considered covered if the centroid of the census block is covered.[7] In this article, our analysis is limited to the top four cellular operators due to their significantly greater prevalence in New Mexico; these operators are also the top four cellular operators in the U.S. more broadly.

Limitations. These coverage maps are generated using predictive models that are proprietary to the operator[2] and not generally reproducible. Furthermore, the publicly available dataset consists of binary coverage and lacks any performance-related data.[e]

Skyhook dataset. Skyhook is a location service provider that uses a variety of positioning tools, including a database of cell locations, to offer precise geolocation to subscribed applications.[4] Through apps that subscribe to Skyhook's location services, user devices report back network information, which is gathered into anonymous logs and used to improve the localization service. Through a data-access agreement, we were able to view the cell-location database, which consisted of location and coverage estimates as well as a list of unique cell IDs along with the cell technology—for example, 3G vs. LTE. The database was originally constructed through extensive wardriving but is now managed and updated through measurements gathered by devices using the Skyhook API for localization. Device measurements with the same cell ID are combined to estimate cell location and coverage in the following manner:

Cell location estimation. A grid-based methodology similar to that proposed by Nurmi et al.[20] is used to estimate the cell tower location. Specifically, Skyhook divides the geographic area into 7-m squares and groups measurements in the same square to obtain a central measure of the square's signal strength. This is done to reduce the bias due to large numbers of measurements coming from the same area—for instance, a popular gathering place. A weighted average of the signal strength is then used to estimate the cell location.

Estimation of cell coverage radius. Skyhook also provides an estimate of the cell's coverage radius using a proprietary method based on path-loss gradient.[26] Path-loss gradient approximates how the wireless signal attenuates as a function of the distance from the transmitter—a radio cell, in this case. The value of the path-loss gradient depends on several factors, including environment (foliage, buildings), geographic topography, and cell signal frequency. Skyhook estimates path-loss gradient using field observations of cell signal strength readings along with their distributed geographic locations. Ideally, the signal attenuation varies based on the direction and distance from the cell. However, to reduce the complexity of coverage estimation, Skyhook's cell coverage estimation heuristic calculates only one path-loss gradient for a single cell. Path-loss gradient is then used in a set of parameterized equations to estimate the cell coverage radius. The parameters in these equations have been determined with careful research and testing over more than 10 years.

The cell-location database is regularly updated with cell-location recalculation and cell-coverage radius using the new device measurements

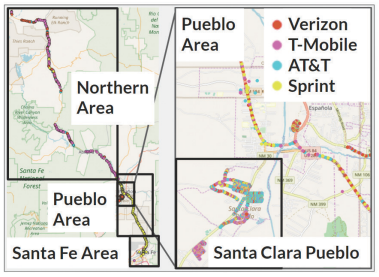
d At the time of this analysis, data from December 2019 was also available on the FCC website. However, we use data from June 2019, as the other two datasets in our analysis are collected around this period.

e The FCC has only recently (December 2019) begun providing speed data along with coverage information.

Table 2. Characteristics and cell ID (CID) counts in selected counties.

County Classification	Region	County Name	Population Density (per square mile)	Skyhook		OpenCellID		Common CIDs
				CIDs (#)	% Overlap	CIDs (#)	% Overlap	
Large Metro	Western	Los Angeles, CA	2,490.3	133,484	28%	39,875	92%	36,816
	Central	Denver, CO	4,683.0	11,061	24%	3,136	86%	2,689
	Eastern	Fulton, GA	1,994.0	27,809	22%	7,225	86%	6,194
Small Metro	Western	Imperial, CA	43.5	1,818	17%	336	93%	311
	Central	Doña Ana, NM	57.1	1,870	32%	663	89%	592
	Eastern	Bibb, GA	613.0	1,953	21%	464	89%	413
Micropolitan	Western	Tehama, CA	21.7	733	17%	158	80%	126
	Central	Rio Arriba, NM	6.7	333	8%	30	87%	26
	Eastern	Pierce, GA	61.3	164	9%	21	67%	14

Figure 2. Map of author wardriving areas in New Mexico.



collected since the last update. For our analysis, we used the cell-location database last updated on June 10, 2019.

Limitations. Since database entries are crowdsourced when the device passes within range of a cell, this dataset is more comprehensive in population centers and highways, where people more often occupy. If there are too few measurements overall, or if measurements are primarily sourced from the same grid section, then the cell-location estimate can be inaccurate.

Targeted measurement campaign. To complement these datasets, we performed a targeted measurement campaign collecting coverage information across 120 miles of Rio Arriba County in New Mexico over a five-day period, beginning May 28, 2019. Figure 2 shows the locations of ground measurements, and the four descriptive area labels we use for this analysis. North area measurements were taken on highways passing primarily through national forest while pueblo area measurements were taken from highways within tribal jurisdiction boundaries. In Santa Clara Pueblo, tribal leadership permitted us to collect additional measurements in residential zones. Finally, the Santa Fe area consists of highway measurements between the pueblos and downtown Santa Fe. While limited in scale, these active measurements provide an important comparison point for coverage and user experience. As already described, we selected these areas of New Mexico for their mix of tribal and non-tribal demographics; tribal lands tend to have the highest coverage overstatements and the most limited cellular availability within the U.S.

Our measurements consist of service-state and signal-strength readings recorded on four Motorola G7 Power (XT1955-5) phones running Android Pie (9.0.0). Service state is a discrete variable indicating whether the phone is connected to a cell. Measurements were collected using the Network Monitor application.[14] An external GlobalSat BU-353-S4 GPS connected to an Ubuntu Lenovo ThinkPad laptop gathered geolocation tags that were matched to network measurements by timestamp. Each phone was outfitted with a SIM card from one of the top four cellular operators in the area: Verizon, T-Mobile, AT&T, and Sprint. The phones recorded service state and signal strength every 10 seconds while we drove at highway speeds (between 40 and 65 mph) in most places and less than 10 mph in residential areas (Santa Clara Pueblo).

Limitations. Our wardriving campaign was intensive in terms of human effort, economic cost, and time, making it difficult to scale. The dataset does not capture any temporal variations in coverage, as the measurements were collected over a short time span. It is possible that driving speed or device configuration impacted the measurements—for example, indicating no coverage when a stationary measurement might have detected coverage.[8] We have no evidence that this occurred, but it might warrant additional investigation.

Analysis

In this section, we evaluate Skyhook as a representative crowdsourced dataset by comparing it with a popular voluntary crowdsourced data from OpenCellID.[21] This is followed by a comparison of coverage across the FCC, Skyhook, and our own wardriving measurement data. Our comparison is guided by the following questions: What is the degree of coverage agreement across the datasets? Where and how do their coverage estimations differ?

Comparison of crowdsourced datasets. We compare the Skyhook dataset with a publicly available crowdsourced dataset—Unwired Lab's OpenCellID.[f] The OpenCellID project provides a publicly available dataset of cell IDs along with their estimated location. The dataset is derived from crowdsourced UE signal-strength measurements similar to Skyhook. However, the UE measurements in this case come from users who voluntarily install the OpenCellID application on their smartphone[21] and manually choose what data to upload. We differentiate this voluntary crowdsourcing method of data collection from Skyhook's incidental crowdsourcing method, where users of the Skyhook API contribute to the data by default. We specifically compare the number of unique LTE cells and the recency of the measurements in both datasets. We

f The OpenCellID Project is licensed under a Creative Commons Attribution ShareAlike 4.0 International License.

consider each of these factors to contribute to the dataset's overall density.

Methodology. While our coverage comparison focuses on New Mexico, we analyzed our selected crowdsourced data more broadly by considering these datasets within a set of counties selected from three areas of the U.S.: western (California), central (New Mexico and Colorado), and eastern (Georgia), each representing varying population densities across the country. Within each region, we considered three different kinds of counties as defined by the National Center for Health Statistics' 2013 Urban-Rural Classification Guide:[19] *large metropolitan* (large), which has a population of at least one million and a principal city; *small metropolitan* (small), which has a population of less than 250,000; and *micropolitan* (micro), which has at least one urban cluster of at least 10,000 but a total population of less than 50,000.

This enabled us to study differences based on population density and geographic region for the crowdsourced datasets. To compare these two datasets, we selected three counties of each population category, for a total of nine counties, which are described in Table 2. For each county, we show the 2018 population density estimated from the U.S. Census Bureau's 2010 census records.[5] We first count the number of unique cell IDs that appear in both datasets for each county. Table 2's "% Overlap" columns show the percentage of each dataset's cell IDs that also appear in the other dataset, and the "Common CIDs" column shows the exact number of common cell IDs.

Results. Overall, Skyhook reports a greater number of cells (from 2.8x to 11.1x more) for all counties. The difference is particularly pronounced in micro counties, which suggests that relying on volunteers to download an application and offer network measurements may not be the most accurate method for assessing LTE coverage in rural areas. Furthermore, Skyhook includes most of the cells that appear in OpenCellID.

We next considered how recently each cell ID record was updated with a new measurement. Figure 3 shows the CDF of the latest measurement date for cells in both datasets, where cells are split into those located in urban and rural census blocks. Almost 60% of the cells in Skyhook were last updated in June 2019, but the most recent update in OpenCellID was in February 2019. Furthermore, cells in rural census blocks were updated less recently than in OpenCellID's urban census blocks, while the difference is negligible in the Skyhook dataset. This suggests that the Skyhook dataset is updated more regularly than OpenCellID, thus making it more likely to

Figure 3. CDF of cell updates in Skyhook (S) and OpenCellID (O).

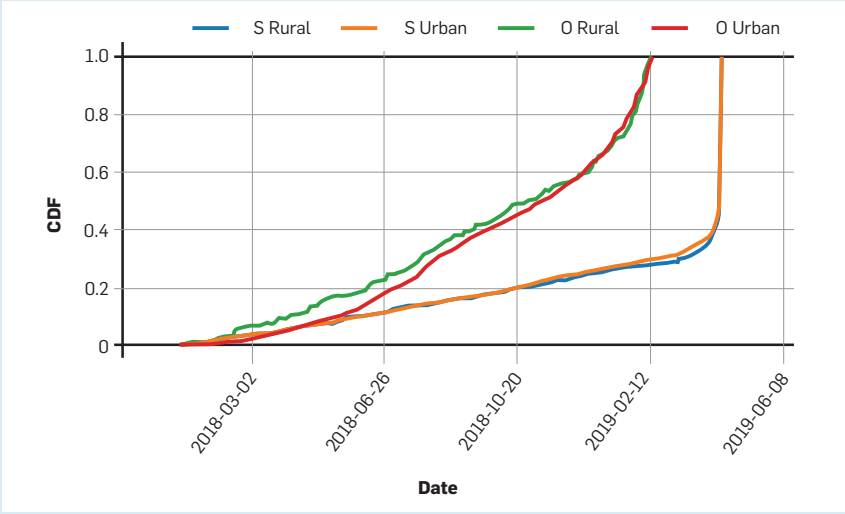
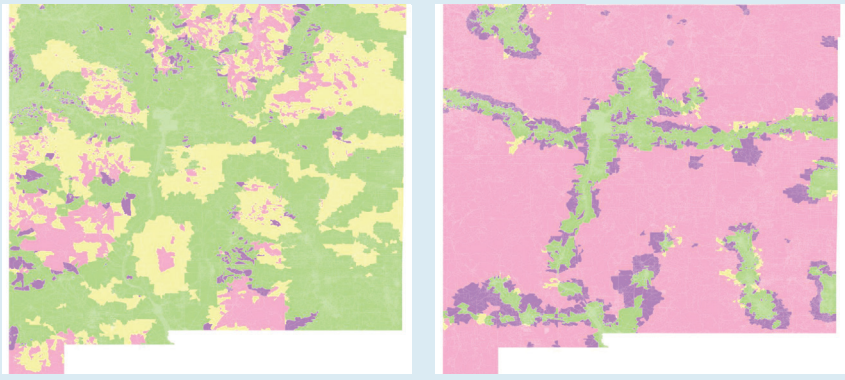


Table 3. Percentage of total census blocks covered, according to FCC and Skyhook.

Census Block Type	Total Census Blocks	Verizon		T-Mobile		AT&T		Sprint	
		FCC	Skyhook	FCC	Skyhook	FCC	Skyhook	FCC	Skyhook
Non-Tribal Rural	93,680	89%	77%	94%	86%	85%	79%	39%	49%
Non-Tribal Urban	41,872	100%	100%	100%	100%	99%	99%	96%	99%
Tribal Rural	30,588	93%	80%	92%	63%	78%	73%	27%	41%
Tribal Urban	2,469	100%	99%	95%	94%	93%	94%	75%	88%
All	168,609	93%	84%	95%	85%	88%	83%	52%	61%

Figure 4. Comparison of LTE coverage maps of New Mexico. Yellow blocks are covered in the FCC map but not in Skyhook; purple blocks are covered in the Skyhook map but not the FCC map. Green blocks are covered in both, and pink blocks are not covered in either map.



(a) Verizon (b) Sprint

represent any changes in the network infrastructure.

Comparison of coverage. We first compared a coverage shapefile generated from Skyhook cell locations and estimated coverage ranges with the FCC map for each operator.

Methodology. We considered coverage at the census-block level for this comparison. In addition to reporting coverage shapefiles, the FCC reports coverage at a census-block level and considers a census block as covered if the centroid of the census block falls within a covered region.[7] We generated a similar census block-level coverage map per operator using Skyhook's estimated coverage. To do so, we first obtained the coverage shapefile for each operator using a cell's estimated location and coverage radius. Then, we used the FCC centroid methodology to generate the Skyhook LTE coverage map at the census-block level. We used the Python GeoPandas 0.8.2 library for the associated spatial operations.[10] To explore whether the degree of agreement of the two datasets varies across these dimensions, we grouped census blocks into four categories: Non-Tribal Urban, Non-Tribal Rural, Tribal Urban, and Tribal Rural. We referenced the U.S. Census Bureau's classification of urban and rural blocks and its boundary definitions of tribal jurisdiction for this categorization.[27] In this analysis, we considered census blocks as tribal if they overlap with any tribal boundaries. We varied the tribal labeling schemes, such as classifying a census block as tribal if the centroid of the block is within a tribal boundary. However, the results remained qualitatively similar and did not impact the findings presented here.

Results: Table 3 shows the percentage of total census blocks covered by each cellular operator, according to FCC and Skyhook data, broken down by census-block type. Among the four operators, T-Mobile covers the greatest number of census blocks based on both FCC and Skyhook data, while Sprint covers the fewest. All four cellular operators have relatively higher coverage for both tribal and non-tribal urban census blocks. However, all operators except Verizon offer their lowest coverage in tribal rural areas. For some operators, the differences between non-tribal rural and tribal rural are as great as 23% (based on Skyhook data) and 11% (based on FCC data).

The extent of LTE coverage differs between the two datasets. For three out of four providers, Skyhook shows lower coverage than the FCC, particularly in the rural census blocks. For instance, FCC T-Mobile data shows coverage in 92% of tribal rural blocks, whereas Skyhook shows coverage in only 63% of such blocks. For Sprint, on the other hand, Skyhook shows more census blocks covered than the FCC. This could have been due to multiple reasons, including: there are differences in the propagation models used by Skyhook and Sprint to estimate coverage, with the former's models being more generous than those of the latter, or Skyhook data is collected across time, and Sprint may have discontinued or temporarily disabled some of the cells, which is challenging to detect from the crowdsourced data.

Figure 4 visually compares the LTE coverage maps from the FCC and the Skyhook datasets for Verizon and Sprint. We more deeply examined the discrepancy, mapped in yellow in Figure 4a. Table 4 shows the number of census blocks where there is coverage according to the FCC but none according to Skyhook for each operator. Coverage claims in both tribal and non-tribal rural census blocks disagree the most. The number of such blocks are particularly high for Verizon (19, 126 overall) and T-Mobile (18, 189 overall). There are two possible reasons for this disagreement: either network

Table 4. Number of census blocks where there is coverage according to the FCC but no coverage according to Skyhook.

Block Type	Total Blocks	Verizon	T-Mobile	AT&T	Sprint
Non-Tribal Rural	93,680	14,013	9,025	8,705	1,355
Non-Tribal Urban	41,872	0	0	213	25
Tribal Rural	30,588	5,109	9,150	3,004	230
Tribal Urban	2,469	4	14	4	0

Table 5. Number of census blocks with LTE coverage according to the FCC, but only 3G coverage according to Skyhook. The numbers in parentheses report the same data as a percentage of total census blocks of the corresponding type.

Block Type	Verizon	T-Mobile	AT&T	Sprint
Non-Tribal Rural	528 (1%)	2,575 (3%)	5,342 (6%)	19 (<1%)
Non-Tribal Urban	0 (0%)	0 (0%)	213 (1%)	0 (0%)
Tribal Rural	2,655 (9%)	2,565 (8%)	2,166 (7%)	0 (0%)
Tribal Urban	0 (0%)	0 (0%)	4 (<1%)	0 (0%)

Table 6. Confusion matrices compare active measurement coverage with FCC and Skyhook. *Total* denotes the number of active measurements in each category.

		FCC		Skyhook	
Active	Total	NC	C	NC	C
No Coverage (NC)	266	19%	81%	32%	68%
Coverage (C)	1,440	0%	100%	5%	95%

(a) Verizon

		FCC		Skyhook	
Active	Total	NC	C	NC	C
No Coverage (NC)	324	6%	94%	21%	79%
Coverage (C)	1,361	0%	100%	5%	95%

(b) T-Mobile

		FCC		Skyhook	
Active	Total	NC	C	NC	C
No Coverage (NC)	568	25%	75%	53%	48%
Coverage (C)	1,095	2%	98%	7%	93%

(c) AT&T

		FCC		Skyhook	
Active	Total	NC	C	NC	C
No Coverage (NC)	231	96%	4%	99%	2%
Coverage (C)	1,122	21%	79%	20%	80%

(d) Sprint

operators lack adequate infrastructure in rural areas but tend to overestimate coverage while reporting it to the FCC, or Skyhook is missing data points from rural census blocks, where fewer people carry UEs. The latter case will lead to either some LTE cells not being detected or an inaccurate characterization of cell coverage due to fewer measurements.

To understand which of these potential reasons for disagreement is more likely, we checked whether Skyhook shows 3G coverage for these census blocks (where the FCC reports LTE coverage but Skyhook does not). If Skyhook reports 3G coverage in these blocks, this suggests that users may have contributed to the Skyhook dataset in these census blocks; therefore, LTE coverage would have been detected if it existed. Note: A more accurate approach would have been to directly consider the location of end-user measurements connected using 3G technology and analyze whether they fall within LTE coverage areas in the FCC data. However, we did not have access to these end-user measurements due to Skyhook's privacy policy. Instead, we considered the 3G coverage maps as a reasonable approximation for our analysis and generated a 3G coverage map at the census-block level for these areas in the same manner as previously described. The number of census blocks that show only 3G coverage according to Skyhook is presented in Table 5. We observed a significant number of census blocks where Skyhook detects 3G coverage, indicating that the FCC LTE coverage claims may be overstated in these areas. The number of such blocks is greater for tribal rural areas (up to 9%), thus indicating a higher mismatch of the two datasets in tribal rural areas.

Active measurements compared to FCC and Skyhook coverage. In this section, we compared our own active measurements with the coverage maps from the FCC and Skyhook described previously. We focused on the geographic region around Santa Clara Pueblo, which lies north of Santa Fe (see Figure 2), a region with a mix of urban, rural, and tribal population blocks.

Methodology. We used the service-state readings collected in our measurements for this analysis (see section called "Targeted Measurement Campaign"). We also collected information about the connected cell's technology (for example, LTE) and the geolocation of the measurements. This information is used to infer whether LTE coverage exists at a location. We consider LTE to be available if the service state shows IN_SERVICE to indicate an active connection and if the associated cell is an LTE cell. We term this the active LTE coverage. We then compared the FCC and Skyhook coverage with the active LTE coverage to see if the datasets agreed. Note: We used the coverage shapefiles for both Skyhook and the FCC in this comparison instead of the census-block centroid approach. This allowed us to more precisely compare coverage for a location, especially if a census block is only partially covered.

Results. Table 6 shows the confusion matrices that compare active LTE coverage with reported coverage from the FCC and Skyhook maps. Both maps show coverage at locations where our measurements did not. In the case of Verizon, 81% of the measurements with no coverage are from locations reported as covered by the FCC. This over-reporting is lowest for Sprint and highest for T-Mobile.

We also observed significant disagreement (up to 79%) between Skyhook coverage and our measurements. Two possibilities may explain this: paucity in Skyhook UE signal-strength readings available for cell location and coverage radius estimation, or an error in the cell propagation model, itself possibly due to variations in environmental conditions, such as the terrain. In either case, Skyhook is more in line with our measurements than the FCC in reporting areas with no LTE coverage. For example, in the case of AT&T, 75% of our measurements with no LTE coverage belong to areas reported as covered by the FCC, compared to just 48% by Skyhook.

Recommendations

In this section, we discuss some of the implications of our experience collecting and analyzing coverage data and offer recommendations and directions for future work based on our findings.

Recommendations for the FCC. Our findings make a case for including mechanisms that validate ISP-reported coverage data, especially in rural and tribal regions. Given the scale of cellular networks, crowdsourcing coverage measurements are a viable approach to validating access as opposed to controlled measurements. Within crowdsourcing, we suggest leveraging incidental rather than voluntary approaches, possibly working with third-party services that collect network measurements as part of their service process (as in the case of Skyhook).

In addition, crowdsourcing alone may not be sufficient for determining coverage in some cases. Even with the more complete datasets provided through incidental crowdsourcing, rural areas tended to receive significantly fewer measurements per tower. In such cases, mechanisms need to be developed to precisely determine the areas of greatest disagreement using sparse crowdsourced datasets. Resources can then be focused to target data collection in these areas instead of a blanket approach that measures coverage everywhere.

Recommendations for crowdsourced data collection. We find some shortcomings in the existing crowdsourced datasets. First, existing datasets only report areas with positive coverage—that is, areas where coverage is observed. This makes it difficult to distinguish areas that lack coverage from areas for which no measurements were gathered. Recording areas that lack a usable signal can enable stronger conclusions from crowdsourced data.

Second, we note that even crowdsourced datasets are prone to overestimation of coverage, potentially due to cell location and coverage estimation errors. Research efforts that effectively use the knowledge of cellular network design are needed for an accurate characterization of coverage from crowdsourced measurements. For instance, existing cell location estimation techniques localize cells independently and are error-prone when there are few end-user measurements.[13] Instead, one can use the fact that a single physical tower in an LTE network hosts multiple cells. Thus, algorithms that jointly localize cells for which the end-user measurements are in physi-

cal proximity may provide higher accuracy even with fewer end-user measurements. Similarly, alternate data sources can also be considered for localizing cell infrastructure, such as using geo-imagery data to identify physical towers, or directly obtaining infrastructure data from entities that build and manage physical cell towers (usually different from cellular ISPs).

Measuring access beyond binary coverage. While the focus of this work is on understanding coverage, we recognize that a binary notion of coverage alone does not necessarily indicate the existence of usable LTE connectivity. Other factors can impact end-user experience in a "covered" area, such as low signal strength or poor middle-mile connectivity. Thus, future coverage-measurement efforts must augment coverage reports with measurements of performance to provide models that better align with user experiences. Measuring such performance metrics poses a greater challenge because end-user experience depends on myriad factors beyond last-mile link quality. We believe that increasing community awareness is the way to tackle this problem—for example, through workshops in public libraries or community meetings on the importance of measuring mobile coverage.

Finally, we also note that access and adoption are different, and there are issues beyond access that might also warrant measurement and consideration as accountability measures for operators. Our collection of ground truth datasets involved five days driving through Rio Arriba County in northern New Mexico. In preparation for the trip, we worked to obtain SIM cards that would enable us to access the networks of the four major U.S. LTE operators. This was surprisingly difficult; over the course of a month leading up to the measurement campaign, we spent a collective 24 hours in various operator kiosks and stores in three states to obtain eight SIM cards (one for each major operator). At one of the Santa Fe stores, we encountered a woman who had to drive an hour from Las Vegas, NM to address some of the issues she was having with her mobile service operator that were preventing her from using her data plan. While these anecdotal experiences mirror the qualitative claims of coverage overestimation, they do introduce a new set of issues that must be considered to effectively reduce the barriers of Internet access for rural communities.

Conclusion

In this article, we quantitatively examined the LTE coverage discrepancy among existing datasets collected using different methodologies. We found that existing datasets display the most divergence when compared with each other in rural and tribal areas. We discussed our findings with respect to their implications for telecommunications policy. We also identified several future research directions for the computing community, including mechanisms to augment existing datasets to precisely determine areas in need of more concerted measurement efforts; improved coverage-estimation models, especially for areas with a lower density of crowdsourced measurements; and accurate and scalable measurement of access beyond a binary notion of coverage.

Acknowledgments

This work is funded in part by National Science Foundation Smart and Connected Communities grant NSF-1831698.

References

1. 2019 Broadband deployment report. Federal Communications Commission. https://www.fcc.gov/reports-research/reports/broadband-progress-reports/2019-broadband-deployment-report.
2. Broadband Internet: FCC's data overstate access on tribal lands. Government Accountability Office (2018), https://www.gao.gov/products/gao-18-630.
3. Connect America Fund (CAF). Federal Communications Commission (2017), https://www.fcc.gov/general/connect-america-fund-caf.
4. Coverage area. Skyhook, https://www.skyhook.com/coverage-map.
5. Decennial census of population and housing by decades. U.S. Census Bureau (2010), https://www.census.gov/programs-surveys/decennial-census/decade.2010.html.
6. Ex parte comments of the National Telecommunications and Information Administration. NTIA (2019), https://www.ntia.doc.gov/files/ntia/publications/ntia_comments_on_modernizing_the_fcc_form_477_data_program.pdf.
7. FCC releases data on mobile broadband deployment. Federal Communications Commission, (2019). https://docs.fcc.gov/public/attachments/DA-16-1107A1_Rcd.pdf
8. Fida, M-R. and Marina, M.K. Impact of device diversity on crowdsourced mobile coverage maps. In *IEEE CNSM* (2018).
9. Form 477 mobile voice and broadband coverage areas. Federal Communications Commission. (2019), https://www.fcc.gov/form-477-mobile-voice-and-broadband-coverage-areas
10. GeoPandas 0.10.2+0.g04d377f.dirty. GeoPandas (2019), http://geopandas.org/
11. ICT facts and figures International Telecommunications Union (2017), https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf.
12. Kahan, J. It's time for a new approach for mapping broadband data to better serve Americans. Microsoft on the Issues (April 8, 2019), https://blogs.microsoft.com/on-the-issues/2019/04/08/its-time-for-a-new-approach-for-mapping- broadband-data-to-better-serve-americans/
13. Li, Z., Nika, A., Zhang, X., Zhu, Y., Yao, Y., Zhao, B.Y., and Zheng, H. Identifying value in crowdsourced wireless signal measurements. In *Proceedings of the 26th Intern. Conf. on World Wide Web* (2017), 607–616.
14. Lubek, B. https://github.com/caarmen/network-monitor
15. Lutu, A., Perino, D., Bagnulo, M., Frias-Martinez, E., and Khangosstar, J. A characterization of the COVID-19 pandemic impact on a mobile network operator traffic. In *ACM Internet Measurement Conference* (2020).
16. Major, D., Teixeira, R., and Mayer, J. No WAN's land: Mapping U.S. broadband coverage with millions of address queries to ISPs. In *ACM Internet Measurement Conference* (2020).
17. Mobile Deployment Form 477 Data. Federal Communications Commission (2017), https://www.fcc.gov/mobile-deployment-form-477-data.
18. Mobile fact sheet. Pew Research Center (2019), https://pewresearch-org-preprod.go-vip.co/pewinternet/fact-sheet/mobile/.
19. NCHS urban-rural classification scheme for counties. U.S. Census Bureau (2013), https://www.cdc.gov/nchs/data_access/urban_rural.htm.
20. Nurmi, P., Bhattacharya, S., and Kukkonen, J. A grid-based algorithm for on-device GSM positioning. In *Proceedings of the 12th ACM Intern. Conf. on Ubiquitous Computing* (2010), 227–236.
21. Open data. OpenCellID (2019), https://opencellid.org/downloads.php.
22. Prieger, J.E. Mobile data roaming and incentives for investment in rural broadband infrastructure. Pepperdine University, School of Public Policy Working Papers, Paper 69 (2017), https://digitalcommons.pepperdine.edu/sppworkingpapers/69.
23. Roberts, E., Beel, D., Philip, L., and Townsend, L. Rural resilience in a digital society. *J. of Rural Studies 54* (2017), 355–359.
24. RWA calls for FCC investigation of T-Mobile coverage data. Rural Wireless Association (2018), https://ruralwireless.org/rwa-calls-for-fcc-investigation-of-t-mobile-coverage-data/.
25. RWA welcomes FCC investigation into violation of mobility fund Phase II mapping rules. Rural Wireless Association (2019), https://ruralwireless.org/rwa-welcomes-fcc-investigation-into-violation-of-mobility-fund-phase-ii-mapping-rules/
26. Tse, D. and Viswanath, P. *Fundamentals of Wireless Communication*. Cambridge University Press (2005).
27. Urban and rural. U.S. Census Bureau (2019), https://www.census.gov/programs-surveys/geography/guidance/geo-areas/urban-rural.html.

Tarun Mangla (tmangla@uchicago.edu) is a postdoctoral scholar at the University of Chicago, IL, USA.

Esther Showalter is a Ph.D. student at the University of California, Santa Barbara, CA, USA.

Vivek Adarsh is a Ph.D. student at the University of California, Santa Barbara, CA, USA.

Kipp Jones is chief technology evangelist at Skyhook, Boston, MA, USA.

Morgan Vigil-Hayes is an assistant professor at Northern Arizona University, Flagstaff, AZ, USA.

Elizabeth Belding is a professor at the University of California, Santa Barbara, CA, USA.

Ellen Zegura is a professor at the Georgia Institute of Technology, Atlanta, GA, USA.