Technical Annex

The project is based on the aggregation of three separate existing datasets, which are combined to allow statistical analysis of the interactions between broadband infrastructure availability, community characteristics, and household Internet adoption. Each of the datasets is described below, along with the procedures and software utilized for analysis/visualization. Table 1 summarizes the more detailed information provided in the rest of the document.

Table 1: Dataset description

Dataset	Description	Period	File format	Source	URL
Wireline Consumer Availability	Contains information about fixed broadband availability (provider and advertised speed) by city block	As of December 2015	Shapefile	CPUC	www.cpuc.ca.gov
American Community Survey (ACS)	Ongoing survey that contains demographic information by census block group	2014 (5-year estimates)	Comma-delimited	Census Bureau	www.census.gov
American Community Survey (ACS)	Ongoing survey that contains estimates about Internet household adoption by PUMA	2015 (1-year estimate)	Comma-delimited	Census Bureau	www.census.gov

Dataset #1: Wireline Consumer Availability

This dataset contains census block-level data on advertised wireline speed (upload and download), provider name, last-mile technology, and total number of providers. It was constructed from the data collected by the California Public Utilities Commission (CPUC) as part of its California Broadband Map initiative. The CPUC issues a data request to California broadband providers on a yearly basis. The data is then geo-processed and validated using a variety of alternative datasets, in particular the FCC's Form 477 data. The 2015 Consumer Wireline geodatabase file can be found in shapefile format here.

After downloading the shapefile data we utilized ArcGIS to map the database and clip out the data of interest, which included geographic boundaries for every block represented within

Los Angeles County (about 98,000 blocks). Specifically, we used ArcGIS link tools to match spatial databases with the relational databases by referencing the standard GEOID system defined by the U.S. Census Bureau.

The attribute table of this database was later exported as a comma-delimited file to calculate new variables such as the number of broadband providers per block and alternative speed tiers (such as the FCC's threshold of broadband as a minimum of 25 Mbps download/3 Mbps upload). In addition, we included new population variables by block (such as total population and density) with data sourced from the Census Bureau. The comma-delimited file was then imported to statistical software (Stata and R). In addition, Gephi was used to create a network graph that illustrates ISP competition at the census block level.

Dataset #2: Wireline Consumer Availability + Community Demographics

This dataset aggregates the variables contained in dataset #1 to the census tract level, to enable merging with demographic data sourced from the ACS. Block population was used as weigh to compute broadband availability variables at the census tract level. There are about 2,340 census tracts in Los Angeles County.

The ACS contains an extensive set of individual and household-level socioeconomic variables, including education attainment, income, race, occupation and many others. For precision, we used the ACS 5-year 2014 estimates rather than the most current (but more imprecise) ACS 1-year 2015 estimates.

However, questions related to Internet adoption were only included in the ACS in 2013. For this reason, disaggregated results at the block level will only be released in 2018 (5-year estimate). Therefore, adoption data was only available at the PUMA level as described below.

Dataset #3: Wireline Consumer Availability + Community Demographics + Internet adoption

Internet adoption questions were not included in the ACS until 2013. As a result, the Census Bureau has so far only released data for 2015 at the PUMA level. PUMAs are large geographical aggregations with rough correspondence to neighborhoods or communities. In Los Angeles County, there are 69 PUMAs, which contain at least 100,000 residents. More info about PUMAs can be found here. 2014 PUMA-level data for Los Angeles County was sourced from the American FactFinder.

The broadband availability in dataset #1 data was aggregated to the PUMA level, weighing by block population. This final dataset thus contains broadband availability data, demographic data, and Internet adoption data at the household level. This is our most important dataset, for it enables the analysis of interactions between broadband supply, community characteristics and Internet adoption.