What We Can Learn from Selected, Unmatched Data: Measuring Internet Inequality in Chicago

James Saxon*[1] and Dan A. Black†[2]

[1]Department of Computer Science, University of Chicago
5730 South Ellis Avenue, Suite 263, Chicago IL 60637, USA
[2]Harris School of Public Policy Studies, University of Chicago
1307 East 60th Street, Chicago, IL 60637, USA

May 5, 2022

Abstract

By integrating a "big" dataset of Internet Speedtest® measurements from Ookla® with data on household incomes from the American Community Survey (ACS), we attempt to measure Internet speeds across income tiers. In the Ookla data, each measurement is technically rigorous but the sample frame is unknown. The ACS provides necessary information on income and Internet access from a known sample frame. Our likelihood combines these data and endogenizes selection effects to identify Internet speed distributions by income tier. We credibly identify the speed distribution for middle and high-income households. However, because the participation rate of low-income households in the Speedtest data is so limited, the speed estimates for these households are not identified.

Keywords: selection effects, Internet, big data, geographic data

1 Introduction

Full participation in modern societies requires usable broadband Internet access. Internet access is used for work, for play, for communication and social interactions, for banking, billing and other economic transactions, for health and for entertainment (see Appendix). This was true before the coronavirus pandemic and it will be true when the pandemic subsides. But the pandemic has underscored the urgency of the "digital divide," between those with and without access to the Internet, especially as it relates to the "homework gap" for children expected to participate in remote learning without appropriate connections or equipment.

*Corresponding author: james.saxon@gmail.com
†danblack@uchicago.edu

The American Community Survey (ACS) tells us that 83% of households nationally have a broadband Internet subscription. These data also highlight variation at the tract level and benefit from a well-controlled sampling strategy. However, they dichotomize access in a fairly rudimentary way: the presence or absence of a connection with a downstream *bandwidth* (data rate) of at least 25 Megabits per second (Mbps) and an upstream bandwidth of at least 3 Mbps. The ACS data cannot tell us how the *quality* of connections varies among households. Internet Service Providers (ISPs) report service plans available in each Census block on the Federal Communication Commission's (FCC's) "Form 477." But the FCC data represent installed infrastructures rather than realized services. Who is getting performant Internet?

In this project we leverage Speedtest Intelligence® performance data, crowd-sourced by Ookla to compare bandwidths (colloquially, Internet *speeds*), used by households in the Chicago area. While these data are the gold-standard of Internet speed measurement, they are not from a probability sample. The data are from a "convenience" sample. In this paper, we ask: What can we learn from "big data" with potentially selected samples? The fundamental problem is that some individuals *choose to run a Speedtest* while others *do not*. This selection problem bears similarities to one familiar to economists: wages are observed only for participants in the labor market. Unfortunately, our measured Internet speeds are not linked to individual-level covariates. This limits the adjustments that can be made through traditional methods.

We begin by describing the assumptions required for inference about the Internet speeds of the *human* population, using the Speedtest on its own. Without appealing to any external dataset, realistic assumptions permit us to construct the relative likelihoods of falling in different Internet service tiers, in Census tracts across the city of Chicago.

We next seek the joint distribution of income and Internet speed. To estimate this, we must combine other data sources into our analysis. In this paper we use data on tracts' income composition from the Census Bureau ACS. Separately, the Speedtest data provide us with the marginal distribution of Internet bandwidths and the Census data provide the marginal distribution of incomes. Naïvely, we might attempt to construct the joint distribution using the Fréchet-Hoeffding bounds (Heckman et al., 1997). However, we have no matching variable with which to narrow these bounds. Moreover, the standard methods do not allow us to address the substantial selection biases that are apparent in our data.

We therefore propose a new model that both endogenizes the selection effect and allows us to estimate the joint distribution. We achieve this by constructing a likelihood that incorporates counts of households by income and Census tract from the ACS with both the volume and distributions of the Speedtest data, again by Census tract. We show that while our method can identify Internet speeds for middle and upper income households, the Speedtest data, despite its "bigness," does not provide us with sufficient information to identify the distribution of speeds for lower income households.

Before concluding, we illustrate how our model may be extended, using the FCC's data on fiber availability to evaluate the impact of that availability on measured speeds.

2 Earlier Literatures

Despite widespread public attention to the digital divide, recent academic work on class-based disparities in Internet accesss and performance in American cities is surprisingly limited and dated. Using a supplement to the Current Population Survey (CPS) on Internet access, devices, and use, as well as their own surveys for Chicago, Mossberger et al. quantified disparities in Internet use in American cities (2013). They usefully delineated the multiple (interacting) levels of the digital divide (skills, infrastructures, access, performance), and the many affected populations (between genders, race or ethnicity, income levels, and regions). In this paper we focus on differential performance between income levels. Hilbert emphasized the need for bandwidth measures and performed this work at the national level (2016), but analogous work has not been performed within communities in the United States. In Britain, Riddlesden and Singleton (2014) used data similar to ours to report mean download speeds within English districts; while they describe the same, substantial variation in sampling rates that we have already noted, they did not explore the implications of this variation for the reported means, or for any inference about populations living *within* those districts. Here, we again take up the challenge, with bandwidth quantified as the maximum achieved data rate in Megabits per second (Mbps), in the downstream direction.[1]

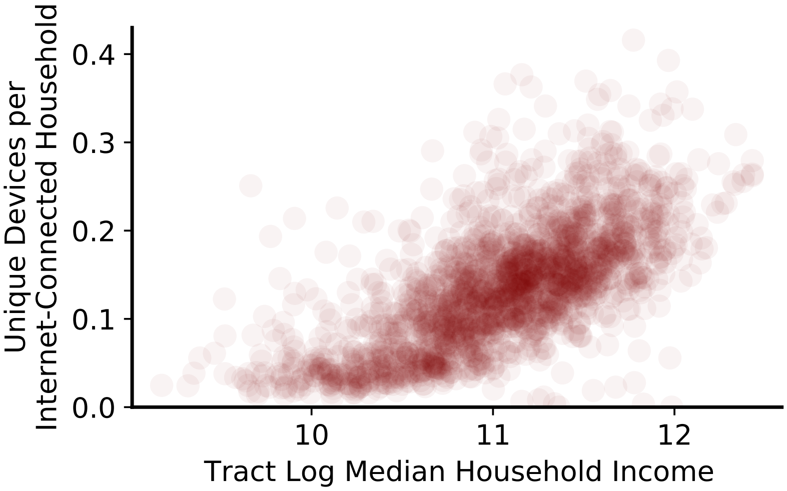
Methodologically, this paper continues an important vein of work on statistical inference with self-selected samples (Winship and Mare, 1992). Although one can naïvely trade accuracy or bias for precision (Elliott and Haviland, 2007), the more-robust and sophisticated strategy is to model the selection effect (Baker et al., 2013). That is the strategy we pursue here. Our work also engages past work on data combination – the construction of a joint distribution from multiple data sources. Like other previous projects in this space, our observable, Internet bandwidth is simply not available in existing controlled samples. At the most conservative level, Fréchet-Hoeffding bounds can be constructed from the univariate distributions, as already noted; if additional "matching" variables are available, these bounds can be tightened (Ridder and Moffitt, 2007). But in our situation (a) no matching variable is available and moreover (b) the univariate distribution of interest has severe selection bias that we must also address (cf. Figure 1).

Finally, readers may recognize parallels between our work and raking or post-stratification methods, in the sense that our procedure determines the observed size of strata from data. Our method differs fundamentally from this literature, however, because the stratification variable (income) is not observed in the Speedtest data.

In short, our formulation of the combined sample selection and data combination problem is distinct from past work. We believe, however, that this structure will find broad applications in the modern, "big data" era, in which researchers must perform rigorous inference with large, unaligned convenience samples.

[1] Download speed is the value typically quoted and is used for our results in the main text. Although we relegate results to the Appendix, upload speeds are in some ways a cleaner measure. The reason for this is that download bandwidths often exceed what consumer equipment viz., Wi-Fi, can actually handle. For that reason, observed upload speeds are more likely than observed download speeds to align with what consumers buy.

Figure 1: ***Number of devices running Ookla Speedtest measurements per Internet-connected household, as a function of Census tract log median household income.*** *Data represent the four-county region centered on Chicago. The y-axis is truncated for visual clarity, excluding 0.4% of tracts.*



3 Data on Internet Performance: Promise and Challenges

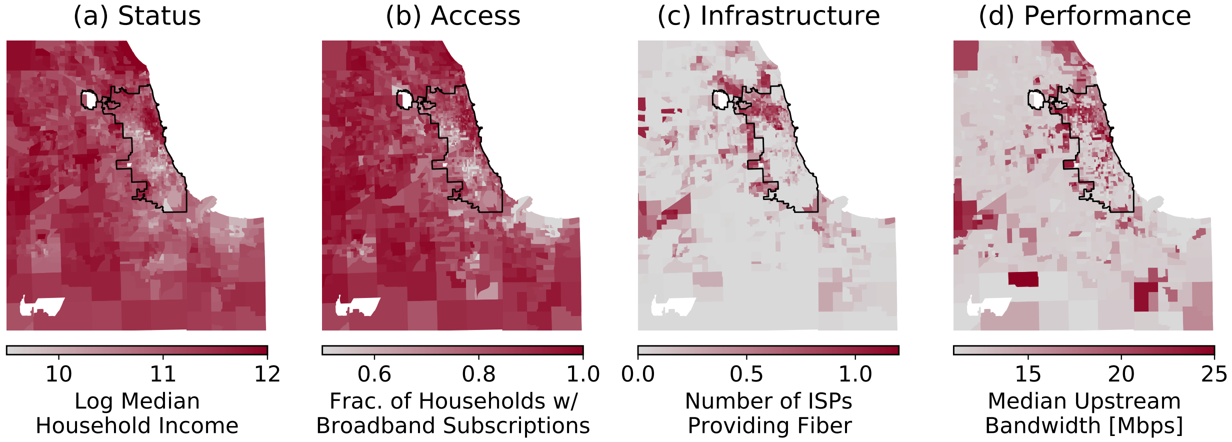
There are several public datasets on Internet access. We discuss several, highlighting the promise and challenges of each. These data cover a variety of measures relating to broadband access. These measures, illustrated in Figure 2, include *access*, understood as broadband subscriptions, *performance*, quantified as bandwidth, and *infrastructure*, represented as the availability of fiber service to consumers. We now describe each in more detail.

3.1 Access: Data from the Census

The American Community Survey (ACS) measures the presence of Internet devices and broadband Internet subscriptions at the Census tract level. The measure of Internet access, the self-reported, dichotomized presence or absence of a subscription with a downstream bandwidth of 25 Mbps and an upstream bandwidth of 3 Mbps, is coarse and captures only the extensive margin. It tells us neither what speeds are actually experienced in a household, nor how much Internet speeds differ between high- and low-income groups. The Census does release tract-level estimates that cross this self-reported Internet access with bins of household income. A great strength of the ACS is of course its rigorous sampling methodology. The 5-year estimates and 2019 vintage of the ACS were used in this analysis.

Similar data are available in a supplement to the CPS for the National Telecommunications and Information Administration (NTIA). Those data are in some ways richer – the CPS elicits respondents' patterns of use (Appendix Table 1) – but because of the limited sample size of the CPS the geographic granularity of the data is coarse (metropolitan regions).

Figure 2: ***Distinct measures of Internet access in Chicago.*** *Each plot presents on "concept" of access and a variable capturing that concept, in the Chicago region. The City of Chicago itself is outlined in black. (a) Shows socioeconomic status or ability to pay, illustrated as median income and (b) presents household broadband subscriptions, both from the American Community Survey. Data reported to the FCC 477 delineate deployed infrastructure, shown here as the availability of fiber (c). Finally, median upstream bandwidths from Ookla Speedtest measurements show performance (d). Plots (a) and (b) are very consistent: access depends strongly on pure ability to pay. Performance depends more closely on infrastructure, as can be seen by comparing (c) and (d); this relationship is exploited in Section 4.3.*


(a) Status
(b) Access
(c) Infrastructure
(d) Performance
10
11
12
Log Median Household Income
0.6
0.8
1.0
Frac. of Households w/ Broadband Subscriptions
0.0
0.5
1.0
Number of ISPs Providing Fiber
15
20
25
Median Upstream Bandwidth [Mbps]

3.2 Infrastructure: Form 477 Reports

On the Federal Communications Commission's (FCC's) Form 477, Internet Service Providers (ISPs) report the services that they offer and contract for in each Census block of the United States. The FCC publishes the data on offerings but not provision. These data have notable limitations. The criterion for "offering" a contract is very weak: it *could* be offered to at least *one* household within a Census block. Recent work comparing 477 reports with ISPs' own online subscription systems has shown that the reports overcount even this loose definition of offers (Major et al., 2020). Earlier work using the same strategy noted close correspondence between 477 reports and Google's Fiber deployments in Provo, UT and Austin, TX Grubesic et al. (2019). Notwithstanding, the data quantify available infrastructures, including the deployment of residential fiber, which we use later in this paper.

3.3 Performance: Ookla Speedtest Intelligence Data

Speedtest Internet performance data are crowd-sourced globally by Ookla. Publicly available data are aggregated at the level of geographic "quadtiles," but Ookla has provided us with disaggregate Speedtest Intelligence data for four counties in the Chicago region (Cook, Will, and DuPage counties in Illinois, and Lake County, Indiana). We use one year of data, from 2020. Each line of data represents one Speedtest. Variables include the unique ID of the device that ran the measurement, the time of the test, its location (latitude and longitude, determined via GPS), the ISP used for the connection (determined from the IP address), and the results of the test: downstream and upstream bandwidths, latency (the round-trip time for a signal to reach a server), and jitter (variation in the latency). The Census tract of each test is determined from its latitude and longitude. We require tests to be executed on a fixed-line (as opposed to cellular) consumer broadband connection, based on the ISP. There are 5.5 million tests on fixed-line connections, from 339 thousand unique GPS-enabled devices (smartphones and tablets) in the four-county region.

We aggregate tests at the device level and assign each device a single "home" tract, determined as its modal tract at night.[2] We then take the median value across tests within the home location, for each device. In doing this, we aim to mitigate the impact of frequent users, some of whom perform hundreds of tests. Note that individuals and households may have several devices; we cannot distinguish this in the data. Further, we do not know *who* runs tests or *why* they choose to run them – either at all, or at the times that they do. For example, users may be motivated to run tests when they are frustrated by their connection, or when they have just set up a service or upgraded equipment. These factors cannot be addressed by our selection methods, below. Speeds observed are those experienced at endpoint devices (usually, post Wi-Fi), which is typically less than the bandwidth delivered to the home and available at the router.

It is worth emphasizing that from a technical and infrastructure perspective, Ookla Speedtest Intelligence is the gold-standard of Internet performance measurement. Performance measures are also publicly available from Measurement Lab (M-Lab), but the test protocol was not historically able to saturate a connection (measure very high speeds, a technical limitation), and the server infrastructure has proven unreliable. Further, M-Lab cannot measure true geographic coordinates; it relies instead on IP geolocation. In general, commercial IP geolocation is not adequately accurate for tract-level, demographic work (Ganelin and Chuang, 2019; Saxon and Feamster, 2021). The FCC's Measuring Broadband America (MBA) also provides performance data. However, the sample is much smaller and in fact no devices are identified within Cook County (where Chicago lies). The MBA sample is stratified by ISP and Census region rather than by population (Office of Engineering and Technology and Consumer and Governmental Affairs Bureau, 2016); like the Ookla data, it is biased towards wealthier neighborhoods.

Speedtest measurements originate disproportionately from wealthier neighborhoods, even taking into account unequal rates of broadband subscriptions across neighborhoods (Figure 1). The correlation in the four-county Chicago region between devices per Internet-connected household and tract log median household income is 0.61. Clearly, the sampling is not random.

[2] Ties are broken based on the tract in which the most tests were executed on weekends and then overall, and finally by the greatest duration between the first and last test in the location.

4 Home Internet Performance in Chicago

4.1 Relative Performance Between Tracts: Non-parametric Estimates

We begin our analysis by considering what may be learned by relying solely on the Speedtest data. Toward that end, let $m_{s,t}$ denote the number of speed measurements observed in a particular tract, t at a given speed s. There are four speed categories $s \in \{0, 1, 2, 3\}$, delimited by $[0, 32)$, $[32, 82)$, $[82, 182)$ and $[182, \infty)$ Mbps; these bounds are defined so that each encapsulates roughly one quarter of tests overall. If we define $m_t = \sum_s m_{s,t}$, then we may write the probability mass function (pmf) of devices' speed tiers as

$$g_t(s) = m_{s,t}/m_t. \tag{1}$$

Denoting the count of households with Internet subscriptions in a particular tract by n_t and the count of such households by Internet speed tier by $n_{s,t}$, the pmf of Internet speeds by Internet-connected *household* is

$$f_t(s) = n_{s,t}/n_t. \tag{2}$$

What is the relationship of speeds $g_t(s)$ to households $f_t(s)$?

It is helpful to define the identity

$$m_{s,t} \equiv a_{s,t}\, n_{s,t} \tag{3}$$

where $a_{s,t}$ is a selection adjustment parameter between Internet-connected households and the number of devices. Note that $a_{s,t}$ is *not* the probability of a household in a tract appearing in the data: households can run tests on multiple devices, each of which would be included once. Two assumptions would give us identification of $f_t(s_t)$ immediately. First, we could assume $a_{s,t} = a$ is a constant. In this case, it is unnecessary to estimate the value of a as it would cancel in our calculation of $g_t(s)$. While providing identification, this assumption would directly contradict the evidence presented in Figure 1.

An alternative that would allow us to trivially match the data in Figure 1 would be to assume that $a_{s,t} = a_t$ so that all of the variation in the rate of running a Speedtest is from Census tract. As above, there would be no need to estimate a_t, since it would cancel in the calculation of $g_t(s)$. In this case, $f_t(s_t)$ would follow immediately from the Speedtest data, as simply $g_t(s_t)$. But while this assumption would be able to account for the heterogeneity depicted in Figure 1, we consider it no more credible. If there is significant heterogeneity in the Speedtest rate as a function of income *between* tracts, it seems unlikely that the rate is homogeneous as a function of household income *within* tracts.

This leaves one last obvious assumption: $a_{s,t} = a_s$. In this case, the rate of running a Speedtest relies on the speed of the Internet connection. This assumption is appealing because performant home Internet and high rates *measuring* Internet performance both evidence an interest in Internet performance. Unfortunately, without auxiliary data there is no way to identify a_s. To see why, consider a possible candidate, say, $a_{s=0} = a^0$. If we double this solution to $a_{s=0} = 2 \times a^0$ and the same time halve the estimate of the size of the population, $n_{s=0,t}$ for each tract, then the observables $m_{s=0,t}$ are unaffected. Without constraints on the number of households at each speed tier within a tract, we may set these sampling rates to what ever value we wish.

However, we can still use any statistic that does not require the parameters a_s to be known. A promising candidate is the ratio of households in high versus low speed tiers, by tract,

$$\frac{\Pr(s=3|t)}{\Pr(s=0|t)} = \frac{f_t(s=3)}{f_t(s=0)} = \frac{n_{s=3,t}/n_t}{n_{s=0,t}/n_t} = \frac{n_{s=3,t}}{n_{s=0,t}}. \tag{4}$$

According to our assumption, this is

$$\frac{\Pr(s=3|t)}{\Pr(s=0|t)} = \frac{m_{s=3,t}/a_{s=3}}{m_{s=0,t}/a_{s=0}} = \frac{m_{s=3,t}}{m_{s=0,t}}\,\frac{a_{s=0}}{a_{s=3}}. \tag{5}$$

Since the adjustment rates a_s are unknown, the value of the ratio is indeterminate. However, because $a_{s=0}/a_{s=3}$ is constant across tracts, $m_{s=3,t}/m_{s=0,t}$ is proportional to the ratio of the probabilities. To fix the constant of proportionality, we normalize the statistic by dividing the number of test devices in each speed tier and tract by the total number devices at that speed tier. Our statistic is then

$$\frac{\Pr(s=3|t)}{\Pr(s=0|t)} \propto R \equiv \frac{m_{s=3,t}/m_{s=3}}{m_{s=0,t}/m_{s=0}}. \tag{6}$$

The statistic R, which remains proportional to the probability ratio, can also be understood as a tract's share of the region's high-tier devices relative to its share of low-tier devices. The share of high- to low-tier devices is clearly correlated with income, though the device-weighted correlations are perhaps lower than anticipated: just 0.28 ($p \ll 0.001$). (The correlation is illustrated in the Appendix.) Our non-parametric approach suggests where speeds are high or low, but it does not tell us *who* it is within tracts that experiences high or low speeds.

4.2 Performance by Income Level: Endogenizing the Selection Effect

We next aim to estimate performance as a function of household income. Technically, we seek the joint distribution of income and Internet bandwidth, $f_t(i_t, s_t)$. The ACS provides us with counts of households with and without Internet subscriptions, by tract and income category. We collapse these counts in three broader bins: less than \$35,000, \$35,000 to \$75,000, and above \$75,000.

Between our two data sources, we have two marginal distributions: the distribution of bandwidths $g_t(s)$ as before, and the distribution of incomes $h_t(i)$, with the income categories $i \in \{0, 1, 2\}$ given above. As above, the sampling rate in the Speedtest data is related to household counts by unknown adjustment parameters $a_{i,s,t}$:

$$m_{i,s,t} \equiv a_{i,s,t} n_{i,s,t}. \tag{7}$$

Again, although we observe the sums $m_{s,t} = \sum_i m_{i,s,t}$ and $n_{i,t} = \sum_s n_{i,s,t}$, we do not observe any of the individual components $m_{i,s,t}$, $n_{i,s,t}$, or $a_{i,s,t}$. What to do?

Economists may appreciate the aim of adapting Heckman-esque problems of self-selection (Heckman, 1974; Winship and Mare, 1992) to the era of big data. There are important parallels here, but also interesting differences: we observe tests for a self-selected subset of people in each tract, rather than incomplete observables for a subset of the observations. We observe groups rather than individuals.

Above we assumed that the adjustment rates a were a function of speed; we now assume that they depend only on income: $a_{i,s,t} = a_i$. The rates are constant within income tiers across speed categories and Census tracts. In view of the correlation between a_i and income shown in Figure 1, as well as our current application, this assumption seems prudent. We also assume that, conditional on income, devices' likelihoods of falling in each speed tier are constant across Census tracts.

Our approach can then be summarized by

$$m_{ts} = \textstyle\sum_i n_{ti}\, a_i\, p_{si}. \tag{8}$$

The expected number of tests in each bin $m_{ts}^{\text{exp.}}$ is derived from the Census' count of households with broadband Internet n_{ti}, multiplied with the income-dependent adjustment parameter a_i, and the probability of a test in an income bin falling in each speed tier, p_{si}. Both m_{ts} and n_{ti} are observed; we will estimate a_i and p_{si}. This equation, which undergirds our subsequent likelihood estimation, is illustrated in Figure 3.

The number of households per income bin is not equal; the bin breaks were constrained by existing divisions from the Census. Since every test falls into one tier, the bin probabilities p_{si} entail three instead of four degrees of freedom per income level. Computationally, we estimate these p_{si} indirectly, via the bin boundaries, denoted b_{si}. The b_{si} represent the value of the CDF of speed conditional on income for the Internet-connected population, at the cut points already given: $\{0, 32, 82, 182, \infty\}$ Mbps. The boundaries and probabilities are related by $p_{si} = b_{si} - b_{s-1,i}$. Clearly, $b_{0i} = 0$ and $b_{4i} = 1$; we also constrain $b_{s+1,i} \geq b_{si}$. These cuts are common across Census tracts, which provides us with p_{si}. The constraints, as well as one enforcing non-negative adjustment rates a_i, are implemented as ad-hoc penalties in the likelihood function. The reason for using penalty functions will become apparent when we discuss our results in Table 3.

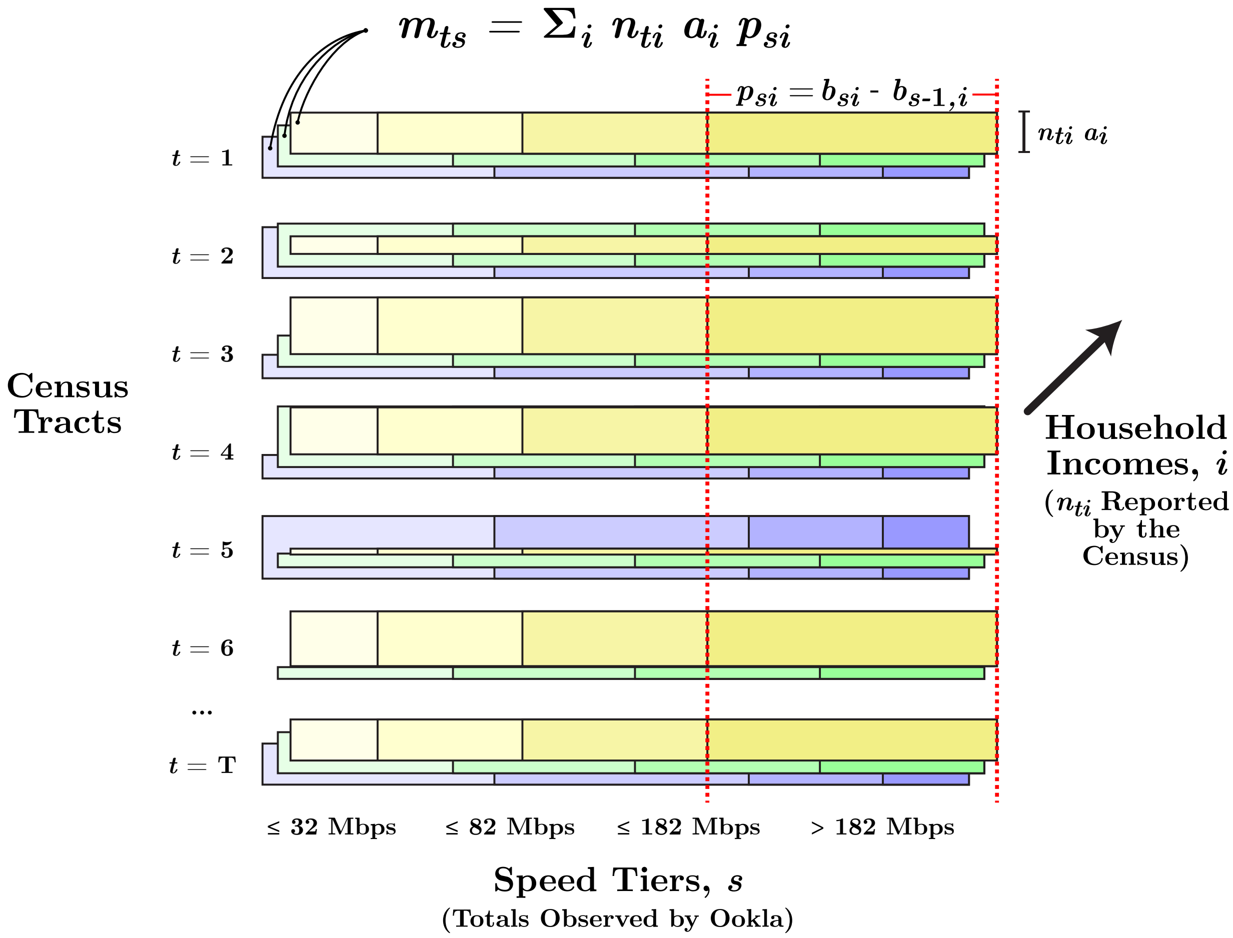
Our approach is to put enough structure on the data to allow us to recover estimates of the joint distribution function without robbing the data of the ability to inform us on the substantive issue. There is, of course, a fundamental tension between the assumptions necessary to recover the joint distribution and the need to extract information from the data.

Now turn to the likelihood. For notational simplicity, allow the vectors $\boldsymbol{n}$, $\boldsymbol{a}$, and $\boldsymbol{b}$ to denote the full sets of bin populations, selection adjustment effects, and speed distribution parameters. The Internet-connected populations $\boldsymbol{n}$ are known from the ACS and fixed; we estimate $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{b}}$. The basic likelihood (without boundary constraints) is then simply the product of the Poisson probabilities of the observations in every tract × speed bin, whose expectations were defined in Equation 8. The negative log likelihood is then

$$-\log \mathcal{L} = -\textstyle\sum_t \sum_s \log[p_{\text{Pois}}(m_{ts}|\boldsymbol{n}, \boldsymbol{a}, \boldsymbol{b})]. \tag{9}$$

In all, there are twelve parameters in our nominal fit: one selection effect and three speed bin boundaries, for each of three income levels.

Figure 3: ***Illustration of the likelihood function, Equations 8 and 9.*** *The observables are the numbers of tests per speed tier in each Census tract m_{ts} and the counts of households at each income tier n_{ti}. By assumption, the proportion of testing devices in each speed tier p_{si} and the number of devices per household a_i, conditional on income, are both constant across tracts.*



For identification, we are relying on heterogeneity in the marginal distributions of income across tracts to identify the selection rate and inform the joint distribution of income by speed. Trivially for Chicago, this requires enough tracts to cover the fitted degrees of freedom. More substantively, estimates require adequate heterogeneity in the income distribution between tracts. Without the income heterogeneity, the test rates per household cannot be isolated by income, and the relative proportions of speeds by income cannot be determined. Again, these requirements are met abundantly in Chicago.

It should be apparent that we can modify our setup by "crossing" income with other indicators, to derive the distribution of speeds conditional on those joint requirements. To illustrate this, we will separate tracts with and without fiber availability, to identify the impact of living in neighborhoods that are or are not served with fiber on upper-income households' Internet performance.

The basic assumption of our model is that the rate at which households generate unique devices in a speed bin is captured by income groups i, and is constant across tracts. Denote the household-level analogs to the variables above by the subscript k. For instance, a_k is one household's number of Speedtest devices. The assumption can then be expressed as:

$$f(s_k|i) = f(s_k|i, t, a_k = 0) = f(s_k|i, t, a_k > 0) \tag{10}$$

The credible part of this assumption is that income does capture a great deal of the variation in both adjustment rates and speed. What is less credible is that the households that *do* generate tests are expressing an interest in Internet performance that we expect may also be in evidence in their purchasing patterns. That would suggest that

$$E[s_k|i_k, a_k = 0] < E[s_k|i_k, r_k > 0]. \tag{11}$$

Indeed, we can confirm empirically that at the tract level, positive residuals for testing rates regressed on income are associated with higher shares of devices in the highest speed tier. Further, our income bins are set by the Census and they are quite coarse. Just as higher-income users generate more tests *across* bins (Figure 1), they may also be expected to do so *within* bins:

$$E[i_k|i, a_k > 0] < E[i_k|i, a_k > 0]. \tag{12}$$

All of these factors would lead to bias towards wealthier users in each bin, potentially biasing speeds upwards. In short, the $a_i n_{ti}$ households who run tests may have different (likely, higher) speeds than the $(1 - a_i)n_{ti}$ households who do not. Our method cannot correct for this error, just as it cannot control for the timing or motivations of users' tests.

There are three further subtleties: independence of observations, the estimation of the standard errors, and why we used broadband penetration rates from the Census. Independence among speed by tract bins is required for the product in the likelihood. But are the bins really independent? As written, the likelihood requires that the Poisson process of the group of people at each speed tier generating a count (running a test) is independent from bin to bin. So the probability of every *observation* is independent. Effectively, the tester "realizes" a count for a theretofore latent variable. The other argument would be that bins are of course *not* independent: a "low" Speedtest moves an observation from a higher bin to a lower bin, violating independence. This is analogous, however, to a "piece of mail"

Table 1: *Downstream Internet speeds in the Chicago region, by household income tier.*

	Sub.	Adj.	CDF at s Mbps		
	ACS		$s = 32$	$s = 82$	$s = 182$
Income	$\mathcal{F}_i$	a_i	b_{1i}	b_{2i}	b_{3i}
Lower	0.62	0.003 (0.000)	0.635 –	0.643 –	0.643 –
Middle	0.84	0.062 (0.001)	0.478 (0.009)	0.636 (0.010)	0.712 (0.010)
Upper	0.94	0.248 (0.001)	0.214 (0.001)	0.483 (0.002)	0.751 (0.001)

Notes: Parameters are from Nelder-Mead maximum likelihood estimation of Equation 9. Standard errors are from the Hessian matrix under assumptions of normality. The "universe" of the CDFs represented by b_{si} is Internet-subscribed households. Values are from Speedtest users, not the entire population. Biases are discussed in Section 4.2. The adjustment parameter for low-income households reaches its lower bound (0), so that the estimates of the speed distribution are not constrained (standard errors undefined). For reference, the shares of households by income tier that subscribe to broadband Internet, $\mathcal{F}_i$, are tabulated from the American Community Survey.

arriving a day "late," in the canonical Poisson arrival process. It is the full reality of the data-generating process that generates counts.

The second issue is that the standard errors are based on the asymptotic estimates of the variance from the Hessian matrix (derivatives in the information matrix). The assumption is then that the parameters (not the counts) are normally distributed. This is defensible in general, but it fails when parameters reach the edge of their bounds, so that the derivative is not well-defined. This happens in practice, for a_0, as we shall see.

Finally: why have we relied on the Census to break out the extensive margin of broadband access, instead of using simply households by income? On its own, the model cannot disentangle selection due to *having Internet* from selection due to *running a test*. In the Speedtest data, those who do not run a Speedtest are indistinguishable from those who do not have Internet service. The Census allows us to break this degeneracy. This means that our estimates represent Internet testing rates *conditional on Internet*. The fraction of households with Internet subscriptions, $\mathcal{F}_i$, can be estimated directly from the ACS.

Results for the model are displayed in Table 1. Several results warrant mention. First, in terms of information only available from the ACS data, having Internet service increases substantially with income. Indeed, for the lower income group, fully 38 percent of households do not have Internet service, but only 6 percent of households from the upper income group do not.

Second, for the lower income group, the number of devices running a Speedtest is extremely low, only three in a thousand: the test rate is converging to its lower limit. Hence, one should have no confidence in the speed estimates for the lower income tier because the Speedtest data simply does not provide enough coverage for low-income households. Indeed, the standard errors on the estimates of b_{s0} are not well-defined. This point nicely illustrates the advantage of combining data. Without appealing to the ACS, it would have been impossible to identify the low-sampling rates of lower income households.

Similarly, higher-income households (incomes in excess of \$75k), generate 3.7 times more tests than middle-income households (with incomes $35k-75k): 0.25 vs 0.062 testing devices per Internet-connected household. The Speedtest data are highly biased toward upper income households, as was apparent in Figure 1.

Finally, devices from high-income households are half as likely as devices from the middle-income group to record speeds in the "basic" connection tier, below 32 Mbps (0.478 vs 0.214). On the other hand, the middle income group records a greater proportion of tests at the top of the distribution, with downstream speeds in excess of 182 Mbps, than the highest income group ($1 - 0.71 = 0.29$ vs $1 - 0.75 = 0.25$).

In Appendix Table 1, we present estimates from a model that treats the selection adjustment rate as a single parameter, shared across income bins. With these estimates, the share of middle income households fitted to have bandwidth less than 32 Mbps falls from 0.48 to 0.29. This behavior is exactly what you would expect, if wealthier households were to generate a disproportionate share of tests. In that case, speeds observed in each tract would be dominated by tests generated by wealthier households. Figure 1 showed that this disproportionate testing rate is the reality. As further shown in the Appendix (Table 1), the gaps between the middle and upper income groups are larger when considering upstream bandwidths, and revert to the "expected" direction: 24% of upper-income Internet-connected households have upstream bandwidths in excess of 32 Mbps, while the estimate for middle-income households is consistent with 0.

4.3 The Impact of Infrastructure: Discontinuous Fiber Deployments

We can embroider the model above, incorporating the plausibly causal impact of fiber deployment on bandwidths measured by consumers. Figure 2(c) shows the availability of fiber broadband subscriptions by tract, while (d) displays median upstream bandwidth. At a local level, fiber follows not demographic lines but a regional park and a highway: "X" marks the spot, in the northwest of the city. This is our spatial discontinuity. Visual inspection shows that it is strongly linked to speed even though these observables are really distinct.

We incorporate this effect in the likelihood by separating the high-income bins (which account for most of the tests) between fiber and non-fiber tracts. A tract is defined as having fiber if more than half of the blocks within it have at least one provider. Fiber infrastructures are associated with a dramatic increase in upstream data speeds (see Figure 2(c-d)). However, the differentials in downstream performance for upper income households in tracts with and without fiber are fairly small: there are reductions in the two slowest speed categories and modest increases in the two fastest categories.

Results for this model are shown in Table 2. Devices associated with upper-income households are 8.4% more likely to register speeds in the top two speed tiers, in neighborhoods where fiber subscriptions are available, than in neighborhoods where they are not. The estimates of the adjustment parameters change very modestly from those in Table 1. Upper income households with access to fiber do have faster Internet service than the middle income category, but upper income households without access to fiber remain a bit less likely to have Internet services at the fastest tier.

Table 2: *Downstream Internet speeds in the Chicago region, separating tracts with and without fiber infrastructures.* *Tracts with and without fiber are also distinguished for upper-income households, based on ISPs offerings as reported on FCC Form 477.*

	Sub.	Adj.	CDF at s Mbps		
	ACS		$s = 32$	$s = 82$	$s = 182$
Income	$\mathcal{F}_i$	a_i	b_{1i}	b_{2i}	b_{3i}
Lower	0.62	0.002 (0.000)	0.003 (0.000)	0.876 –	0.876 –
Middle	0.84	0.066 (0.001)	0.505 (0.009)	0.618 (0.009)	0.731 (0.008)
Upper w/o Fiber	0.94	0.243 (0.001)	0.221 (0.001)	0.500 (0.002)	0.758 (0.001)
Upper w/ Fiber	0.95	0.253 (0.001)	0.170 (0.002)	0.416 (0.003)	0.700 (0.002)

Notes: In this model, the parameter for the CDF of speeds for low-income households has reached its boundary constraint at zero. The SE is therefore defined even though the estimate is not meaningful. (See also notes to Table 1.)

5 Discussion

Venturing into the world of "big data," with self-selected, unmatched samples, we obtain estimates that are unavailable from traditional sources. In our application, we exploit a crowd-sourced sample of Internet speed measurements. At first glance, these data appear to offer an excellent opportunity to measure the variation in Internet speeds across income classes: the "digital divide." We show, however, that when we combine the Speedtest data with Census data from the ACS and construct a simple selection model, the rate of sampling from low-income households is so low that performance cannot be measured for this group. In our view, this finding highlights the need for data combination and care for selection effects. The low testing rate of low-income households is also apparent, with the natural consequence that the CDF of Internet performance is unconstrained for that group. At the same time, we do extract meaningful estimates of Internet performance for middle- and upper-income households. We have also shown how other plausible assumptions – a testing rate dependent on speed instead of income – allows for useful non-parametric estimates.

Research on the geography of Internet access, and data combination in general, is relevant to an evolving theoretical understanding of poverty measurement. Recent work across economics and sociology has emphasized the multidimensional nature of poverty and the need for measurement of the distinct, interlocking resources necessary for a complete, flourishing life in a neighborhood (Heckman and Mosso, 2014; Sharkey and Faber, 2014; Sampson, 2011). Internet access is one such facet. Other applications may include ride-share trips, package deliveries, or social engagement as proxied through 311 calls. None of these datasets can be linked to Census microdata, but they can often be aggregated geographically to the Census tract level. This is exactly what we have done here, and we hope that our selection model will find broad applications.

Along the lines of classic work on health care and resource accessibility, access is determined not only by infrastructures but by a cascading set of economic and social factors; *pro forma* availability is not on its own a complete metric or, typically, even the relevant one (Harvey, 2009; Aday and Andersen, 1974). Shifting from the FCC's 477 "offerings" to realized performance is a step forward.

Our goal in this work has not been to provide consistent estimates of the underlying joint distribution function of Internet speed and income. In our view, the inherent limitations of the Speedtest data and the aggregation of the ACS data preclude consistent estimates of the relevant distribution functions. Instead, we aim to reduce errors by offering an approximate model.

Our approach does, however, caution against two reactions to "big data." First, naïvely applying data with unknown sample frames can result in serious errors of inference. Without combining data, one might mistakenly have assumed that the Speedtest data had coverage of low-income households. The second view, which we consider equally naïve, is to dismiss quality data out of hand, if the sample frame is not known. Doing so obviously precludes any possibility of inference and seems equivalent to asserting there is no information in the data.

References

Aday, L. A., Andersen, R., 1974. A framework for the study of access to medical care. Health Services Research 9 (3), 208–220.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1071804/

Baker, R., Brick, J. M., Bates, N. A., Mattaglia, M., Couper, M. P., Dever, J. A., Tourangeau, R., June 2013. Report of the AAPOR task force on non-probability sampling. Tech. rep., American Association for Public Opinion Research, Alexandria, VA, retrieved Dec. 23 2021 from .

Elliott, M., Haviland, A., 12 2007. Use of a web-based convenience sample to supplement a probability sample. Survey Methodology 33, 211–215.

Ganelin, D., Chuang, I., 2019. Ip geolocation underestimates regressive economic patterns in mooc usage. In: 11th International Conference on Education Technology and Computers. Association for Computing Machinery, New York City, New York, pp. 268–272.
URL https://dx.doi.org/10.1145/3369255.3369301

Grubesic, T. H., Helderop, E., Alizadeh, T., 2019. Closing information asymmetries: A scale agnostic approach for exploring equity implications of broadband provision. Telecommunications Policy 43 (1), 50–66.
URL https://www.sciencedirect.com/science/article/pii/S0308596118300363

Harvey, D., 2009. Social Justice and the City, revised edition Edition. University of Georgia Press, Athens, Georgia.

Heckman, J., 1974. Shadow prices, market wages, and labor supply. Econometrica 42 (4), 679–694.

Heckman, J. J., Mosso, S., 2014. The economics of human development and social mobility. Annual Review of Economics 6 (1), 689–733.

Heckman, J. J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. The Review of Economic Studies 64 (4), 487–535.

Hilbert, M., 2016. The bad news is that the digital access divide is here to stay: Domestically installed bandwidths among 172 countries for 19862014. Telecommunications Policy 40 (6), 567–581.

Major, D., Teixeira, R., Mayer, J., 2020. No wan's land: Mapping u.s. broadband coverage with millions of address queries to isps. In: Proceedings of the ACM Internet Measurement Conference. IMC '20. Association for Computing Machinery, New York, NY, USA, p. 393419.

Mossberger, K., Tolbert, C., Franko, W., 2013. Digital Cities: The Internet and the Geography of Opportunity, kindle Edition. Oxford University Press, Oxford, UK.

Office of Engineering and Technology, Consumer and Governmental Affairs Bureau, 2016. 2016 technical appendix: Measuring broadband america fixed broadband. Tech. rep., Federal Communications Commission, Washington, DC, retrieved Feb. 28 2022 from https://data.fcc.gov/download/measuring-broadband-america/2016/Technical-Appendix-fixed-2016.pdf.

Ridder, G., Moffitt, R., 2007. The econometrics of data combination. In: Heckman, J. J., Leamer, E. E. (Eds.), Handbook of Econometrics. Vol. 6. Elsevier, Amsterdam, pp. 5469–5547.

Riddlesden, D., Singleton, A. D., 2014. Broadband speed equity: A new digital divide? Applied Geography 52, 25–33.
URL https://www.sciencedirect.com/science/article/pii/S0143622814000782

Sampson, R. J., 2011. Neighborhood effects, causal mechanisms and the social structure of the city. In: Demeulenaere, P. (Ed.), Analytical Sociology and Social Mechanisms. Cambridge University Press, New York, pp. 227–249.

Saxon, J., Feamster, N., 2021. Gps-based geolocation of consumer ip addresses.

Sharkey, P., Faber, J. W., 2014. Where, when, why, and for whom do residential contexts matter? moving away from the dichotomous understanding of neighborhood effects. Annual Review of Sociology 40 (1), 559–579.

Winship, C., Mare, R. D., 1992. Models for sample selection bias. Annual Review of Sociology 18 (1), 327–350.

A Common Uses of the Internet

The Current Population Survey's supplement for the National Telecommunications and Information Administration (NTIA) elicits broadband and technology penetration, and common uses of the Internet. Table 1 This most recent cycle of this survey was in November 2019, before the Coronavirus pandemic; more recent work-from-home numbers would presumably be higher.

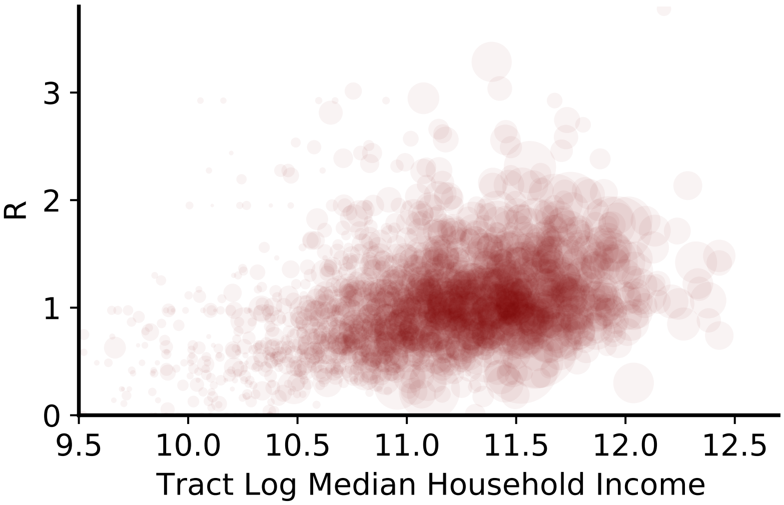
Table 1: ***Proportion of US adults who use various Internet technologies and services.*** *Data from National Telecommunications and Information Administration's November 2019 supplement to the Current Population Survey.*

Adults who		Common Activities	
access the Internet		among Internet Users	
Anywhere	0.81	Texting	0.92
At Home	0.76	Email	0.91
At Work	0.39	Watching Videos	0.74
		Social Media	0.73
Adults who		Finance/Banking	0.72
use devices		Calling	0.50
Smartphone	0.76	Services (Uber, &c)	0.35
Computer	0.62	Remote Work	0.25

B Illustration of the Non-Parametric Variable, R.

Figure 1 shows R as defined in Equation 6, as a function of tract log median household income. R is the ratio of a tract's share of the region-wide low-speed and high-speed devices. The correlation between R and neighborhood income is 0.28.

Figure 1: ***Relative shares of low- and high-speed devices, as a function of tract log median household income.*** *Shares are normalized by the region-wide rates, see Equation 6 and discussion. Point sizes represent the number of test devices in each tract.*



C Results for a Model with a Single Adjustment Rate

Here, we manipulate the model defined in Equations 8 and 9, making a single adjustment rate or proportionality between tracts' Internet-connected households and test devices. Results are shown in Table 1. With respect to the nominal estimates of Table 1, this model fits fewer middle income households as falling within the lowest speed tier: the CDF at 32 Mbps ($b_{1,1}$) as 0.29 instead of 0.48. This is as expected. This model, which we consider less credible, treats all households as generating an equal share of data. Empirically, we know that wealthier neighborhoods generate more data, we expect that this is also true at the household level. If that is true, then the devices affecting the "middle-income" estimates in Table 1, are biased towards wealthier and presumably faster connection tiers.

Table 1: *Downstream Internet speeds in the Chicago region, under the assumption of constant sampling rate.* *By contrast with the nominal model (Table 1), estimates of the bin boundaries for the low-income group are defined. As expected, speed estimates for the middle-income group are also higher than for the nominal model: fewer households in the lowest speed bin.*

	Sub.	Adj.	CDF at s Mbps		
	ACS		$s = 32$	$s = 82$	$s = 182$
Income	$\mathcal{F}_i$	$a_i = a$	b_{1i}	b_{2i}	b_{3i}
Lower	0.62	0.149 (0.000)	0.424 (0.008)	0.690 (0.009)	0.841 (0.008)
Middle	0.84	0.149 (0.000)	0.286 (0.007)	0.503 (0.008)	0.703 (0.007)
Upper	0.94	0.149 (0.000)	0.180 (0.002)	0.450 (0.002)	0.739 (0.002)

(See notes to Table 1.)

D Results for Upstream Bandwidths

Results for upstream bandwidths are presented in Table 1.

For downstream bandwidths, the boundaries between speed tiers were set (at 32, 82, and 182 Mbps) so that each tier captured roughly one quarter of tests overall. For upstream bandwidths, the bin boundaries are set in the same way. The bin bounds thus shift to to $[0, 9)$, $[9, 13)$, $[13, 24)$, and $[24, \infty)$ Mbps.

Table 1: ***Upstream Internet speeds in the Chicago region, by household income tier.***

	Sub.	Adj.	CDF at s Mbps		
	ACS		$s = 9$	$s = 13$	$s = 24$
Income	$\mathcal{F}_i$	a_i	b_{1i}	b_{2i}	b_{3i}
Lower	0.62	-0.000 (0.000)	0.022 (0.000)	0.262 –	0.262 –
Middle	0.84	0.074 (0.001)	0.566 (0.008)	0.816 (0.008)	1.000 (0.000)
Upper	0.94	0.243 (0.001)	0.214 (0.001)	0.468 (0.001)	0.759 (0.001)

(See notes to Table 1.)