Online Appendix for "The Internet as a Tax Haven?"

David R. Agrawal

A Appendix (For Online Publication)

This online appendix provides information on data sources, data construction, and supplementary material / results mentioned in the text of the paper.

A.1 Baseline Data

A.1.1 Taxes

The raw sales tax data were acquired from Pro Sales Tax (2003-2011) and assembled in Agrawal (2015) and Agrawal (2014). To summarize that assembly procedure, Agrawal (2015) merges a cross-section of the tax data to Census data. Thus, Agrawal (2015) restricts the sample to municipalities that are identified Census Places.[27] When doing this, Agrawal (2015) name merges data provided by the American Community Survey (ACS) to the tax data. These cross-sectional data are extended to a panel data setting in a later paper, Agrawal (2014). These data have complete coverage of all local sales taxes in the United States at the monthly frequency from 2003 to 2011. The sales tax data used in this paper and the data merging procedure are described in detail in these two prior papers.

A.1.2 Driving Distances

Using the geo-spatial network data in Agrawal (2015), which modifies Lovenheim (2008) to calculate distance to the border, I also data on the driving time to the nearest state border major road crossing. Distance to the border is the time (in minutes) that minimizes the driving time from the population weighted centroid of a town to the closest state border and a major road intersection. By using minutes rather than miles, these data are able to capture the true cost of driving to the border. Agrawal (2015) then identifies a town as being in a high-tax state if its state has a higher state sales tax rate than the nearest neighboring state; a town is in a low-tax state if its state sales tax rate is lower in the own state than

[27] A Census Place is generally an incorporated place with an active government and definite geographic boundaries such as a city, town, or village. In some western states, a Census Place may be an unincorporated place that has no definite boundaries or government. Census Places contain some locations that may not have legal authority or jurisdiction to set sales taxes.

the nearest neighboring state. The driving data assembled in Agrawal (2015) were merged to the panel data in Agrawal (2014) by Agrawal and Mardan (2019).

A.1.3 Primary Internet Data from Broadband Map

Data on local internet penetration comes from the July 2011 version of the National Broadband Map, which is collected by the National Telecommunications and Information Administration (NTIA) in conjunction with the Federal Communication Commission (National Broadband Map 2011). These data are supplemented with state level data from Form 477 (FCC 2003-2011) and pricing data from the FCC Reference Book (FCC 2008).

A.1.4 Internet Usage at the Local Level

Unlike the *penetration* data in the Broadband Map (which is available at the town level), internet *usage* is not available at the town level. For every Census tract, the FCC releases binned data on the percentage of households with a fixed internet connection (FCC 2008-2011), ι such that:

$$\iota = \begin{cases} 0 & \text{if } I^* = 0 \\ 1 & \text{if } 0 < I^* \leqslant .20 \\ 2 & \text{if } .20 < I^* \leqslant .40 \\ 3 & \text{if } .40 < I^* \leqslant .60 \\ 4 & \text{if } .60 < I^* \leqslant .80 \\ 5 & \text{if } .80 < I^* \leqslant 1 \end{cases} \tag{A.1}$$

where I^* is the true fraction of households with an internet subscription. Because these are tract level data and tracts can cross town lines, I aggregate up to the town level by assuming that the value of ι is uniformly distributed within the tract. I calculate the fraction of the town area in any Census tract using ArcGIS software and Census mapfiles. I then calculate the municipal-level value of ι by aggregating up using the percent of area in the Census tract as weights. This yields a continuous value of internet usage between zero and five. I then assume that each integer value of ι is the mid-point between its bin's extreme values. Given the town level measures constructed using area weights are not necessarily integers, I calculate a value of internet usage based on how far the decimal is from each of the two nearest integers. This gives me a continuous measure of internet usage between zero and 0.90. Note that this variable contains measurement error because: (1) ι contain error and the midpoint assumption also introduces error, and (2) for towns that are smaller than a Census tract, the value in the FCC data may be substantially off if the distribution is not uniform within a tract.

A.1.5 Computer and Internet Usage at the State / National Level

Measures of computer and internet usage come from the Current Population Survey. Aggregate state information is from CPS (2003-2011a) and micro information is from CPS (2003-2011b). I supplement this with information on online shopping from Census (1998-2012).

A.1.6 Nexus Data

Bruce, Fox and Luna (2015) construct nexus data by hand. Bruce, Fox and Luna (2015) visit the websites of the largest e-tail firms and attempt to make a retail purchase from each state. They code the firm as having nexus if retail sales taxes are levied on that transaction. These authors provide me with summary data as to whether the state has an above average nexus variable along the with the quartiles in the nexus distribution for each state.

Finally, I ask someone with access to Compustat data to count of the number of firms in traditional retail sectors that are headquartered within a state. This measure does not directly get at nexus as in Bruce, Fox and Luna (2015); rather, it is designed to proxy for nexus by state. The underlying assumption is that states with more retail firms headquartered in the state are more likely to have more firms with nexus. To construct this, I classify firms by their NAICS codes as "retail firms" – firms traditionally remitting retail sales taxes on most purchases. The count is the total number of retail firms headquartered in a state as listed in the Compustat database. This variable is used in the state level panel data regression.

A.1.7 Controls and Revenue Data

All baseline controls come from the 2000 Census (Census 2000) or various American Community Surveys (Census 2005-2011). Political controls are downloaded from a database at MIT (MIT Election Data and Science Lab 2018). Sales tax revenue data comes from the Census of Governments (Census 1967-2012).

A.1.8 Summary Statistics

Figure A.1 shows examples of areas serviced by four or more broadband providers (2011) in a large metro area and in smaller towns. The providers often elect to service areas that do not start or end at municipal borders. As is clear, providers make decisions to enter particular parts of a municipality, likely based on historical infrastructure.

A.2 Theoretical / Policy Justification for Proxy Variable

Why does the fraction of households with access to four or more providers a good proxy for internet access? The existing industrial organization literature provides some theoretical and empirical evidence that increased competition by broadband companies will increase take-up of the internet. For example, Faulhaber and Hogendorn (2000) shows that "the subgame equilibrium capacity and price strategies depend only on the number of networks to which a household has access." Thus, the number of providers serving an area (the outcome of the first stage) is, from a theoretical perspective, the most important determinant of price in this industry. Second, as shown in Distaso, Lupi and Manenti (2006) and verified empirically, inter-platform competition such as DSL versus cable technologies (rather than intra-platform competition), increase internet usage. If individuals have more types of choices – and they will in places with more providers – then they are more likely to adopt a particular technology. Prieger and Hu (2008) also show empirically that competition in broadband markets is an important contributing factor of the Digital Divide that exists across races even though prices do not vary substantially across various markets; the authors provide evidence that suggests more intense competition increases internet usage because companies compete more intensely on installation, service fees, and other charges.

All of this evidence taken together suggests that markets with more intense competition will have higher internet usage rates and that penetration is also correlated with online purchases, which should then feedback into the tax setting behavior of the jurisdictions. The economics literature is also complemented by the views of the NTIA, who write in the National Broadband Map, "The primary factors that people consider when deciding what type of broadband internet service to subscribe to include service availability, connection speed, technology and price." The United States National Broadband Plan studies some of the data on competition and notes that competition in residential broadband markets "provides consumers the benefits of choice, better service and lower prices."

A.3 Implementation of Lubotsky and Wittenberg (2006)

As noted in the text, I can aggregate up to a single coefficient of interest using:

$$\beta^{\rho} = \sum_{n=1}^{N} \beta^{n} \frac{cov(\tau_i,\, I_i^n)}{cov(\tau_i,\, I_i^1)}. \tag{A.2}$$

The expression is normalized by $cov(\tau_i,\, I_i^1)$. This means that the procedure is an interpretation procedure where the coefficient is scaled such that a one unit increase corresponds to a one unit increase in I_i^1. In order to be able to compare this procedure to my other results, I select this normalization such that the results are comparable to the fraction of the population with access to four or more providers. Lubotsky and Wittenberg (2006) show that attenuation bias will be most reduced when estimating β^{ρ} and Bollinger and Minier (2015) show that including all proxy variables in the regression minimizes the bias on other coefficients in the regression as well.

To implement this, I use variables indicating the percent of consumers with access to one or more, ..., eight or more providers, download speeds greater than 768k, ..., download speeds greater than 1gig, upload speed greater than 10,000k, upload speeds greater than 50,000k, the total number of providers in the jurisdiction, the total number of residential providers in the jurisdiction, and the total number of broadband providers for various speeds. The latter of these are constructed from form 477 tract level data. The "..." imply that I use all data for values in between the given range. As noted in Lubotsky and Wittenberg (2006), the procedure is not a license to include every variable the researcher may think is a proxy variable. Proxy variables can affect other control variables (Bollinger 2003) and "adding proxies that absorb the effects of covariates rather than proxying for the latent variable will be particularly damaging." For this reason, I exclude most of the type of technology variables (dsl, optical fiber, copper, etc.).

A.4 IV Variable Construction

A.4.1 TV and Phone Usage

TV and phone usage at the county and state level are obtained from the 1956 City and County Data Book and 1960 City and County Data Book (Census 1955-1960). These data are at the county rather than municipal level.

A.4.2 Lightning as an IV for IT at the State Level

Andersen et al. (2012) show that lightning strikes are a powerful predictor of IT usage (at the state level) in the United States during the period from 1996 to 2006. Andersen et al. (2012) argue that in places with high lightning density, more power disturbances occur. These power disturbances increase the cost of investing in IT, which then lowers IT investment and internet usage.

As an instrumental variable, I construct a measure of the flash density of lightning using the National Oceanic and Atmospheric Administration's Severe Weather Database. I use data on the annual number of ground strikes from 1986 to 2011 to construct the per year average number of strikes per square mile. Define the flash density of lightning in a jurisdiction as

$$lightning_i = \frac{(\sum_{t=1986}^{2011} strikes_{i,t})/T}{area_i}, \tag{A.3}$$

where $strikes_{i,t}$ is the number of lightning strikes in jurisdiction i in year t, T is the total number of years, and where $area_i$ is the area of the jurisdiction. At the state and county level, this variable can be constructed from publicly released data (NOAA 1986-2012a) at each geographic level; but at the local level, it must be constructed using grid level data on lightning strikes.

To do this at the local level, I obtain grid level data on all lightning strikes from 1986 to 2011 from the National Oceanic and Atmospheric Administration (NOAA 1986-2012b). The data I obtained provide me the precise 4km grid cell that the lightning strikes hit on the map. I then aggregate from the grid level up to the municipal level accounting for the fact that some grids can cross municipal boundaries. I do this aggregation by weighting by the grid area within a municipality and assuming the lighting strikes were distributed uniformly within the grids.

A.4.3 Internet Backbone.

Durairajan et al. (2015) and Durairajan and Barford (2016) construct maps of the internet backbone.[28] However, these maps are not eligible for public use outside of a secure portal due to national security reasons. For this reason, using the data maps constructed by these authors, which are mapped into counties in Durairajan and Barford (2016), I first determine if a county has internet infrastructure running through it. Then, I calculate the crow-flies distance (or linear distance) from the population weighted centroid of each town to the nearest county containing this physical network assuming the long-haul network runs through the midpoint of the county. For towns in a county with such infrastructure, I set this variable to zero. The crow-flies distance avoids any possible correlation with infrastructure.

A.5 Additional Data Sources

Agrawal, David R. 2015. Unpublished data from “The Tax Gradient: Spatial Aspects of Fiscal Competition.” American Economic Journal: Economic Policy, 7(2): 1–30.

Agrawal, David R. 2014. Data from Table 1,2, and Appendix Table 4-10 of “LOST in America: Evidence on Local Sales Taxes from National Panel Data.” Regional Science and Urban Economics, 49: 147–163.

[28] See also Durairajan et al. (2013).

Bruce, Donald, Fox, William F., and Luna, LeAnn. 2015. Unpublished data from "E-Tailer Sales Tax Nexus and State Tax Policies".

Compustat Industrial. 2003-2011. Available: Standard & Poor's/Compustat. Accessed approximately 2013.

CPS. 2003-2011a. "Computer and Internet Usage Tables" https://www.census.gov/topics/population/computer-internet/data/tables.html (Accessed approximately 2013)

CPS. 2003-2011b. "Current Population Survey, Computer and Internet Usage Supplement" Accessed at https://data.nber.org/data/current-population-survey-data.html (accessed approximately 2014)

Einav, Liran, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan. 2014. Data from Table A6 of "Sales Taxes and Internet Commerce." American Economic Review, 104(1): 1–26.

FCC. 2003-2011. "Form 477 State Data" https://www.fcc.gov/internet-access-services-reports (Accessed approximately 2014)

FCC. 2008-2011. "Form 477 Tract Data" https://www.fcc.gov/internet-access-services-reports (Accessed approximately 2014)

FCC. 2008. "Reference Book" https://www.fcc.gov/oea-archived-data-and-statistical-reports (Accessed approximately 2014)

FCC. 2011. "National Broadband Map" https://broadbandmap.fcc.gov/#/ (Accessed approximately 2013)

MIT Election Data and Science Lab, 2018, "countypres_2000-2016.tab", County Presidential Election Returns 2000-2016, https://doi.org/10.7910/DVN/VOQCHQ/HEIJCQ, Harvard Dataverse, V6, UNF:6:ZZe1xuZ5H2l4NUiSRcRf8Q== [fileUNF] (accessed 2019)

NOAA. 1986-2012a. "County and State Summaries" https://www.ncdc.noaa.gov/data-access/severe-weather/lightning-products-and-services (accessed approximately 2015)

NOAA. 1986-2012b. "Gridded Summaries" https://www.ncdc.noaa.gov/data-access/severe-weather/lightning-products-and-services (accessed approximately 2015)

Pro Sales Tax. 2003-2011. "Pro Sales Tax Monthly Database." https://www.prosalestax.com/

United States Census. 1955-1960. "Census County and City Databook" https://www.icpsr.umich.edu/web/ICPSR/studies/12 (Accessed approximately 2016)

United States Census. 1998-2012. "Quarterly e-Commerce Reports" https://www.census.gov/retail/ecommerce/historic_releases.html (Accessed approximately 2014)

United States Census. 2000. "2000 Census" Accessed at https://www.socialexplorer.com/explore-maps (accessed approximately 2014)

United States Census. 2005-2011. "American Community Survey, 5 Year Estimates"

Accessed at https://www.socialexplorer.com/explore-maps (accessed approximately 2013)

United States Census. 1967-2012. “Census of Governments, Revenue Data” Accessed at https://www.census.gov/programs-surveys/gov-finances/data/historical-data.html (accessed approximately 2016)

A.6 Additional Works Cited in the Appendix

Bollinger, Christopher R. 2003. “Measurement Error in Human Capital and the Black-White Wage Gap.” *The Review of Economics and Statistics*, 85(3): 578–585.

Bollinger, Christopher R., and Jenny Minier. 2015. “On the Robustness of Coefficient Estimates to the Inclusion of Proxy Variables.” *Journal of Econometric Methods*, 4(1): 101–122.

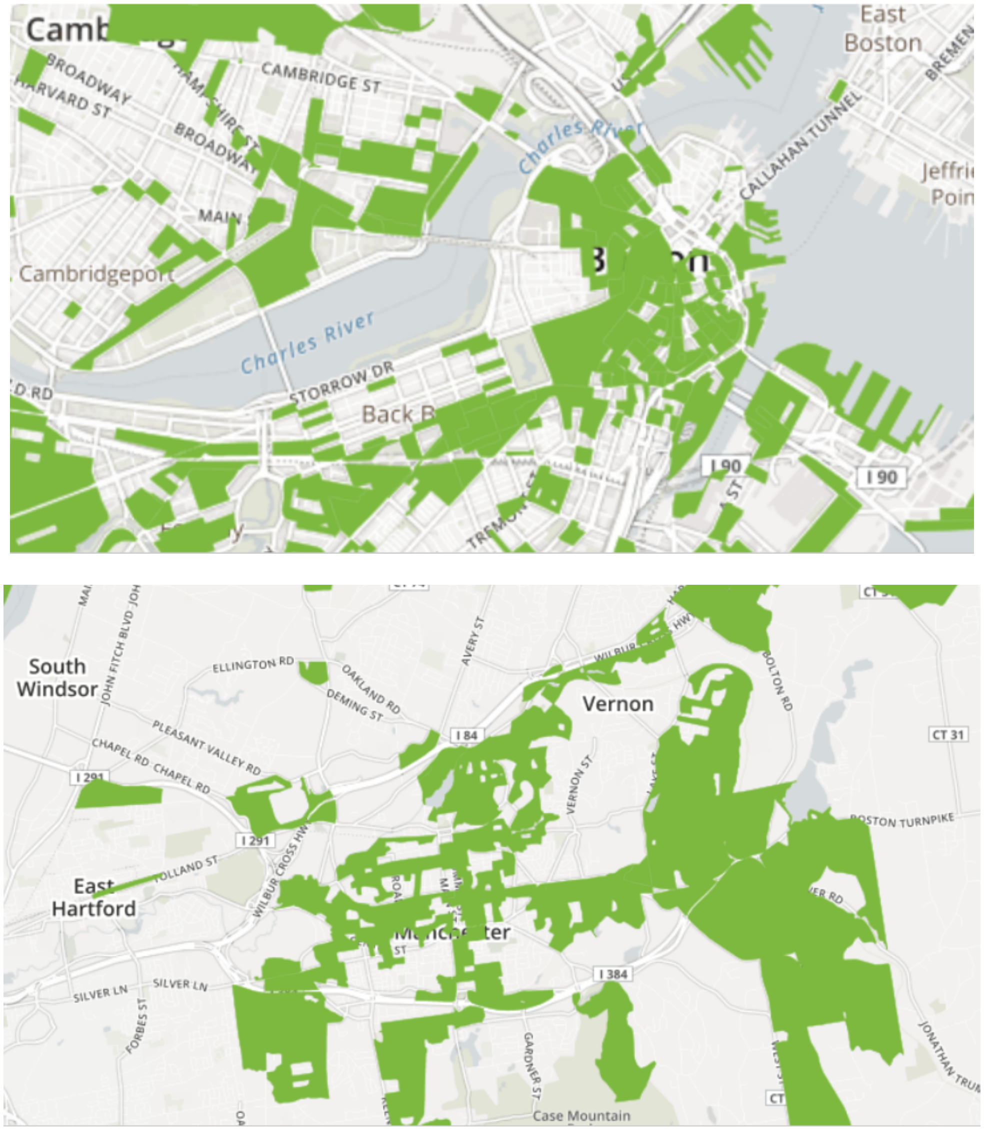
Distaso, Walter, Paolo Lupi, and Fabio M. Manenti. 2006. “Platform Competition and Broadband Uptake: Theory and Empirical Evidence from the European Union.” *Information Economics and Policy*, 18(1): 87–106.

Faulhaber, Gerald R., and Christiaan Hogendorn. 2000. “The Market Structure of Broadband Telecommunications.” *The Journal of Industrial Economics*, 48(3): 305–329.

Lovenheim, Michael F. 2008. “How Far to the Border?: The Extent and Impact of Cross-Border Casual Cigarette Smuggling.” *National Tax Journal*, 61(1): 7–33.

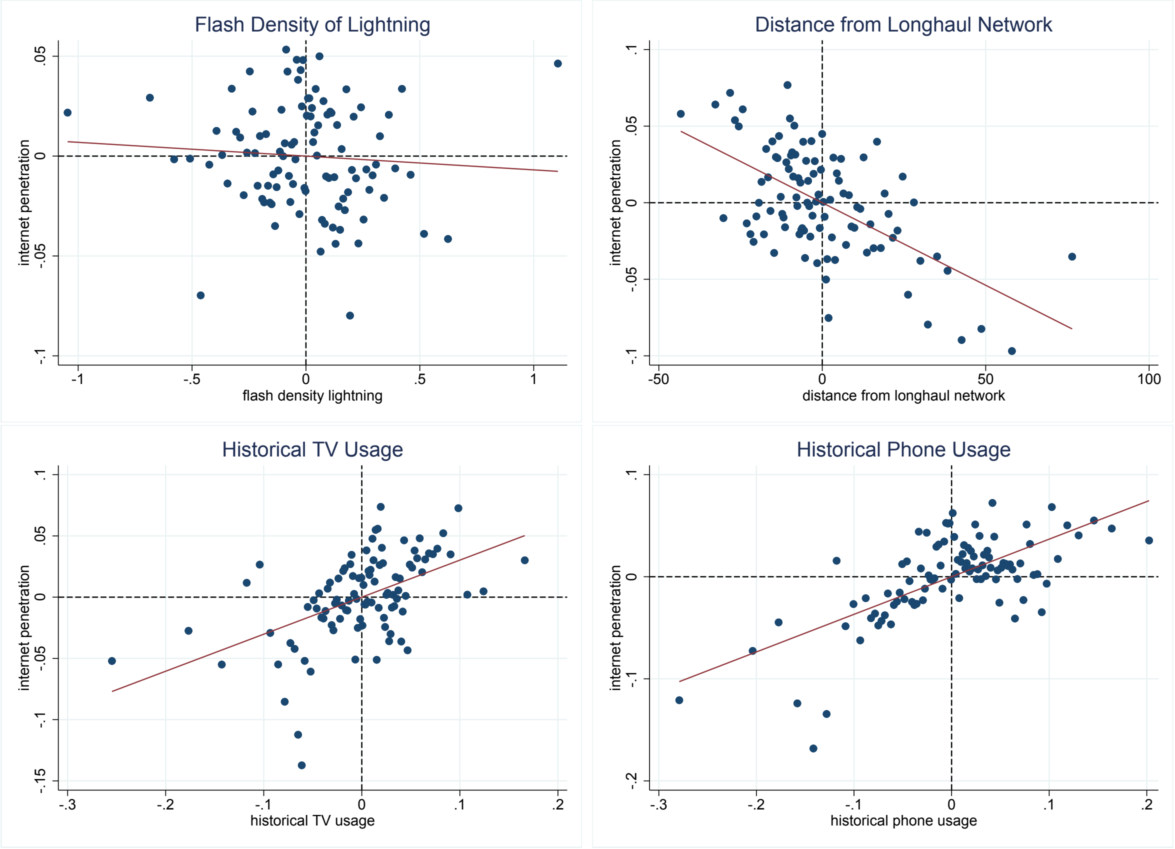
Prieger, James, and Wei-Min Hu. 2008. “The Broadband Digital Divide and the Nexus of Race, Competition, and Quality.” *Information Economics and Policy*, 20(2): 150–167.

Figure A.1: Examples of Service Provider Maps



The upper figure shows areas of Boston and Cambridge, MA that are serviced by more than four broadband providers. The second figure shows a similar map for the smaller towns of South Windsor, East Hartford, Manchester, and Vernon, CT. The provider areas serviced by four or more providers clearly do not correspond to town boundaries.

Figure A.2: Relationship Between Instruments and Internet Penetration at the Local Level



The graph shows the relationship between the fraction of households with access to four or more providers and each instrumental variable in the local cross-sectional data. To make this figure, I regress internet penetration on state fixed effects and control variables from the regression; I repeat this procedure each instrument. I predict the residuals and then I then bin the data into 100 equally sized bins and plot the line of best fit. Note, the slope here is not the precise first stage coefficient because the first stage includes all four instruments.

Table A.1: Regression of Usage on Penetration

Standard Errors in ( ) and R^2 in [ ].

	(1) Usage	(3) eBay	(3) Local Usage
A. Candidate Proxy Variable			
% Households with Access to Any Service	0.051	0.376***	0.132***
	(0.152)	(0.079)	(0.015)
	[0.003]	[0.144]	[0.015]
% Households with Access to Speed ⩾ 6000k	0.296**	0.513***	0.309***
	(0.118)	(0.179)	(0.031)
	[0.087]	[0.262]	[0.094]
% Households with Access to Speed ⩾ 3000k	0.017	0.405***	0.169***
	(0.145)	(0.061)	(0.021)
	[0.000]	[0.166]	[0.027]
% Households with Access to Providers⩾ 1	-0.078	0.331***	0.107***
	(0.091)	(0.042)	(0.013)
	[0.006]	[0.112]	[0.010]
% Households with Access to Providers ⩾ 2	0.108	0.400***	0.180***
	(0.146)	(0.132)	(0.022)
	[0.012]	[0.162]	[0.031]
% Households with Access to Providers ⩾ 3	0.348**	0.498***	0.253***
	(0.144)	(0.181)	(0.027)
	[0.121]	[0.249]	[0.063]
% Households with Access to Providers ⩾ 4	0.446***	0.560***	0.317***
	(0.129)	(0.135)	(0.032)
	[0.199]	[0.312]	[0.100]
% Households with Access to Providers ⩾ 5	0.409***	0.578***	0.347***
	(0.141)	(0.117)	(0.032)
	[0.168]	[0.327]	[0.121]
% Households with Access to Providers ⩾ 6	0.380***	0.442***	0.309***
	(0.122)	(0.124)	(0.031)
	[0.145]	[0.187]	[0.096]
B. Details and Statistics			
N	51	50	29,130
Unit of Analysis	State	State	Town

Each cell represents a different regression. Each row of columns (1) - (3) reports the coefficient on the variable listed, the standard error in (), and the R^2 in [] from the univariate regression of the form $I^*_s = \theta + \delta I_i + \nu_i$ with standard errors robust to heteroskedasticity in columns (1)-(2) and clustered at the state level in column (3). Both I^*_i and I_i are standardized such that δ represents the effect of a one standard deviation increase in penetration. The dependent variable in column (1) is the fraction of homes in a state with internet access at home as measured by the CPS. The dependent variable in column (2) is the per capita number of eBay purchases in the state measured by Einav et al. (2014). In column (3) the dependent variable is local internet usage, which is constructed in section A.1.4. ***99%, **95%, *90%

Table A.2: Correlation of Providers and Prices

A. Specification				Placebo
Internet Penetration Variable:	(1) Residential	(2) Total	(3) $\geq$ 4	(4) Mobile
	Providers	Providers	Providers	Providers
No Controls	-0.577***	-0.161**	-2.730***	-0.168
	(0.200)	(0.076)	(0.257)	(0.312)
	[0.050]	[0.048]	[0.026]	[0.004]
Control for ln(population)	-0.485**	-0.143	-2.128***	0.198
	(0.198)	(0.155)	(0.472)	(0.410)
	[0.070]	[0.048]	[0.052]	[0.040]
Control for ln(population), demographics	-0.383*	-0.200	-2.106***	0.216
	(0.207)	(0.155)	(0.606)	(0.376)
	[0.093]	[0.095]	[0.089]	[0.078]
B. Details and Statistics				
N	94	94	94	94
Unit of Analysis	Locality	Locality	Locality	Locality
Average Price in 2007	\$15.27	\$15.27	\$15.27	\$15.27

Each cell represents a different regression. Standard errors are in ( ) and the R^2 in [ ]. Each row adds various controls. Each cell reports the coefficient on the variable listed in the column heading, the standard error and the R^2 from a town level regression of the form $\varphi_i = a + bI_i + \nu_i$ with robust standard errors. The price of internet services excluding government taxes is φ_i per month and I_i is internet penetration. Column (1) uses the number of residential providers and column (2) uses the number of residential and commercial providers. Column (3) uses the percent of households with access to four or more providers. As a placebo test, column (4) uses the number of cell phone providers. The average monthly price in 2007 is given in the last row of the table. ***99%, **95%, *90%