HARVARD | BUSINESS | SCHOOL

Where the Cloud Rests: The Location Strategies of Data Centers

Shane Greenstein
Tommy Pan Fang

Working Paper 21-042

Where the Cloud Rests: The Location Strategies of Data Centers

Shane Greenstein
Harvard Business School

Tommy Pan Fang
Harvard Business School

Working Paper 21-042

Funding for this research was provided in part by Harvard Business School.

Where the Cloud Rests: The Location Strategies of Data Centers

Shane Greenstein and Tommy Pan Fang[1]

June 2022

This study provides an analysis of the entry strategies of third-party data centers in the United States. We examine the market before the pandemic in 2018 and 2019, when supply and demand for data services were geographically stable. We compare with the entry strategies of major cloud-based data centers for services on demand, which include those known as cloud services. We conclude that third-party firms and cloud providers have different entry strategies. The former favor urban settings more, and appear sensitive to buyer demand for proximity. They trade-off costs of supply, which vary with density; and economies of scale, which cannot be achieved without large volumes of demand. We also find that data center firms providing specialized services display an urban bias. Cloud providers display a lower propensity to locate in urban areas, and they tend to concentrate their building in a small number of locations. We see little evidence to suggest cloud providers will spread their data centers to any but a small number of low-density locations. Our findings support speculation about the likely direction of changes as demand shifts to the cloud, and the location decisions begin to concentrate in the hands of cloud providers.

[1] Harvard Business School, Soldier Field, Boston, MA. 02163. Emails: sgreenstein@hbs.edu and tpanfang@hbs.edu.
Thanks to Patrick Clapp, Christine Snively, and Maggie Kelleher for assistance. Thanks to seminar audiences, Michael Rechtin, Aaron Kulick, Victor Bennett, Alphonso Gambardella, Igal Hendel, Jonathan Timmins, Tim DeStefano, and anonymous reviewers for comments and suggestions. All errors are our responsibility.

I. Introduction

A data center provides storage and computing at a large scale, sharing the cost savings and high performance across many users. Users may either own or rent the facility and equipment, which is housed at a separate location, and may manage it themselves or pay for services. Third-party suppliers can also accommodate a variety of specialized requests, including high reliability (e.g., 99.999% uptime) or privacy regulations, such as those in financial and medical usage. A few large providers have built data centers for their own use—e.g., Google, the first to do so, in 2003. Today some firms rent capacity to users as cloud services. The largest today is Amazon Web Services (AWS), which started in 2006. Its revenues exceeded $62 billion in 2021. Users turn the cloud on or off at will and have the option to grow their data demands, use default software services, and rent additional applications.

Despite their importance and recent growth, no research has investigated the strategies that shape data centers. This study brings attention to the topic. It is the first study to gain insight into the different determinants of data center location strategies. What are the key tensions determining the locations of different types of data centers? The study treats location as endogenous, determined by the demand for a local facility and the costs of a location. Evidence reveals that the third-party data center industry considers trade-offs between buyer demand for proximity and the reduced costs of operating in less expensive locations. The study provides a statistical description of the differences between the third-party data center industry and cloud providers, and it identifies some previously unrecognized trade-offs. While it focuses on measuring the determinants of managerial concerns, the insights also have implications for public policies that aspire to shape managerial choice.

The setting is the pre-pandemic market for data centers. This time period displays comparatively stable geographic demand because most employment took place on-premises. For our purposes, the pre-pandemic era enables us to statistically track determinants of local demand and local supply, which provides insight into underlying determinants of the location of data centers. Just after this period, work from home during the Covid pandemic altered data traffic patterns dramatically. In some locations, buyers and suppliers of data center services responded with new infrastructure plans. The movement for "distributed

architectures" also began to take off just after the period we examine, and with the benefit of hindsight, some of its antecedents will appear in our data and analysis.

The study collects and analyzes a new data set about data centers offered by third parties in the US in 2018 and early 2019, and supplements it with information about the largest cloud providers. The unit of analysis is counties, which vary in their attractiveness to any data center provider. Attractiveness varies due to the size of local industries with high demand for data centers, such as financial activities and information technology, and features of the area that drive up operating costs and construction expenses, such as industrial electricity prices and construction wages.

The study finds a pervasive urban bias in the location of third-party data centers. For example, we find that all large metropolitan areas with over 700,000 population have at least one supplier. Less dense areas may or may not have any. Moreover, local entry rises with the presence of local information industries and intensive data users, such as finance, insurance, and real estate. Because less supply locates in the areas with lower density, a high fraction of buyers in small and medium-sized locations must get their services from non-local suppliers—likely located in the closest major city. Relatedly, we also find supply of more specialty services in denser and more competitive locations. We interpret all these patterns as the result of tension between economies of scale and user preference for proximity.

We find evidence of different strategies between third-party supply and the major cloud providers. The major cloud suppliers display less propensity to locate their own data centers in urban areas. Cloud suppliers are comparatively less responsive to local demand and more responsive to costs, such as electricity, construction wages, and land prices. Cloud providers also tend to concentrate their building in a small number of locations. Because our study focuses on an early moment in the shift towards the cloud, we interpret these differences as an indicator of a shift in where infrastructure locates at the time, and where it will locate in the future. Decisions at large providers will determine the small set of areas affected by this shift.

Contributions

While the supply of internet access has received attention from researchers, only a few studies examine the supply chain of services that make up modern internet architecture (Greenstein, 2020). Very little

research examines infrastructure location. The only instances are a study of Netflix's content delivery networks (CDNs)[2] and a study of the value users place on closer physical facilities for cloud services (Wang et al., 2019). As in this prior research, we do not directly observe demand, but we infer it from observable supplier behaviors, such as their actual entry choices (Xiao and Orazem, 2011). We also share motivation with analysis and measurement of the cloud market (Byrne, Corrado, and Sichel, 2018; Coyle and Nguyen 2018; DeStefano et al. 2019), a new class of infrastructure that contains assets worth hundreds of billions of dollars and that drives major productivity changes. Research focuses on how services enable flexible and inexpensive uses of storage and computation, remove frictions in access to big data applications, and support applications for a mobile labor force (DeStefano et al. 2019; Jin and McElheran 2019). We spotlight an overlooked topic: suppliers providing services in close proximity to users. No research other than Wang et al. (2019) addresses the topic, and they focus on the demand for proximity faced by one cloud provider, while we focus on the broad patterns across all suppliers. We both find evidence that buyers possess a preference for proximity.

We contribute to the studies about urban bias in the strategies of information technology suppliers (Forman et al, 2018). Urban bias arises in the market for broadband (DeStefano, Kneller, and Timmis, 2018; Pew, 2019), business applications (Ceccagnoli et al., 2012; Forman et al., 2009), and markets for technical talent (Tambe, 2014). Urban bias also arises from buyer desire to locate near local third-party vendors who substitute for internal providers (Forman et al., 2008) and from supplier desire to benefit from the increased capacity of a region to innovate (Delgado et al., 2010). We overlap with studies of general purpose technologies that discuss specialty suppliers of upstream internet services (Bresnahan and Gambardella, 1997). In contrast, our study stresses private incentive to meet such demand by spreading fixed costs among nearby users.

Though our study focuses on managerial strategic concerns, it has implications for policy-focused research about the internet's global growth (OECD, 2014). Such research neglects data centers. This gap requires attention. Localities compete for centers with tax abatement and subsidy programs. Regulatory pressures, such as those affiliated with the General Data Protection Regulation (GDPR), also devote

[2] Böttger et al. (2016) find 4600+ servers in 233 locations in 2016.

considerable attention to regulating the location of storage and computation of data (Zhuo et al., 2020). No statistical research backs up or confirms the premises behind these policies.

Section II presents a brief outline of the determinants of private data center locations just prior to the pandemic. Section III describes the data analyzed in the study. Section IV presents the estimation results, and Section V draws several implications for the economics and policy of infrastructure.

II. Determinants of the location of data centers

Size. The industry measures capacity as the maximum electricity required by a building.[3] A minimal data center has a capacity of five megawatts (MW), while the largest, sometimes called "hyperscale," far exceed this size. Potentially limiting factors on the size of a data center are equipment to support and cool servers efficiently and the availability of sufficient land in a dense area. As an example, AWS has encountered trade-offs between costs and performance at hyperscales and, consequently, limits its data centers to 25–32 megawatts.[4] After AWS reaches that capacity, if it wishes to expand it does not add more capacity to an existing building. Instead, AWS commissions the building of another structure.

Cost of construction. A new data center requires at least a year to build, with longer times for enormous projects and specialized designs.The construction costs of a 5MW data center exceed a hundred million dollars, even in the least expensive locations, while the largest data centers can cost several billion (CBRE, 2015). Many factors vary with location. Costs increase when construction wages rise, when land is more expensive, and when the costs of the first generation of equipment increases. It also increases when the costs of building high-capacity connectivity to fast backbone internet lines rises, when the costs of building to (multiple) electrical lines increases, and when tax subsidies and abatements do not defray these expenses. Special features also raise costs, such as reinforced structures to silence vibrations, raised floors to prevent flooding, backup generators to achieve reliability, special cooling equipment, and insulation to mitigate hot weather, and lines to support connectivity to multiple networks.

[3] That capacity measure accommodates different uses for the servers, whether computation or storage. As a rule of thumb, an 8MW data center would typically house just under fifty thousand servers.
[4] See https://www.infrastructure.aws/, accessed December 2020.

Operating costs. Electricity for servers and air conditioning comprises the largest fraction of operating expenses.[5] Operational costs rise (decline) in hot (moderately cool) and humid (dry) regions and vary seasonally. The costs of replacing equipment can accumulate, and abatements for sales tax and property taxes can reduce those costs. The operational costs of a small (large) data center exceed tens (hundreds) of millions of dollars a year and vary by at least several (tens of) millions between locations (CBRE, 2015).

Usage and the preference for proximity. Several factors shape user demand for proximity. Lower latency, which the user experiences as faster response time, motivates suppliers to locate in close proximity to some users. That also motivates locating close to high-speed data lines that reliably support many users and avoid congestion. Financial users and highly skilled programmers of frontier data science have intense demand for lower latency. So do uses that require large volumes of transactions or "big data" applications.[6] Such demands arise in many applications, e.g., video conferences, multi-user streaming, interactive gaming, complex simulations, or other high volume electronic transactions. New use cases continue to emerge in machine learning, where users derive insights from iterating between (big) data input, model development (with millions of parameters), and iteration and analysis.

Other factors also shape the preference for proximity. It can result in the colocation of firms in the same place to facilitate connectivity between them, and some business users intensively value that connectivity. This desire may also be due to less technical behavior, often labeled as "server hugging"—i.e., executives want to be near facilities to manage the physical assets, either to monitor shared facilities for performance (e.g., to adopt cooling or wiring or server design) or retain the option to visit the facility in the event of an urgent situation (e.g., a weather disaster).

Some users do not desire proximity, and virtually all such users possess little desire to visit a facility. One type of user desires geographic diversity to support high reliability in the face of risks of disasters, such

[5] An 8MW data center in a moderate climate approximately splits electricity between servers and AC, for example.
[6] For example, if a big data application uses two data centers, Amazon keeps those within a hundred miles in order to limit the latency between them to one millisecond. See Peter DeSantis, AWS Re: Invent 2020. https://www.youtube.com/watch?v=AaYNwOh90Pg, accessed December 2020.

as fires, earthquakes, and hurricanes. Another type desires generic and non-urgent computing, such as backup storage, where costs primarily drive choices and usage can tolerate delay.

Ownership. Third-party suppliers offer users options to "collocate" in a building shared with others. Collocating enables the buyer to use innovations offered by the supplier and redeploy their engineers to other activities. Owning enables a buyer to deploy their own engineers to iterate on issues without coordinating with the provider. In practice, firms with technically challenging and unique use cases tend to favor owning many or all facets of their data centers and configuring them to their purposes. Some of these same firms—Microsoft, Amazon, and Alphabet—use their facilities to offer cloud services on a rental basis.

III. Data about Data Centers

We collected data in February 2018 and February 2019 to assemble a cross-sectional overview of the market-based supply of facilities. We scrape data from Baxtel.com and Datacenters.com, two of the largest information resource sites for third-party providers.[7] In all but a small number of situations (e.g., to supply services to the military), suppliers want to make their services known, so our data collection efforts result in a census of all third-party data centers between 2018 and 2019. The data source provides a snapshot of the industry at a point in time and provides limited data on entry and capacity over time.

We supplement the data with information about the facilities of the four largest cloud providers.[8] We provide information on 1,386 data centers from third-party providers. Adding the major cloud supplier centers, we get 1,433 data centers in February 2019. Figure 1 provides a map. A county can have a *Downtown*, *Suburban*, *Footloose*, or *Cloud* supply of data centers. The first three are defined by their distance from the nearest downtown area containing concentrations of businesses, who are the principal buyers of services from data centers. Within each MSA, we identify the downtown hubs to estimate the largest clusters of economic activity within the nearby area. Each MSA contains at least one downtown hub, although some MSAs have up to three. We geocode the longitude and latitude of each downtown and data center and calculate the distance between each data center and its nearest downtown area. A third-party data center is

[7] Our data sources (Baxtel.com and Datacenters.com) are two known sources of information on third-party data centers. Datacenters.com is the largest marketplace for colocation data center services (in transaction volume and revenue).
[8] These providers are Alphabet, Amazon, Facebook, and Microsoft.

"Downtown" if it is within 5 kilometers of the nearest downtown. "Suburban" centers are between 5 and 30 kilometers from downtown. "Footloose" centers are further than 30 kilometers. The "Cloud" data centers are owned by the cloud providers.

The map illustrates an urban bias. Almost all third-party data centers reside in Metropolitan Statistical Areas (MSAs), namely, large urban and suburban areas with large populations. Over 96.4% of counties with at least one third-party data center are part of an MSA.[9] The census classifies 37.2% of US counties as MSAs. More than 2,000 US rural and micropolitan[10] counties have no provider. Major cloud supplier data centers are less likely to locate in urban areas—73% of counties with at least one cloud supplier data center are MSAs (see appendix).[11]

Entry of third-party data centers results in geographically concentrated supply. It occurs in only 218 counties (i.e., 18.6% of MSAs). Only 76 counties have only one data center, and 34/23/17 counties have two/three/four data centers, respectively, which account for 281 data centers. There are more than 1,100 third-party data centers in the 68 counties that have five or more data centers. Twenty-six counties have 10 or more data centers, and 13 counties contain twenty or more. In other words, the vast majority of third-party data centers locate in one of a small number of the largest MSAs with many competitors, and the vast majority of small and medium-sized urban locations do not experience any entry at all.[12]

Because information about maximum electrical usage is not available for all data centers, we create a close proxy for capacity, termed *Gross SQFT*, that measures the total data center capacity in square feet. About 25% of the data centers provide the total square footage. We estimate the remainder based on images from Google Maps.[13] The aforementioned data provides clues about the order of magnitude of the

[9] The first calculation is the number of counties in MSAs over the total number of counties (1,173/3,150). The second calculates the number of counties in MSAs with at least one data center against all counties with at least one data center (218/26).
[10] Micropolitan areas have less than 50k population.
[11] Of the 26 counties with at least one cloud data center, 19 of them are in MSAs (73%).
[12] Santa Clara County (a.k.a. Silicon Valley) has the largest number, with 71, unweighted by size. It is followed by Los Angeles County, CA, Dallas County, TX, Cook County, IL, New York County, NY, Maricopa County, AZ, Fulton County, GA, Harris County, TX, Middlesex County, MA, Miami-Dade County, FL, and Loudoun County, VA (with 21).
[13] In almost all cases, square footage corresponds with total capacity. The rare exceptions are retrofitted buildings, such as the RRDonnelly building 350 E. Cermak in Chicago, which contains 1.1 million square feet. It is the largest third-party data center in North America. Other exceptions are data centers fit into dense urban settings with extremely expensive land, such as Manhattan or San Francisco, where we obtain total square footage from industry sources.

replacement value of these assets covered by third-party providers and cloud suppliers.[14] It takes a simple calculation to show that it must be on the order of magnitude in the hundreds of billions of dollars.[15] We will discuss other features of capacity later in the study.

We focus our statistical analysis on counties in MSAs, and we add a few additional counties with cloud supplier data centers. Removing most rural counties has no practical consequences for the estimation.

Explaining location. To test among different determinants of entry we estimate logits for data center entry in a county. We perform separate analyses on third-party and cloud supplier data centers, and our measure of entry is an indicator variable for whether there is at least one data center of that type operating in the county. With only industry beliefs as a guide, we opt to err on the side of testing many exogenous variables (without multicollinearity). Table 1 summarizes the definitions of variables.[16]

Local demand. The 2012 economic census provides the level of employment, by county, for two-digit NAICS industries. We focus on the most IT-intensive: Information, Healthcare, Manufacturing, Education, and Finance, Insurance, & Real Estate (FIRE). We calculate the percentage of employees in a county employed in that industry. We also include the percentage of the population with at least a bachelor's degree. While we interpret these as measures of demand, these measures may pick up elements of the supply of skilled labor.[17] Population and population density are highly correlated, so we cannot include both, and both are correlated with the prevalence of unobservable types of demand. We include population density, measured by the population per square mile. We also measure heterogeneity in population density within a county by creating measures for the lowest and highest population density areas by ZIP code within a county.[18]

---

[14] Replacement value is a standard concept for asset valuation, valuing an asset at reconstruction costs from scratch.
[15] For a conservative estimate, value the footloose and cloud data centers (average size: 450k sq ft and $465 sq ft) at $500m, value the suburban data centers (average size: 375k sq ft) at $400m, and value the downtown data centers (average size: 345k sq ft) at $800m. There are 103 downtown, 149 suburban, 47 footloose, and 26 cloud data centers. Hence, a conservative estimate is 103*800 + 149*400 + 47*500 + 26*500 = $183.5 billion. Reiterating, we do not offer that as a precise estimate of asset value but, rather, as an illustration of the order of magnitude.
[16] We listed these variables as part of demand, operational costs, and set-up costs, and we offer interpretations organized by category. In practice, a few variables lend themselves to interpretations as both a demand and cost factor (e.g., density).
[17] We are somewhat doubtful of the interpretation of supply based on industry interviews. Most third-party data centers employ small work forces, and shortages of inputs other than skilled labor receive more frequent mention.
[18] Population and density are also highly correlated in ZIP code data.

Operating costs. The Energy Information Administration provides 2016 state-level average electricity prices.[19] The CDC Wonder North America Land Data Assimilation System provides daily air temperatures from 2016, by county. We use average temperature during *Hot* and *Cold* months. Following studies of software (Bennett and Hall, 2020), we obtain the median hourly wage of network engineers in 2016 from the Bureau of Labor Statistics (BLS). For annual property taxes, we use state-level real estate tax (RET) rates from the National Association of Home Builders. Given the coarser level of this data, we create indicators for whether a county is located in a state in the top (bottom) quartile of RET: *High (Low) Real Estate Tax*.

Set up costs. We approximate land costs by collecting county-level data on the median home value by square foot from Zillow. To account for land cost variance within a county, we create measures for the lowest and highest median home value areas by ZIP code within the county. We obtain MSA-level sales tax (RET) rates in 2010 from the Tax Foundation to consider taxes on materials and IT hardware involved in building a data center. We use this data to create indicators for whether a county is in the top (bottom) quartile of sales tax: *High (Low) Sales Tax*. We also consider the median hourly wage of construction laborers in each county in 2016 from the BLS.[20] Finally, we use data from the CBRE on tax incentives for the 30 largest enterprise markets. We create *Tax Break MSA* for whether a county is based in an MSA that provided tax breaks to data centers in 2015, the most recent available year. Though listed as a set-up cost, it applies to many operational costs, too.[21]

Summary statistics of variables. Table 2 shows summary statistics. As noted earlier, only 18.6% of counties have at least one data center. When we examine the types of data centers, we observe several interesting patterns. On average, there is more suburban capacity than any other type of capacity, followed by urban and then footloose capacity. The amount of suburban capacity in a county is almost three times higher than the footloose capacity (47,622 SQFT vs. 17,869 SQFT). We also observe that our capacity measurements appear to have a skewed distribution. Similarly, we observe that our density variables appear to have a skewed

[19] We also test the prices to commercial customers. Our results do not change when we use alternative measures.
[20] For counties that did not provide an estimate of the hourly wage of network engineers or construction laborers, we used averages from each MSA or each state.
[21] Tax abatement and incentive programs can apply to property, sales, and personal taxes. CBRE (2013).

distribution. We apply a log-transformation for both exogenous and endogenous variables in our analysis. We also calculate correlations for the exogenous variables (see appendix).[22]

Descriptive patterns. Figure 2a shows a scatter plot of the number of data centers and the logged population of each county. There is a positive correlation of 0.42. All locations with more than 700,000 people have at least one third-party provider. There is no entry of third-party data centers below a population level corresponding to approximately 160,000 people. Between these two population levels, the experience is mixed.

Figure 2b shows the relationship between logged capacity and logged population. The figure includes only counties with at least one data center. Third-party data center capacity and population display a positive correlation of 0.596. Capacity is not synonymous with the number of entrants, however. The county with the largest capacity, Loudon, VA, ranks as the county with the tenth-largest number of entrants, for example. The capacity in Loudon partly arises due to demand among users in the greater Washington, DC area, which contains a large concentration of data-intensive work. This location also contains several data centers affiliated with AWS. Altogether this location contains features consistent with agglomeration economies due to demand for connectivity, as in a transportation hub (Alcácer and Chung, 2014).[23] Notably, we do not observe such agglomeration at a similar scale at any other location. [24]

Among cloud providers we do observe some geographic concentration of buildings. Facebook has located a significant amount of capacity (i.e., eight buildings) in Prineville, OR. While AWS has located many buildings in Ashburn, VA., and Microsoft in Quincy, WA, neither has located as much of its infrastructure in one place. Google does not display a similar tendency and has large facilities in several locations.[25]

[22] Median household income was not included in the analysis due to potential multicollinearity.
[23] Known as the "data center capital" and "interconnection capital" of the US, it contains excellent connectivity between proximate data centers, between the facilities of cloud providers, and between data centers and high-speed internet lines leading to other locations.
[24] That location was home to MAE-East, one of the earliest data-exchange points on the commercial internet, and today contains a number of Amazon's data centers. As it turns out, the firms that locate in Loudon operate some of the very largest data centers in the US. This location is followed in rank (of capacity) by Dallas, TX, Santa Clara, CA, Cook, IL, Los Angeles, CA, Maricopa, AZ, Middlesex, NJ, Harris, TX, King, WA, NY, NY, and Fulton, GA.
[25] We should note that these observations pertain only to facilities in North America. All providers maintain facilities in multiple locations outside North America, but we have no insight into the strategies for their locations. It is a topic for further study.

Table 3 asks whether the prevalence of third-party vendors with specialized services rises in the largest markets. The table shows the entry of data centers that expended extra costs to comply with two industry-specific standards, comparing markets with more or fewer than 750k population, labeling them as major and minor. Some data centers protect personal health information in compliance with HIPAA (Health Insurance Portability and Accountability Act of 1996) and HITECH (Health Information Technology for Economic and Clinical Health Act of 2009). Some protect financial data in compliance with Service Organizational Controls (SOC), namely, SOC 1 and SOC 2.[26]

The table shows that major markets attract more suppliers who meet these specialized needs than minor markets. These results are consistent with the hypothesis that data centers are responsive to local demand when entering a market. The major markets tend to host more finance and medical firms, and in such markets there tend to be more of them. There are 3.78 (1.05) specialists on average in markets with over (under) 750K population, and more than 75% (43%) of large (small) markets provide at least one data center that has adopted a specialized standard. The data centers in major markets find it worthwhile to incur the expenses required to meet these demands from local firms.[27] We also find an urban bias in the location of specialized data centers.

IV. Estimation results

Table 4 presents the estimates on the entry of data centers. All estimates focus on the threshold between zero and one. Column 1 presents a logit model where the endogenous variable is an indicator variable for whether a county has a third-party entrant. Column 2 shows a logit model where the endogenous variable is an indicator variable for whether a county has a cloud entrant. All logit estimates are displayed as log-odds ratios. Experiments with alternative specifications yield qualitatively similar results (see appendix). The third column of Table 4 uses the estimates of Column 1 and presents the change in probability of third-party entry due to a one standard deviation increase in the exogenous variable, simulated at the mean values for all

[26] SOC1 is concerned primarily with internal financial controls related to financial reporting, and SOC2 is concerned with IT security, confidentiality, privacy, processing integrity, and availability. Source: https://www.ssae-16.com/soc-1/.
[27] Consistent with the hypothesis that urban and suburban data centers are responsible to local demand, these findings are more pronounced when the footloose data centers are excluded from the sample.

exogenous variables. With dummy variables, it displays the change in probability from turning it on. The fourth column provides a similar simulation for the estimates in Column 2.

First, we discuss third-party entry strategies. Column 1 shows that some local demand factors lead to more entry. The proportion of Information and FIRE employees predicts entry and is statistically significant. Column 3 shows that a one standard deviation increase in the supply of information and FIRE workers is associated with, respectively, a 1.9% and 3.1% increase in the likelihood of data center entry into a county. Their similar direction and magnitude are consistent with the interpretation that they measure local demand. None of the other types of users matter, including manufacturing or health workers. Column 3 shows that having a population density one standard deviation above the mean level is associated with a 7.4% increase in the likelihood of entry into a county. There is a negative association between the size of the minimum population density in a county and the likelihood of entry. The former and latter results are consistent with the interpretation that data centers select counties that are highly dense (and part of populous MSAs) and then locate within a county with the lowest density (e.g., parcel of large available land).

Column 1 shows that an increase in median home values is associated with an increase in the likelihood of entry. This goes in the opposite direction of predictions and likely reflects unobserved demand affiliated with dense locations with high income, which correlates with the higher value for homes. However, consistent with the finding about minimum density, we find that data center entry decreases as the minimum land value increases. Column 3 shows that having a minimum land value one standard deviation above the mean level is associated with a 3.1% decrease in the likelihood of entry into a county. Most of the other estimates for variable and fixed costs are statistically insignificant except for hot weather, which has a small discouraging effect.

We compare the entry patterns between third-party and cloud provision in Columns 1 and 2. None of the measures of local demand in Column 2 play an important role in predicting cloud entry, except local information industries, and its quantitative effect is much smaller. The proportion of FIRE employees does not statistically predict cloud entry. There is also a small negative association between the likelihood of entry and the percent of the population with Bachelor's degrees and the minimal density, but its importance is small. We

are unable to distinguish between an interpretation of local industry that favors demand or supply of labor, but in both cases, we conclude that the comparative role of proximity is much smaller for cloud suppliers.

For the variable and fixed costs, we see several comparatively different outcomes, i.e., that higher electric prices discourage cloud entry, and lower median construction wages encourage entry from major cloud providers. The same is true for the minimum land value. Although none of these are large, they differ markedly from the estimates for third-party data centers, providing evidence that cloud supplier entry strategy is sensitive to different cost factors.

For third-party data centers, we simulated the difference in fixed costs due to observable factors (see appendix). We compared markets with a single entrant with those with more than five entrants. The estimates say the fixed costs are 24% higher in the competitive locations due to observable factors. We infer it is likely that all providers in downtown and suburban locations get a benefit from the more competitive locations that make up for the higher setup costs. In other words, local demand considerations have a more central role than operation considerations in covering the fixed costs of entry of data centers in a dense urban location.

Limits to inference

The inferences discussed arise from estimating the entry at the margin of entry or no entry. That leaves unaddressed whether further inferences are possible from examining the extent of entry. We explored this possibility and found little else—as a statistical matter—beyond the obvious relationship between more population, and more entrants. In other words, once other factors are controlled for, approximately the same incremental increases in population levels determine the second, third, and fourth entrants (see appendix).[28]

Statistical analysis of capacity does not yield much additional insight into firm strategies. We perform a Heckman correction to predict capacity after conditioning on entry. The demand-side effects for the Information and FIRE industries do not predict demand, once we control for selection of any entry (see appendix). We do find that population density is associated with a positive and significant increase in urban and suburban capacity. A 1% increase in population density is associated with a 1.059% increase in urban capacity

[28] Due to the high correlation between population levels and population density for large cities, the same incremental increases in population density also determine the second, third, and fourth entrants.

and a 1.416% increase in suburban capacity. We find little evidence that footloose capacity responds to lower variable costs except in one area: as expected, lower industrial electric prices are associated with higher footloose capacity. Overall, only a few coefficients predict capacity, once we control for entry, so only weak conclusions emerge. Though we see some differences across the coefficients for local and footloose equations, the differences are not dramatic except for electricity. We conclude that entry behavior is more informative about local demand and supply than quantity demanded as measured by capacity.

The biggest outlier to the relationship between population and capacity occurs in Ashburn, Virginia, the location in the US with the highest capacity. Industry consensus believes the entry of firms and their growth are driven by demand for connectivity with others. Ashburn benefits from its unique history as the location for the first public data exchange point in the NSFNET, chosen, in part, due to a propitious location just outside the large number of users in the Washington, DC MSA, and on lines that carry traffic along with the cities of the mid-Atlantic, Southeast, and Northeast. Consistent with this view, it is also the location with the largest collection of AWS data centers. But one observation does not make a statistical relationship, and we were unable to identify any other location that industry consensus had identified in similar terms.

We also tested for differences in the behavior of urban and suburban firms and found no substantial differences. We also tested for differences between cloud data centers and footloose third-party data centers. We found no statistical differences between them, but this finding was underpowered due to a small number of observations. While this suggests differentiation between firms shapes the geographic patterns of entry, this hypothesis requires considerably more analysis than this first analysis can provide. It is a topic for future work.

V. Conclusion

The most striking finding is differences in the behavior of cloud and third-party data centers. Statistical evidence suggests different degrees of responsiveness to features of local demand and costs. For example, cloud supplier data centers are comparatively more responsive to costs such as electricity and construction wages, and comparatively less responsive to demand associated with high population density.

Our findings are consistent with a role for localized demand as an explanation for variance in entry strategy at third-party data centers. All large metropolitan areas (with over 700,000 population) have at least one local third-party supplier of data services, only a fraction of small and medium-sized population areas have local suppliers of third-party data services, and no city below a certain size has any data centers (approximately a little more than 160,000 population). In between these two population levels, some areas have entrants but most do not. These patterns are consistent with the tension between economies of scale and localized demand across US cities. A high fraction of small and medium-sized locations must get their services from a non-local supplier (likely located in the closest major city).

The overall patterns point toward a pervasive urban bias in the location of third-party data centers. Our analysis suggests third-party data centers are more prevalent in locations with local users who demand their services, and we find evidence that third-party entrants select areas with lower costs. For example, local entry rises with the presence of information-intensive industries and FIRE, consistent with the presence of localized demand. We also find evidence that specialized providers in the market become more prevalent in larger markets. Once again, users in small and medium-sized locations have fewer nearby choices among specialized suppliers, if they have any at all.

There is scant evidence during this pre-pandemic era that data centers from cloud providers spread to any but a countably small number of non-urban locations. At most, our finding suggests cloud data centers search for low electricity prices and lower construction costs next to good access to internet lines, and that leads them to make different location choices than third-party data centers.

If proximity begins to be important for cloud providers, we expect cloud providers with hyperscale data centers in less urban locations will adopt architectures with points of presence close to customers. This is sometimes called "distributed architectures." To achieve a distributed architecture, cloud providers would either have to build their own centers in many locations, acquire third-party data centers in many locations, or negotiate colocation agreements in third-party data centers. As demand spreads to the cloud, we will see which options are chosen. Given the existing locations of third-party data centers, we might expect many

colocation agreements to be the easiest way to become close to potential customers to relieve concerns about network congestion.

Strategic implications

This study offers four implications. First, this study quantifies how data center managers trade-off between the setup and operational costs of running a facility and capturing local demand. While supplier proximity to users who demand data center services alleviates a buyer's "distaste for distance," these markets are also associated with higher setup costs. The tradeoff between proximity and setup cost appears fundamental and irreducible. As long as a sufficiently large number of users are willing to pay a premium for close proximity, suppliers will incur the extra setup costs to locate nearby.

Second, this study also offers evidence that data center firms that provide specialized services display an urban bias. When there is enough local demand from potential buyers, this provides more opportunities to provide differentiated services within a data center in proximity to users. In this respect, specialized data centers resemble plenty of other specialized services in information technology, whose availability is more prevalent in dense urban locations.

Third, the differences between the location of major cloud suppliers and third-party firms suggest the cloud providers will encounter challenges in satisfying demand for local suppliers. We would expect this demand to create demand for distributed cloud architectures. These are architectures where multiple clouds meet user needs and requirements. They tend to have facilities built to support limited computing "at the edges" with strong connectivity back to the central cloud provider.[29]

Finally, from a buyer perspective, we expect that firms that operate businesses in small and medium cities will be most affected by the geographic choices of vendors. They either must build for their own use, and not gain the benefits of scale economies from sharing infrastructure with others, or find supply from distant locations and give up the benefits of close proximity. Deeper research can address some of the tension. For example, though managers prefer a local supply of infrastructure when it is available, it may be

[29] https://www.gartner.com/smarterwithgartner/the-cios-guide-to-distributed-cloud.

possible to use remote data centers or cloud storage to manage the potential latency issues associated with congestion on data lines and to cater to specialized needs.

Further research

This research speaks to two distinct views that animate open policy questions about digital infrastructure supply. An outlook that could be labeled as "optimistic" anticipates experimentation in a few places, followed by more diffusion to more users, more regions, and a larger set of applications. It interprets the state of digital infrastructure at a point in time as temporary and transient, and in the midst of wider diffusion. In contrast, an outlook that might be labeled as "pessimistic" stresses that digital infrastructure has achieved higher productivity in dense locations. That arises due to economies of scale in equipment, due to increased productivity from the colocation of many related activities, and due to the availability of skilled labor in urban areas in developed economies. Overall, the experience with third-party data centers supports the less optimistic view, due to the concentration of supply around urban cities, and the persistent demand for local supply. The shift to cloud suppliers does not alleviate these concerns. One major open question is whether such patterns will persist after the providers adjust to the post-pandemic demands for their services.

We undertook this study in the absence of other statistical research on the strategies of these suppliers. This study brings attention to the broad topic. But it focused on only one aspect of the strategies pursued by firms. Many topics remain uncovered. Additional studies could focus on the build and acquisition strategies of the largest third-party providers. There is room for careful study of the strategies of firms who specialize in meeting specific industry needs in, for example, finance and health care. Enough time also has passed to observe the consequences of the building of cloud infrastructure in less dense locations, and evaluate the consequences for the locations that subsidized their entry. Finally, many of the cloud providers have undertaken significant investments outside the US. All the strategic issues behind those investments are open questions.

References

Alcácer J, Chung W (2014) Location Strategies for Agglomeration Economies. *Strateg. Manag. J.* 35(12):1749–1761.

Bennett VM, Hall TA (2020) Software Availability and Entry. *Strateg. Manag. J.* 41(5):950–962.

Böttger T, Cuadrado F, Tyson G, Castro I, Uhlig S (2016) Open Connect Everywhere: A Glimpse at the Internet Ecosystem through the Lens of the Netflix CDN. *CoRR*.

Bresnahan TF, Gambardella A (1997) The Division of Labor and the Extent of the Market. *Work. Pap.*:1–34.

Byrne, D, Corrado, C, & Sichel, DE (2018) The Rise of Cloud Computing: Minding Your P's, Q's and K's. National Bureau of Economic Research Working Paper Series, No. 25188.

Cardona M, Kretschmer T, Strobel T (2013) ICT and Productivity: Conclusions from the Empirical Literature. *Inf. Econ. Policy* 25(3):109–125.

CBRE (2013) Impact of Taxes & Incentives on Data Center Locations. https://f.tlcollect.com/ fr2/813/17870/Impact_of_Taxes_and_Incentives_on_Data_Center_Locations_(2013).pdf, accessed January 2022.

CBRE (2015) Site Selection Strategies for Enterprise Data Centers. CBRE Research, https://www.cbre.us/research-and-reports, accessed January 2022.

Ceccagnoli M, Forman C, Huang P, Wu DJ (2012) Co-creation of Value in a Platform Ecosystem: The Case of Enterprise Software. *MIS Q.* 36(1):263–290.

Coyle D, Nguyen D (2018) Cloud Computing and National Accounting. *ESCoE Discuss. Pap. 2018-19* (December):1–59.

Delgado M (2012), Porter, ME, Stern S (2010) Clusters and entrepreneurship. *Journal of Economic Geography* 10(4): 495-518.

DeStefano T, Kneller R, Timmis J (2018) Broadband infrastructure, ICT Use and Firm Performance: Evidence for UK firms. *J. Econ. Behav. Organ.* 155:110–139.

DeStefano T, Kneller R, Timmis J (2019) Cloud Computing and Firm Growth. *Nottingham Cent. Res. GEP 20/02*: 1-68.

Forman C, Ghose A, Goldfarb A (2009) Competition between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live. *Manage. Sci.* 55(1):47–57.

Forman C, Ghose A, Wiesenfeld B (2008) Examining the Relationship between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets. *Inf. Syst. Res.* 19(3):291–313.

Forman C, Goldfarb A, and Greenstein S (2018) How Geography Shapes—and Is Shaped by—the Internet. in (ed.) Gordon Clark, MaryAnn Feldman, Meric Gertler, and Dariusz Wojcik, *The New Oxford Handbook of Economic Geography*, Oxford University Press. Pp. 269–285.

Goldfarb A, Tucker C (2019) Digital Economics. *J. Econ. Lit.* 57(1):3–43.

Greenstein S. (2020) Digital Infrastructure. in (ed.) Edward Glaeser and James Poterba, *Economics of Infrastructure Investment*. NBER Book. University of Chicago Press.

Jin W, McElheran K (2019) Economies Before Scale: Survival and Performance of Young Plants in the Age of Cloud Computing. *SSRN Work. Pap.*:1–77.

OECD (2014) *Measuring the Digital Economy, A New Perspective*. OECD Publishing, Paris. https://doi.org/10.1787/9789264221796-en.

Pew Research Center (2019) Mobile Fact Sheet. June 12, 2019. https://www.pewresearch.org/ internet/fact-sheet/mobile/, accessed January 2022.

Tambe P (2014) Big Data Investment, Skills, and Firm Value. *Manage. Sci.* 60(6):1452–1469.

Wang S, LaRiviere J, Kannan A (2019) Spatial Competition and Missing Data: An Application to Cloud Computing. *Work. Pap.*:1–23.

Xiao M, Orazem PF (2011) Does the Fourth Entrant Make any Difference?: Entry and Competition in the Early U.S. Broadband Market. *Int. J. Ind. Organ.* 29(5):547–561.

Zhuo R, Huffaker B, Claffy KC, Greenstein S (2020) The Impact of the General Data Protection Regulation on Internet Interconnection. *NBER Work. Pap.* No. 26481.

Table 1: Definition of Demand & Supply Variables.

Variable	Definition
Dependent Variables	
Third-Party Providers	Indicator for county with at least one third-party data center provider in the county.
Cloud Providers	Indicator for county with at least one cloud data center provider in the county.
Demand Shifters	
% Bachelor's Degree	% of population with at least a bachelor's degree.
% Information	% of county's employees in Information sector (NAICS: 51).
% Health	% of county's employees in Health sector (NAICS: 62).
% Education	% of county's employees in Education sector (NAICS: 61).
% FIRE	% of county's employees in FIRE.
% Manufacturing	% of county's employees in Manufacturing sector (NAICS: 32–33).
Population Density	Population density (pop./square mile).
Min. Density	Lowest population density ZIP code in a county.
Max. Density	Highest population density ZIP code in a county.
Variable Cost Shifters	
Industrial Electric Price	State-level average price of electricity to industrial customers (cents/kilowatt-hour) in 2016.
Low Real Estate Tax	Indicator for a county that has a RET (state-level) in the bottom quartile.
High Real Estate Tax	Indicator for a county that has a RET (state-level) in the top quartile.
Median Network Engineer Wage	2016 median hourly wage of network engineers in the county.
Cold	Average temperature during three-month period of December 2015, January, February 2016.
Hot	Average temperature during three-month period of June, July, August 2016.
Fixed Cost Shifters	
Median Construction Wage	2016 median hourly wage of construction laborers in the county.
Tax Break MSA	County in MSA that gave out tax breaks to data centers (source: CBRE).
Low Sales Tax	Indicator for a county that is an MSA with a 2016 sales tax rate in the bottom quartile.
High Sales Tax	Indicator for a county that is an MSA with a 2016 sales tax rate in the top quartile.
Median Home Val. (per SQFT)	Estimated median home value/SQFT in a county.
Min. Median Home Val.	Minimum home value/SQFT ZIP code in a county.
Max. Median Home Val.	Maximum home value/SQFT ZIP code in a county.

Table 2: Summary Statistics of Demand & Supply Variables.

	mean	s.d.	min	max
Dependent Variables				
Third-Party Providers	0.186	0.389	0	1
Cloud Providers	0.022	0.147	0	1
Demand Shifters				
% Bachelor's Degree	17.371	7.108	4.854	55.464
% Information	1.085	1.193	0	15.066
% Health	10.126	4.373	0	44.326
% Education	0.210	0.253	0	3.008
% FIRE	3.180	2.198	0	29.758
% Manufacturing	8.241	6.489	0	40.011
Population Density	644.237	2912.159	0.8	72158
Min. Density	23.057	163.762	0	2793.67
Max. Density	4342.058	10916.470	0.814	172373.4
Variable Cost Shifters				
Industrial Electric Price	6.978	2.011	4.66	22.92
Low Real Estate Tax	0.251	0.434	0	1
High Real Estate Tax	0.225	0.418	0	1
Median Network Engineer Wage	30.440	4.466	18.25	46.87
Cold	29.247	8.120	7.163	55.975
Hot	65.710	7.322	38.188	79.25
Fixed Cost Shifters				
Median Construction Wage	14.52	3.98	8.21	29.38
Tax Break MSA	0.163	0.369	0	1
Low Sales Tax	0.255	0.436	0	1
High Sales Tax	0.211	0.408	0	1
Median Home Val. (per SQFT)	117.638	74.574	36	590
Min. Median Home Val.	84.444	43.580	18	437
Max. Median Home Val.	139.142	113.642	24	1142
Counties	1173			

Table 3: **Specialization Analysis.** The table shows the mean number of data centers within a county that are compliant with various standards and the proportion of counties with at least one specialist data center. HIPAA and HITECH are relevant to organizations that handle personal health information in any capacity. SOC 1 and SOC 2 are a set of standards relevant to demonstrate that an organization has the appropriate controls in place to protect and account for financial data. A major market is defined as a county with a population greater than 750,000.

	Major Market				Minor Market			
	HIPAA	HITECH	SOC 1	SOC 2	HIPAA	HITECH	SOC 1	SOC 2
Mean Data Centers	3.684	0.776	1.605	2.184	1.048	0.267	0.381	0.571
Counties with Specialist (%)	76.32	39.47	51.32	61.84	42.86	17.14	19.05	26.67

Table 4: Entry Analysis of Data Centers by Type. Columns 2.1 and 2.2 are logit models and the coefficients are in odds-ratios. The dependent variable is an indicator for data center entry of a particular type. Columns 1.3 and 1.4 show the change in probability of entry from a one standard deviation increase, calculated at the mean values for all exogenous variables.

	(2.1) Third-Party Providers	(2.2) Cloud Providers	(2.3) Simulated Gains Third-Party Entry	(2.4) Simulated Gains Cloud Entry
Demand Shifters				
Percent Bachelor's Degree	1.024	0.895*	0.009	-0.002
	(0.026)	(0.045)		
Percent Information	1.314**	1.860***	0.019	0.005
	(0.136)	(0.303)		
Percent Health	0.940	0.958	-0.013	-0.001
	(0.033)	(0.067)		
Percent Education	1.038	0.681	0.000	0.000
	(0.652)	(0.966)		
Percent Finance, Insurance,	1.248***	1.135	0.031	0.002
& Real Estate	(0.069)	(0.085)		
Percent Manufacturing	0.953	1.02	-0.014	0.001
	(0.027)	(0.042)		
Log Population Density	1.887***	1.462	0.074	0.003
	(0.284)	(0.429)		
Log Minimum Population Density	0.805*	0.460*	-0.013	-0.002
	(0.069)	(0.147)		
Log Maximum Population Density	1.750***	0.974	0.091	0.000
	(0.202)	(0.214)		
Variable Cost Shifters				
Industrial Electric Prices	1.071	0.402*	0.007	-0.003
	(0.061)	(0.155)		
Low Real Estate Tax	0.696	0.253	-0.018	-0.004
	(0.217)	(0.196)		
High Real Estate Tax	0.738	1.393	-0.014	0.001
	(0.208)	(0.871)		
Median Network Engineer Wage	0.939	1.142	-0.013	0.004
	(0.034)	(0.092)		
Cold	1.035	0.946	0.016	-0.001
	(0.019)	(0.047)		
Hot	0.952*	0.959	-0.016	-0.001
	(0.019)	(0.038)		
Fixed Cost Shifters				
Median Construction Wage	1.046	0.804*	0.010	-0.002
	(0.040)	(0.087)		
Tax Break MSA	2.227*	1.552	0.054	0.002
	(0.719)	(0.941)		
Low Sales Tax	0.686	0.755	-0.018	-0.002
	(0.222)	(0.49)		
High Sales Tax	1.420	0.936	0.014	0.000
	(0.369)	(0.625)		
Median Home Value	1.006*	1.001	0.03	0.000
(per SQFT)	(0.003)	(0.007)		
Min. Home Value	0.981***	1.018*	-0.031	0.005
(per SQFT)	(0.004)	(0.009)		
Max. Home Value	1.000	0.995	-0.002	-0.002
(per SQFT)	(0.001)	(0.004)		
Observations	1173	1173	1173	1173

Notes: Columns 1.1 and 1.2 show exponentiated coefficients; standard errors in parentheses. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1

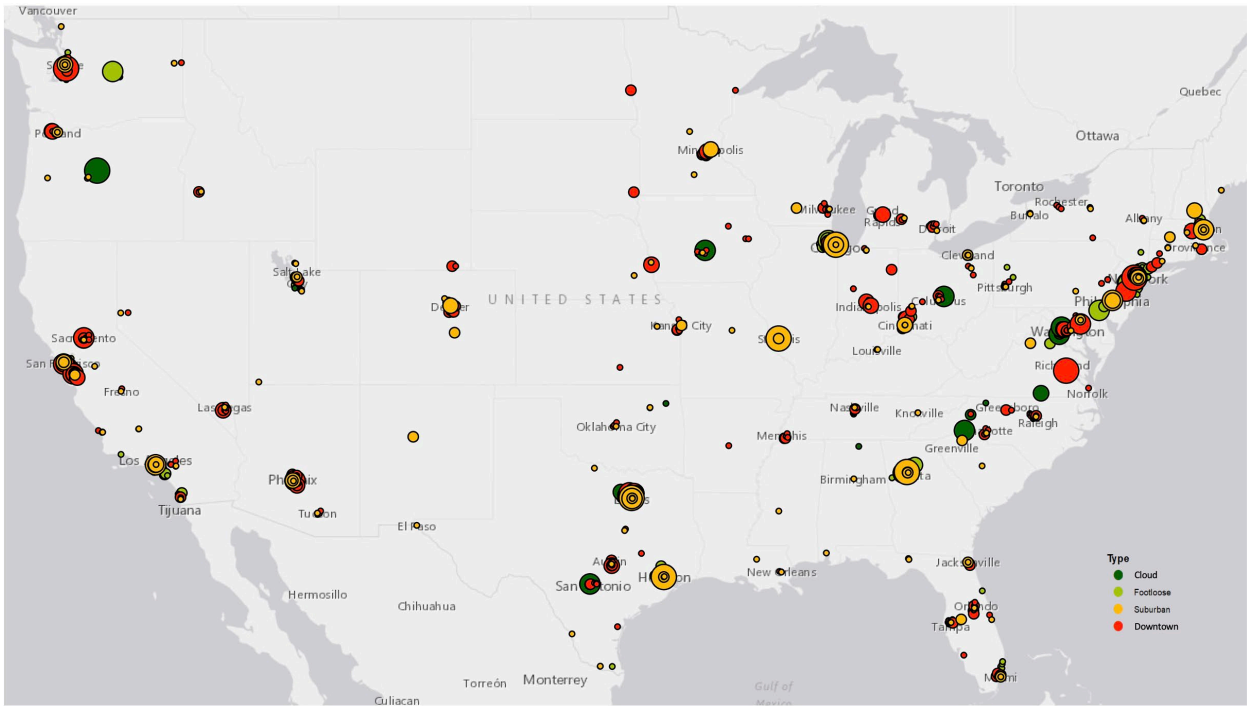

Figure 1: Geography of Data Centers. This figure maps out the geographic locations of data centers. Each point represents a data center and is color-coded by type and weighted by the total capacity of the facility (in SQFT).

Figure 2

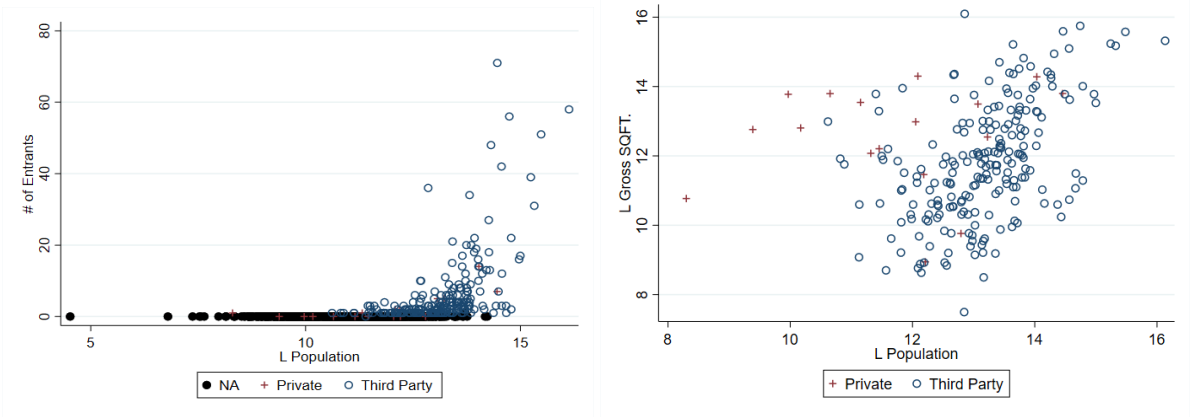

Figure 2: Population vs. Entry and Capacity. The first graph demonstrates the relationship between county population and the number of data center entrants within the county. A county is assigned a type based on what type of data center has the largest share of capacity within the county. The second graph demonstrates the relationship between county population and the capacity within the county, conditional on some data center entry. Once again, a county is assigned a type.