informs
https://pubsonline.informs.org/journal/stsc

STRATEGY SCIENCE
Vol. 10, No. 4, December 2025, pp. 404–420
ISSN 2333-2050 (print), ISSN 2333-2077 (online)

Where the Cloud Rests: The Economic Geography of Data Centers

Tommy Pan Fang,[a,*] Shane Greenstein[b]

[a] Jones Graduate School of Business, Rice University, Houston, Texas 77005; [b] Harvard Business School, Boston, Massachusetts 02163
*Corresponding author
Contact: tpanfang@rice.edu, https://orcid.org/0000-0003-2332-7121 (TPF); sgreenstein@hbs.edu (SG)

Received: May 1, 2024
Revised: December 17, 2024; June 14, 2025
Accepted: August 5, 2025
Published Online in Articles in Advance: October 14, 2025

https://doi.org/10.1287/stsc.2024.0225

Abstract. This study provides an analysis of the entry strategies of third-party data centers in the United States. We examine the market prior to the pandemic, in 2018 and 2019, when supply and demand for data services were relatively stable geographically. We compare these patterns 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 for this digital infrastructure. The former favors urban settings more and appears 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 tend to display a lower propensity for locating in urban areas, concentrating their buildings 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.

Funding: The authors are grateful to Rice University and the HBS Division of Research and Faculty Development for financial support.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/stsc.2024.0225.

Keywords: technology strategy • entry strategy • economic geography • data centers • spatial competition

1. Introduction

Did digitization change the strategic role of proximity? Proximity confers a competitive advantage through a range of different mechanisms (Alcácer and Chung 2014). It lowers transportation costs, enables easier access to skilled labor and innovative complementors, and assists firms in enjoying the benefits they share with others. The increasing digitization of goods and services potentially alters these mechanisms. Digitization lowers the frictions associated with moving information (Goldfarb and Tucker 2019). It enables the creation of services with high "shippability," which means they can be delivered to and consumed from any connected location (Gervais and Jensen 2019), and it lowers barriers to trading with distant customers with similar tastes (e.g., Sinai and Waldfogel 2004, Blum and Goldfarb 2006).

Data centers are an ideal setting for studying how digitization shapes proximity. The services provided by a data center play an essential role in the digital services used and managed by virtually every firm with a large-scale digital presence. On the supply side, data centers display economies of scale, but only after the expenditure of extraordinarily high fixed costs for structures that occupy outsized plots of land (Cassiman et al. 2022). If data centers are cost-focused, we would expect them to enter locations with less competition and where the fixed and operational costs associated with the data center facility are low. While they may benefit from locating adjacent to pools of network engineers, operating these facilities does not require large numbers of employees. Such a minimal labor requirement provides considerable flexibility in determining their location and would discourage entry into dense urban areas. On the demand side, the services of data centers travel at the speed of light, which makes them highly shippable and accessible without being geographically proximate.

However, buyers for data center services may exhibit a distaste for distance and prefer services from nearby suppliers instead of those who are further away. Such distaste could arise from a mix of factors discussed in the text. If distance does matter, it should result in data centers in expensive locations to be near customers, with supply and demand determining where and how much. In light of these factors, we ask the question: How does proximity to local demand influence the entry strategies that determine the location of U.S. data centers?

To answer this question, we collect and analyze a novel data set detailing the location of U.S. data centers in 2018 and 2019, a period of relative geographic stability before the pandemic. For our purposes, the prepandemic era enables us to statistically track the determinants of local demand and local supply, providing insight into the underlying factors that influence the location of data centers. Immediately following this period, the COVID-19 pandemic led to a shift in data traffic patterns, with many people working from home (Choudhury et al. 2021, Bloom et al. 2024).

Our data provides a census of third-party data centers and includes the facilities of the largest private cloud providers. We compare the entry strategies of "third-party" data centers that lease resources to multiple firms with "cloud" data centers owned and operated by a single major provider. Using the county as our unit of analysis, our primary estimation strategy employs a logit model to analyze the likelihood of data center entry. We test how entry decisions are influenced by a range of county-level variables, including measures of local demand (such as the level of employment within the Information and Finance, Insurance, & Real Estate sectors), operating costs (like industrial electricity prices), and setup costs (such as construction wages and land values).

Our findings reveal that third-party firms and cloud providers pursue different location strategies. We find a pervasive urban bias in the location of third-party data centers, as they favor dense metropolitan areas. Data center entry into a county is significantly predicted by local demand from information-intensive industries, as they trade off the costs of supply against the benefits of proximity to their buyers for each potential site. All large U.S. metropolitan areas with over 700,000 people host at least one third-party supplier, while less dense areas may or may not have any entry. Thus, a high fraction of the buyers in small and medium-sized locations must get their services from nonlocal suppliers. Relatedly, we also find a supply of more specialty services in denser and more competitive locations. Ultimately, a higher-quality network appears to play a modest yet significant role in explaining regional disparities in data center locations. We interpret all these patterns as the result of tension between producers' desires to realize economies of scale and users' preference for proximity to their suppliers. In sharp contrast, major cloud providers exhibit a lower propensity to locate in urban areas and are comparatively more sensitive to costs, such as electricity prices and construction wages. Cloud providers also tend to concentrate their facilities in a small number of locations, displaying less responsiveness to geographically dispersed local demand. Because our study focuses on an early moment in the shift toward providing cloud data services, we interpret these differences as explaining where infrastructure is currently located and identifying a trend in where it will be located in the future.

This study is the first to statistically analyze the determinants of data center location strategies, making several contributions to the literature on technology strategy and economic geography. We show that for data centers, the strategic role of proximity has not been eliminated by digitization but has instead bifurcated. Our findings reveal that third-party data centers confirm prior work on "urban bias" where proximity to urban demand remains paramount (e.g., Forman et al. 2009, Greenstein et al. 2018). Yet, the supply-side logic of cloud providers implies that the role of proximity is conditional on a firm's business model, extending recent work on this topic (Wang et al. 2019). Our analysis also highlights a significant departure from the clustering patterns common in other technology sectors. In contrast to agglomeration theories that emphasize the pooling of skilled labor (e.g., Alcácer and Chung 2014), we find that data center location is largely decoupled from specialized human capital. This implies that for industries with separable physical and digital assets, the drivers of clustering may shift from labor dynamics to unique infrastructure characteristics. Theoretically, our work also extends the literature on infrastructure strategy (Bresnahan and Gambardella 1998). We demonstrate how upstream suppliers must navigate a spatial tradeoff between achieving economies of scale and locating near downstream users to spread high fixed costs. Finally, these findings inform critical debates in both managerial practice, by clarifying the strategic tradeoffs of a key component of digital infrastructure, and public policy, by providing an evidence base to evaluate policies such as tax incentives (CBRE 2013, 2015) and rural broadband subsidies (Rosston and Wallsten 2022).

Section 2 briefly outlines the determinants of data center locations just before the pandemic. Section 3 describes the data analyzed in the study and the estimation strategy. Section 4 presents the estimation results, Section 5 provides additional extensions, Section 6 draws several strategic and policy implications for this infrastructure, and Section 7 concludes.

2. Determinants of the Location of Data Centers

2.1. Entry Decisions for Data Centers

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

Despite their importance and recent growth, it appears that no research has investigated the strategies that shape the location decisions of these 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 fundamental tensions determining the locations of various types of data centers? The study treats location as endogenous, determined by the demand for a local facility and the costs of a location. We find evidence suggesting that the third-party data center industry balances the trade-off between buyer demand for proximity and the reduced costs of operating in less expensive locations.

2.1.1. Ownership. Third-party suppliers offer users options to "colocate" in a shared building. Collocating enables the buyer to use innovations offered by the supplier and redeploy their engineers to other activities. In contrast, owning a private data center enables buyers to deploy their own engineers to troubleshoot and resolve issues without coordinating with the provider. Some large firms today own and manage their own centers for their company's purposes. Most visibly, firms with large and technically challenging computing needs, such as Facebook, Apple, Microsoft, Amazon, and Google/Alphabet, own all facets of their data centers and configure the hardware and software to their purposes. Some of these same firms—Microsoft, Amazon, and Alphabet—utilize their facilities to offer cloud services to users on a rental basis, allowing users to turn the service on and off at will.

2.1.2. Local Demand and the Preference for Proximity. Several factors influence user demand for proximity. Proximity plays a crucial role in reducing latency, which users perceive as a faster response time. This motivates suppliers to locate nearer to some users. Proximity also encourages the selection of locations close to high-speed data lines that can reliably support a large number of users and prevent congestion. Financial users and highly skilled programmers of frontier data science have a strong need for lower latency. This is also true for uses that involve large volumes of transactions or "big data" applications.[1] Such demands arise in many applications, for example, video conferences, multiuser 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"—that is, 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 disaster risks, such as fires, earthquakes, and hurricanes. Another type desires generic and nonurgent computing, such as backup storage, where costs primarily drive choices, and usage can tolerate delay.

2.1.3. Operating Costs. Electricity for servers and air conditioning comprises the largest fraction of operating expenses.[2] Operational costs often vary seasonally, and costs can be relatively higher in hot and humid locations or lower in temperate and dry locations. The costs of replacing equipment can accumulate, and abatements for sales tax and property taxes can reduce those costs. The operational costs of a smaller data center exceed tens of millions of dollars a year, while a large data center can cost hundreds of millions of dollars. Further, cost variations across locations can exceed several tens of millions between locations (CBRE 2015).

2.1.4. Connectivity. The availability of internet trunk lines is a crucial factor in the location of data centers, working in tandem with other elements that support low latency, such as proximity to users. All data centers must be connected to the internet, while superior connectivity with high-bandwidth lines enables the data center to handle large data volumes. Applications that can tolerate delays do not necessitate as much bandwidth. In practice, major broadband lines run along all major highways and serve all major population centers, ensuring data centers rarely face connectivity issues. However, depending on the location and quality of existing lines, some localized upgrading may be necessary.

2.1.5. Size and Electricity. The industry measures capacity as the maximum electricity required by a data center building.[3] A minimal data center has a capacity of five megawatts (MW), while the largest, sometimes referred to as "hyper-scale," far exceeds this size. Potentially limiting factors on the size of a data center are the equipment needed to support and cool servers efficiently, as well as the availability of sufficient land in a densely populated area. As an example, AWS has encountered trade-offs between costs and performance at hyperscales, consequently limiting its data centers to 25–32 megawatts.[4] If AWS wishes to expand after reaching that capacity, it will not add more capacity to an existing building. Instead, it will commission the construction of another structure.

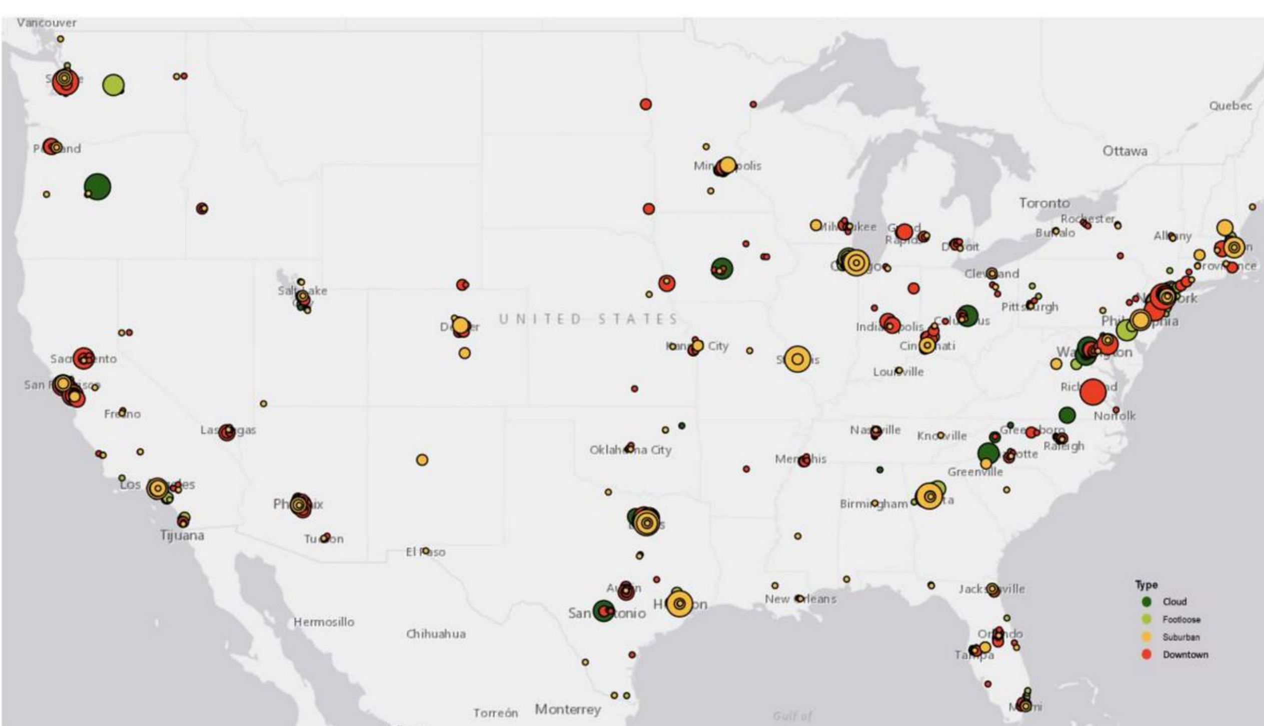
2.1.6. Cost of Construction. Building a new data center typically requires at least a year, with longer timelines for larger projects and those with specialized designs. The construction costs of a 5MW data center exceed $100 million, even in the least expensive locations, while the largest data centers can cost several billion dollars (CBRE 2015). Many factors vary with location. Costs increase when construction wages rise, land is more expensive, and the acquisition costs of the first generation of equipment increase. It also increases when the costs of building high-capacity connectivity to fast backbone internet lines rise, when the costs of building (multiple) electrical service lines increase, and when tax subsidies and abatements are insufficient to defray these expenses. Special features also increase costs, such as reinforced structures to reduce vibrations, raised floors to prevent flooding, backup generators for reliability, specialized cooling equipment and insulation to mitigate hot weather, and lines to support connectivity to multiple networks.

3. Data and Estimation on 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 about third-party providers.[5] Suppliers want to make their services known in all but a small number of situations (e.g., to supply services to the military), so our data collection efforts resulted in a census of all third-party data centers in 2018 and 2019. The data source provides a snapshot of the industry at a specific point in time, but offers limited data on entry and capacity over time.

We supplement the data with information about the facilities of the four largest private cloud providers.[6] We provide information on 1,386 data centers from third-party providers. We finished with 1,433 data centers in February 2019 by adding the major cloud supplier centers. Figure 1 provides a map. A county can have a *Downtown*, *Suburban*, *Footloose*, or *Cloud* supply of data centers.[7] The first three are defined by their distance from the nearest downtown

Figure 1. (Color online) Geography of Data Centers



Notes. This figure maps out the geographic locations of data centers. Each point represents a data center and is coded by type and weighted by the facility's total capacity (in SQFT).

area, which contains concentrations of businesses that are the principal buyers of services from data centers. Within each Metropolitan Statistical Area (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 data center and calculate the distance between each data center and its nearest downtown area.

The map illustrates an urban bias. Almost all third-party data centers reside in 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 a MSA.[8] The census classifies 37.2% of U.S. counties as MSAs. More than 2,000 U.S. rural and micropolitan[9] counties have no providers. Major cloud supplier data centers are less likely to be located in urban areas, as 73% of counties with at least one cloud supplier data center are MSAs (see Table A.2b of the Online Appendix).[10]

Entry of third-party data centers results in a geographically concentrated supply. It occurs in only 218 counties (i.e., 18.6% of MSAs). Only 76 counties have a single data center, while 34, 23, and 17 counties have two, three, and four data centers, respectively, accounting for a total of 281 data centers. More than 1,100 third-party data centers reside in the 68 counties that have five or more data centers. Twenty-six counties have ten or more data centers, and 13 contain twenty or more. In other words, the vast majority of third-party data centers are located in one of a small number of the largest MSAs with many competitors.[11] However, the vast majority of small and medium-sized urban locations do not experience any entry.

Because information about maximum electrical usage is unavailable for all data centers, we create a close proxy for capacity, *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.[12] The aforementioned data provides clues about the order of magnitude of the replacement value of these assets, which are covered by third-party providers and cloud suppliers.[13] It takes a simple calculation to show that it must be on the order of magnitude in the hundreds of billions of dollars.[14] We will discuss other features of capacity later in the study.

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

3.1. 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 sectors: Information, Healthcare, Manufacturing, Education, and Finance, Insurance, and 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.[15] Further, population and population density are highly correlated. While both variables are correlated with the prevalence of unobservable types of demand, we cannot include both in the model specification. Therefore, 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 areas with the lowest and highest population densities, as defined by ZIP code, within the county.[16]

While we interpret population density as a measure of local demand, we acknowledge it may also proxy for the availability of skilled labor. However, because data centers require relatively few specialized employees for ongoing operations, we expect the effect of labor availability to be secondary to the more significant effect of local customer demand.

3.2. Operating Costs

The Energy Information Administration provides 2016 state-level average electricity prices.[17] The CDC Wonder North America Land Data Assimilation System provides daily air temperatures from 2016 by county. We also use the average temperature during the *Hot* and *Cold* months. Following studies of the software industry (Bennett and Hall 2020), we obtained the median hourly wage of network engineers in 2016 from the Bureau of Labor Statistics (BLS). We use state-level real estate tax (RET) rates from the National Association of Home Builders for annual property taxes. 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*.

3.3. Set Up Costs

We approximate land costs by collecting county-level data from Zillow's median home value by square foot. 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 obtained MSA-level sales tax 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.[18] Finally, we use data from the CBRE (2013) on tax incentives for the 30 largest enterprise markets.

We create *Tax Break MSAs based on* whether a county is part of an MSA that provided tax breaks to data centers in 2015, the most recent available year. Though listed as a set-up cost, it also applies to many operational costs.[19]

Observing the variance in broadband line availability is a challenging task. Security concerns have made it challenging to obtain public information about high-speed backbone lines.[20] However, while we cannot directly observe bandwidth availability, we can proxy for the quality of local bandwidth using household information about broadband usage. We add three proxies: the percentage of household adoption of the internet supported by DSL, cable modem, and fiber optics. Although data centers tend not to locate in residential areas, a better network among residences—that is, more cable or fiber—correlates with a better network among nearby businesses, and likely saves the data center costs associated with upgrading the existing connectivity. We expect these factors to be salient for third-party networks that offer low latency. Table 1 summarizes the definitions of variables.[21]

3.4. 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, suburban capacity exceeds that of any other type, followed by urban and footloose capacity. A county's suburban capacity is almost three times higher than the footloose capacity (47,622 SQFT versus 17,869 SQFT). We also observe that our capacity measurements appear to have a skewed distribution. Similarly, our density variables appear to have a skewed distribution. In our analysis, we apply a log transformation to both exogenous and endogenous variables to reduce the skewness and allow for an elasticity-based interpretation of the results. We also calculate correlations for the exogenous variables (see Table A.1 of the Online Appendix).[22]

3.5. Descriptive Patterns

The upper graph of Figure 2 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 a population of more than 700,000 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.

Table 1. Definition of Demand & Supply Variables

Variable	Definition
Dependent variables	
Third-Party Providers	Indicator for the 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.
Common Ownership	Indicator for county with at least one firm with multiple data centers in the county.
Demand shifters	
% Bachelor's Degree	% of population with at least a bachelor's degree.
% Information	% of county's employees in the Information sector (NAICS: 51).
% Health	% of county's employees in the Health sector (NAICS: 62).
% Education	% of county's employees in the Education sector (NAICS: 61).
% FIRE	% of county's employees in FIRE.
% Manufacturing	% of county's employees in the 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

Variable	Mean	S.D.	Min	Max
Dependent variables				
Third-Party Providers	0.186	0.389	0	1
Cloud Providers	0.022	0.147	0	1
Common Ownership	0.072	0.259	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	2,912.159	0.8	72,158
Min. Density	23.057	163.762	0	2,793.67
Max. Density	4,342.058	10,916.470	0.814	172,373.4
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	1,142
Counties	1,173			

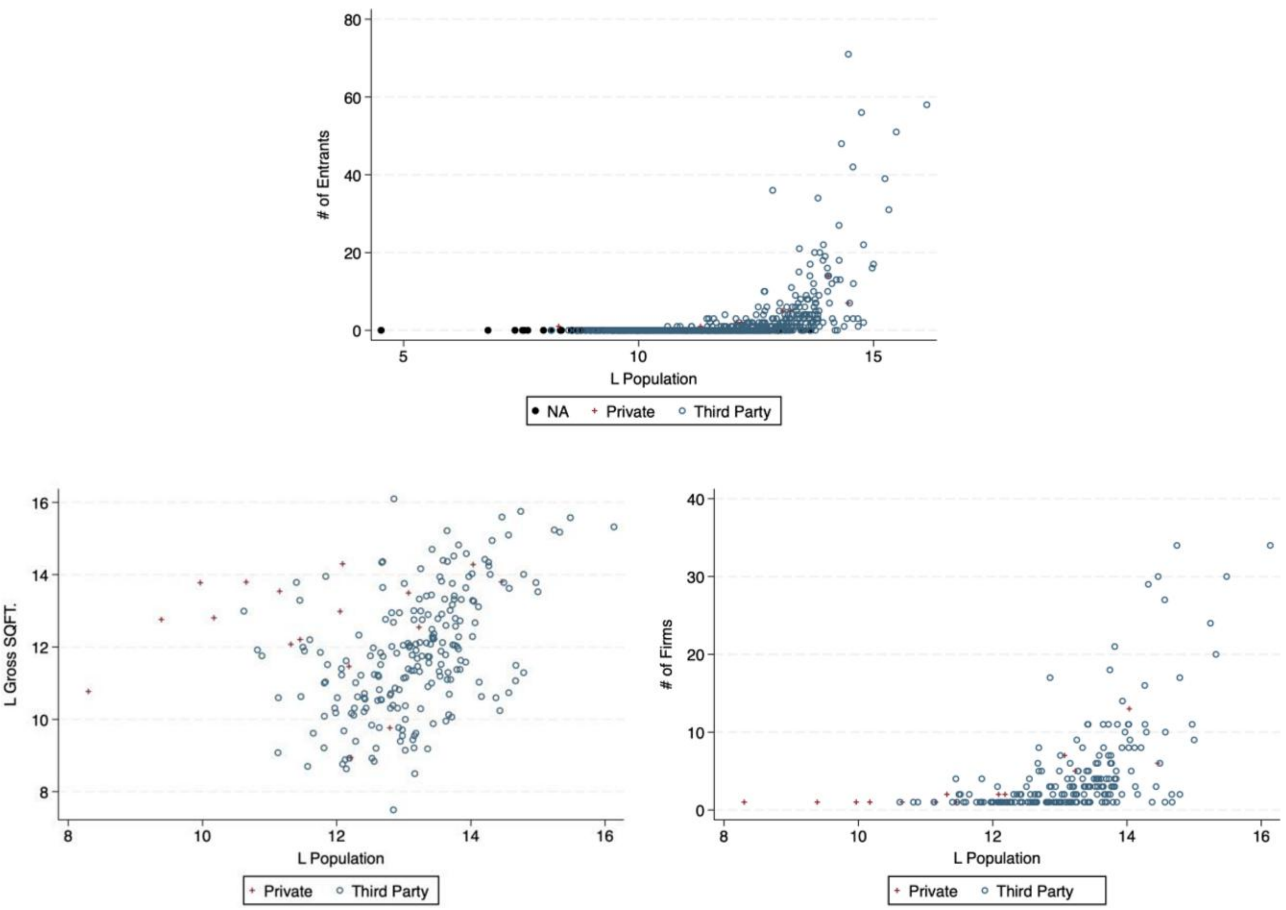
The bottom-left graph in Figure 2 illustrates the relationship between logged capacity and logged population. The figure includes only counties with at least one data center. Third-party data center capacity and population display a positive correlation of 0.596. Capacity is not synonymous with the number of entrants, however. For example, the county with the largest capacity, Loudoun, VA, ranks as the tenth-largest county in terms of entrants. The capacity in Loudoun 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 the 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 concentrations 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 firm has located as much of its infrastructure in one place as Facebook. Google does not display a similar tendency, as it has large facilities in several locations.[25]

Finally, the bottom right graph of Figure 2 shows the relationship between the number of unique firms in a county and the logged population. The figure includes only counties with at least one data center. There is a positive correlation of 0.46. We also find that many of the locations with multiple data centers are the same locations with common ownership and more than one firm entrant.

Table 3 examines whether the prevalence of third-party vendors offering specialized services increases in the largest markets. The table shows the entry of data centers that expended extra costs to comply with two industry-specific standards. It compares markets with populations of more than or fewer than 750,000 and labels them as major and minor. Some data centers protect personal health information in compliance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) and the Health Information Technology for Economic and Clinical Health Act of 2009 (HITECH). Some protect financial data in compliance with Service Organizational Controls (SOC), namely, SOC 1 and SOC 2. SOC1 is primarily concerned

Figure 2. (Color online) Population vs. Entry, Capacity, and Number of Firms



Notes. The upper 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 bottom left graph demonstrates the relationship between county population and the capacity within the county, conditional on some data center entry. The bottom right graph demonstrates the relationship between county population and the number of unique firms within the county, conditional on some data center entry. In both plots, a county is assigned a type.

with internal financial controls related to financial reporting, while SOC2 focuses on IT security, confidentiality, privacy, processing integrity, and availability.

The table shows that major markets attract more suppliers who meet these specialized needs than minor markets. These results suggest that data centers are responsive to local demand when entering a market. The major markets tend to host more finance and medical firms, and there tend to be more of them in such markets. When markets having at least one data center are divided into those with greater than 750K population and those with less, we find that 75% of the larger markets have at least one data center that has adopted a specialized standard, with an average of 3.78 of these specialist providers in each market. However, only 43% of the smaller markets have such specialist providers, with an average of 1.05 per market. The data centers in major markets find it worthwhile to incur the

Table 3. Specialization Analysis

	Major market				Minor market			
Measure	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

Notes. 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 that demonstrate an organization has the appropriate controls in place to protect and account for its financial data. A major market is defined as a county with a population greater than 750,000.

expenses necessary to meet the demands of local firms. As evidence that urban and suburban data centers are responsible for local demand, these findings are more pronounced when the footloose data centers are excluded from the sample. We also find an urban bias in the location of specialized data centers.

3.6. Estimation Strategy

Our primary estimation strategy analyzes the factors that predict the entry of a third-party data center into a U.S. county. For this main analysis, we collapse the different types of third-party data centers (Downtown, Suburban, and Footloose) into a single category to understand the overall drivers of entry. Our dependent variables are indicator variables, Third-Party Providers and Cloud Providers, that equal 1 if a county i contains at least one third-party data center (cloud data center), and 0 otherwise. Our analysis is conducted on a sample of 1,173 U.S. counties that are part of a Metropolitan Statistical Area (MSA).

Given the binary nature of this outcome, we employ a logit model. The probability of data center entry in county i, $P(Y_i = 1)$ is modeled as a function of a vector of explanatory variables $\boldsymbol{X}_i$. The logistic regression model is specified as:

$$\ln\left(\frac{P(Y_i = 1)}{1 - P(Y_i = 1)}\right) = \beta_0 + \beta_1 \boldsymbol{X}_i$$

The vector of explanatory variables, $\boldsymbol{X}_i$, includes county-level measures of local demand (e.g., % Information employees), variable costs (e.g., Industrial Electric Prices), and fixed costs (e.g., Median Construction Wage), as detailed in Table 1. We run separate models to assess the probability of third-party data center entry and the probability of cloud data center entry. To provide meaningful interpretations, we also simulate how a one standard deviation increase in an exogenous variable, evaluated at the mean values of all other variables, influences the probability of data center entry. With only industry beliefs as a guide, we opt to err on the side of testing many exogenous variables (without multicollinearity).

4. Main Results

Table 4 presents the estimates on the entry of data centers. Column 1 presents a logit model where the endogenous variable indicates whether a county has a third-party entrant. Column 2 presents a logit model in which the endogenous variable indicates whether a county has a cloud entrant. All logit estimates are displayed as log-odds ratios. Column 3 presents a logit model where the variable equals one if there is an entry by commonly owned data centers (i.e., when a firm has multiple data centers located in a county), and it equals zero otherwise. It tests whether the same factors influence both commonly owned data centers and third-party data centers. Alternative specifications using OLS yield qualitatively similar results (see Table A.4 of the Online Appendix). Columns 4 and 5 of Table 4 show the simulated changes in the probability of entry for third-party and cloud entrants, respectively.

First, we discuss third-party entry strategies. Column 1 indicates that certain local demand factors lead to increased entry. The proportion of Information and FIRE employees predicts entry and is statistically significant. Column 4 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 estimate the degree of local demand. None of the other types of users matter, including manufacturing or health workers. Column 4 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. A negative association exists between the size of the minimum population density in a county and the likelihood of entry. Taken together, these results are consistent with the interpretation that data center firms are balancing competing priorities: they select highly dense counties (and part of populous MSAs) to serve strong local demand, but within those counties, they seek out less dense parcels to minimize land costs and secure larger tracts of 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 associated 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 4 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.[26]

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 for local information industries, and their 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 percentage of the population with

Table 4. Entry Analysis of Data Centers by Type

Variable	(4.1) Third-party providers	(4.2) Cloud providers	(4.3) Common ownership	(4.4) Simulated gains Third-party entry	(4.5) Simulated gains Cloud entry
Demand Shifters					
ACTPercent Bachelor's Degree	1.024	0.895*	1.022	0.009	−0.002
	(0.026)	(0.045)	(0.039)		
Percent Information	1.314**	1.860***	1.548***	0.019	0.005
	(0.136)	(0.303)	(0.185)		
Percent Health	0.940	0.958	0.879	−0.013	−0.001
	(0.033)	(0.067)	(0.058)		
Percent Education	1.038	0.681	0.975	0.000	0.000
	(0.652)	(0.966)	(0.860)		
Percent Finance, Insurance, & Real Estate	1.248***	1.135	1.185**	0.031	0.002
	(0.069)	(0.085)	(0.067)		
Percent Manufacturing	0.953	1.02	0.904	−0.014	0.001
	(0.027)	(0.042)	(0.057)		
Log Population Density	1.887***	1.462	2.274***	0.074	0.003
	(0.284)	(0.429)	(0.519)		
Log Minimum Population Density	0.805*	0.460*	0.703**	−0.013	−0.002
	(0.069)	(0.147)	(0.095)		
Log Maximum Population Density	1.750***	0.974	1.801**	0.091	0.000
	(0.202)	(0.214)	(0.392)		
Variable Cost Shifters					
Industrial Electric Prices	1.071	0.402*	1.181	0.007	−0.003
	(0.061)	(0.155)	(0.103)		
Low Real Estate Tax	0.696	0.253	1.687	−0.018	−0.004
	(0.217)	(0.196)	(0.862)		
High Real Estate Tax	0.738	1.393	1.695	−0.014	0.001
	(0.208)	(0.871)	(0.715)		
Median Network Engineer Wage	0.939	1.142	1.007	−0.013	0.004
	(0.034)	(0.092)	(0.056)		
Cold	1.035	0.946	1.062*	0.016	−0.001
	(0.019)	(0.047)	(0.032)		
Hot	0.952*	0.959	0.950	−0.016	−0.001
	(0.019)	(0.038)	(0.033)		
Fixed Cost Shifters					
Median Construction Wage	1.046	0.804*	0.992	0.010	−0.002
	(0.040)	(0.087)	(0.057)		
Tax Break MSA	2.227*	1.552	1.769	0.054	0.002
	(0.719)	(0.941)	(0.526)		
Low Sales Tax	0.686	0.755	0.940	−0.018	−0.002
	(0.222)	(0.49)	(0.282)		
High Sales Tax	1.420	0.936	2.568***	0.014	0.000
	(0.369)	(0.625)	(0.600)		
Median Home Value (per SQFT)	1.006*	1.001	0.999	0.030	0.000
	(0.003)	(0.007)	(0.002)		
Min. Home Value (per SQFT)	0.981***	1.018*	0.985**	−0.031	0.005
	(0.004)	(0.009)	(0.006)		
Max. Home Value (per SQFT)	1.000	0.995	0.999	−0.002	−0.002
	(0.001)	(0.004)	(0.002)		
Observations	1,173	1,173		1,173	1,173

Notes. Standard errors in parentheses. Columns 4.1, 4.2, and 4.3 are logit models, and the coefficients are log odds ratios. The dependent variable in Columns 4.1 and 4.2 is an indicator for data center entry of a particular type. The dependent variable in Column 4.3 is an indicator for common ownership in a county (a firm with multiple data centers in a county). Columns 4.4 and 4.5 show the change in probability of entry from a one-standard-deviation increase, calculated at the mean values for all exogenous variables.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Bachelor's degrees, as well as the minimal density, but its importance is minimal. We are unable to distinguish between an interpretation of local industry that favors the demand or supply of labor. Still, in both cases, we conclude that the comparative role of proximity is much less influential for cloud suppliers.

We observe several distinct outcomes for the variable and fixed costs, specifically that higher electricity

Table 5. Entry Analysis of Data Centers by Type with Broadband Adoption Measures

Variable	(5.1) Third-party providers	(5.2) Cloud providers
Broadband availability		
Percent cable modem	1.021*	0.964
	(0.009)	(0.025)
Percent DSL	0.980	0.987
	(0.014)	(0.017)
Percent fiber	1.039***	1.016
	(0.008)	(0.014)

Notes. Standard errors in parentheses. Columns 5.1 and 5.2 are logit models, and the coefficients are log odds ratios. The dependent variable is an indicator for data center entry of a particular type. The other coefficients are similar to those in Table 4. See Table A.8 in the Online Appendix for all estimates.

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

prices discourage cloud entry, and lower median construction wages encourage entry from major cloud providers. The same applies to the minimum land value. Although none 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.

In Column 3, we examine the determinants of common ownership in a county where a firm has multiple data centers in the same area. These results suggest that demand shifters are an important determinant in increasing the likelihood of firms opening multiple establishments in the same area. These results are qualitatively similar to the behavior of third-party data centers in Column 1. The same factors influence the entry of commonly owned data centers in the same direction and to approximately the same degree. They are consistent with the interpretation that entry patterns are similar at the firm and establishment levels.

Table 5 shows the coefficients for the broadband estimates, displaying the log-odds estimates from Table A.8 in the Online Appendix. The estimates are consistent with the hypothesis that a better network plays a small but meaningful role in explaining regional differences in data center locations. As expected, these factors play no role in predicting cloud data centers because proximity to users and low latency are not considered.

5. Additional Analyses

Local demand should not shape the decision of footloose and private cloud providers, who, by assumption, satisfy national demand, and, hence, locate in response to localized costs but not local demand. We tested for differences in the behavior of various types of data centers, including Downtown, Suburban, Footloose, and *Private* data centers. Our results appear in Table A.5 of the Online Appendix. We do not find evidence to reject the hypothesis of equality between the urban and suburban models. We perform a similar test for footloose and private ("cloud") data centers and do not find enough differences to distinguish between the models. While the entry behavior of urban and suburban data centers is similar, they differ from footloose and private data centers. This is consistent with localized demand varying in importance for different types of providers.

Our main results focus on the decision of whether to enter a county. That leaves unanswered whether further inferences can be drawn from examining the extent of data center entry. We explore this by adopting a flexible functional form for demand. This approach simultaneously estimates a threshold for the marginal entrant at zero/one, one/two, two/three, three/four, four/five, or more. Our results (see Table A.6a of the Online Appendix) primarily emphasize the relationship between more population and more entrants. In other words, once other factors are controlled for, the second, third, and fourth entrants will be determined by approximately the same incremental increases in population levels. 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. We also simulated the difference in fixed costs due to observable factors for third-party data centers (see Table A.6b of the Online Appendix). We compared markets with a single entrant with those with more than five entrants. The estimates indicate that fixed costs are 24% higher in competitive locations due to observable factors. Larger markets with the greatest incidence of entry tend to have the highest fixed costs and the lowest profit margins. We interpret this as evidence that locating in a major metropolitan area, and near more users, makes up for these more challenging cost conditions and more competitive margins. In other words, local demand considerations play a more central role than operational considerations in covering the fixed costs of entry for data centers in densely populated urban locations. We provide additional details about this model in our Online Appendix.

A key econometric challenge in our analysis of data center capacity is the potential for endogenous sample selection. This issue arises if unobserved factors that influence a firm's initial entry decision—the selection mechanism—are correlated with unobserved factors that determine its subsequent capacity choice. For example, a county's unobserved protechnology disposition might both attract firms to enter and encourage them to build larger facilities. Failing to account for such nonrandom selection would yield biased and inconsistent estimates (Wooldridge 2010, Certo et al. 2016). To address this type of endogeneity, we employ a Heckman two-stage model, which explicitly models the selection process to correct for the bias in the outcome equation. Estimating capacity requires a two-equation approach, where the first equation accounts for selection by threshold into a situation in which positive entry occurs, and the second equation predicts total demand, conditional on entry having occurred. To satisfy the exclusion restriction and to avoid issues of multicollinearity, we include a set of fixed-cost shifters in the first-stage entry model that do not appear in the second stage (Certo et al. 2016). We argue these variables plausibly affect the initial decision to enter a market without directly determining the scale of operating once entry occurs.

We find that the demand-side effects for the Information and FIRE industries do not predict demand once we control for the selection of any entry (see Table A.7b of the Online Appendix). We 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 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. Only a few coefficients predict capacity once we control for entry, so only weak conclusions emerge. Although we observe some differences in the coefficients for local and footloose equations, these 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 most significant outlier to the relationship between population and capacity is found in Ashburn, Virginia, located in Loudoun County, which has the highest capacity in the United States. Industry consensus holds that firms' entry and growth are driven by the 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 promising 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 most extensive collection of AWS data centers. However, one observation does not make a statistical relationship, and we could not identify any other area that the industry consensus had identified in similar terms.

6. Discussion

The study shows that third-party data centers and hyperscale cloud providers follow distinct location strategies. We find that third-party firms are sensitive to local demand and favor urban locations, while cloud providers are comparatively more sensitive to input costs, such as electricity prices. In this section, we discuss the implications of our findings for strategy literature and practice. We first examine how the distinct location strategies of data center providers inform and extend existing theories of urban bias, agglomeration, and general-purpose technologies. We then detail the consequences of this economic geography for managers and policymakers.

6.1. Theories of Proximity and Urban Bias

Our findings contribute to the literature on the economic geography of IT by revealing a strategic divergence in how firms approach proximity. While third-party providers locate near to customers, reinforcing standard theories of urban bias, cloud providers demonstrate a novel logic driven by input costs, offering a crucial counterpoint. Consistent with this division in entry strategies, we cannot reject the hypothesis that the entry strategies of footloose data centers resemble those of cloud providers and differ from those of suburban and urban data centers.

The location strategy of third-party data centers aligns with and builds upon existing theories that emphasize the persistent advantages of urban locations. These providers exhibit a strong urban bias, a finding in line with studies of other information technology suppliers (Forman et al. 2008, Greenstein et al. 2018). This bias arises from the concentration of information-intensive activities in urban centers and the private incentive to spread fixed costs by locating near users (Forman et al. 2009, Tambe 2014). The "distaste for distance" among buyers—whether due to latency concerns or a desire for managerial oversight—compels these suppliers to enter high-cost urban markets, even when the services are highly "shippable." This finding is consistent with the model of a single supplier's demand presented in Wang et al. (2019), and our work extends this finding to the broader third-party market. The concentration of specialized services (e.g., HIPAA and SOC compliant) in major markets further reinforces this point.

In contrast, the strategy of hyperscale cloud providers offers a crucial nuance to the "death of distance" narrative. In a pattern with no clear equivalent in other IT markets, these firms exhibit a heightened sensitivity to supply-side costs, such as electricity and construction wages, while being far less influenced by local demand or population density. This does not mean geography is irrelevant; instead, its importance shifts from proximity to users to proximity to low-cost inputs. Enabled by highly "shippable" services (Gervais and Jensen 2019) and low information friction (Goldfarb and Tucker 2019), cloud providers can reach customers from anywhere. This frees them from the need to produce everywhere, allowing them to strategically concentrate their massive facilities in a few optimized, low-cost locations. This cost-driven logic stands in stark contrast to the demand-driven strategy employed by third-party providers.

6.1.1. Agglomeration Economies. A key finding of our study is that data center location is largely decoupled from specialized human capital, a pattern that contrasts with clustering in other technology sectors. While recent work has pointed to an increased concentration of specific technologies in clusters at the regional level (Puga 1999, Alcácer and Chung 2014, Alcácer and Delgado 2016, Delgado et al. 2016) and the city level (Kalyani et al. 2025), we find little evidence of similar widespread agglomeration for data centers except in Ashburn, VA. Since their location is not driven by skilled labor markets (Le Bas and Miribel 2005), the fact that the "dog did not bark" in other potential hubs with high demand suggests that for this type of infrastructure, traditional agglomeration forces are either absent or outweighed by other constraints.

The one significant instance of agglomeration in Ashburn, Virginia, appears to be driven by unique infrastructure characteristics and network path dependencies rather than local labor markets. Its emergence as a hub is consistent with its unique history as a public data exchange point on major data traffic routes. This aligns with Alcácer and Chung (2014) on how specific infrastructure, such as a transportation hub, can be a primary driver of clustering. The lack of similar agglomeration elsewhere implies that limiting factors can be significant constraints on the development of such clusters. Examples would include the high cost and scarcity of available land, particularly in counties with dense populations, extending the literature by highlighting these powerful countervailing forces. Despite this different clustering logic, our results also reinforce a key tenet of agglomeration theory: firms located within established clusters still reap benefits. We find that business users in major metropolitan areas with a concentration of data centers receive superior access to infrastructure and specialized services. This finding is consistent with literature showing that firms in strong clusters benefit from better access to inputs and related industries (Delgado et al. 2010, Alcácer and Delgado 2016, Delgado et al. 2016).

6.1.2. Digital Infrastructure Location. This study provides one of the first statistical analyses of data center location, addressing a notable gap in the literature on the economic geography of critical, but often invisible, internet infrastructure (Cardona et al. 2013, Greenstein 2021). Our research contributes to the literature that analyzes the factors that shape the "supply chain" of services that make up the internet. Our analysis emphasizes the less visible services that complement internet access. In contrast to extant strategy literature that is focused on other technical markets, we find that this upstream technology infrastructure operates under different conditions and exhibits a different pattern of firm entry. For instance, a large majority of small and medium-sized areas must rely on nonlocal suppliers for data center services, a stark contrast with the patterns for broadband and other advanced IT services, which are available at much lower population density levels (Xiao and Orazem 2011, DeStefano et al. 2018, Greenstein 2021).

The study also adds an interesting wrinkle to the literature on general-purpose technologies (GPTs). While GPT theory often focuses on how the extent of the market encourages entry (Bresnahan and Gambardella 1998), our study emphasizes the spatial trade-off for upstream suppliers. We show how these firms must balance the pursuit of economies of scale with the private incentive to locate near downstream users to spread high fixed costs. The observed tension between distinct demand-driven versus cost-driven strategies we observe offers empirical detail on how different types of firms providing a similar underlying technology navigate this fundamental trade-off. We share motivation with the analysis and measurement of the cloud market (Byrne et al. 2018, Coyle and Nguyen 2018, DeStefano et al. 2023, Jin and McElheran 2024), a new class of infrastructure that contains assets worth hundreds of billions of dollars and drives significant productivity improvements. We spotlight an overlooked topic: suppliers providing services in close proximity to users. No research other than that of Wang et al. (2019) addresses this topic. They focus on the demand for proximity faced by one cloud provider, while we focus on the broad patterns across all suppliers.

6.2. Strategic Implications

This study provides insights into strategic tradeoffs and offers some speculative observations about long-term trends for data center managers. These implications can guide their decision-making processes and help them navigate the complex landscape of data center location strategies.

6.2.1. Data Center Managers. Supplier proximity to users who demand data center services alleviates a buyer's "distaste for distance." Still, high-density markets are associated with higher setup costs. The trade-off between the gains from proximity and the costs of urban density appears fundamental and irreducible. The resulting location behavior aligns with the model of a single supplier's demand in Wang et al. (2019), in which a firm's buyers are willing to pay more for services if they perceive performance improvements are associated with the proximity of the physical assets that support those services. Whereas they examined only the demand for one cloud provider firm, we find evidence that this model generalizes. We see behavior consistent with the demand for proximity across many third-party data center providers.

Relatedly, but more speculatively, these findings also align with industry forecasts about the potential for new applications that can tolerate delay. This is an important open question for future location decisions. For example, a recent (high profile) entrant into the data center market, Crusoe Energy (https://crusoe.ai/), locates data centers next to natural gas supplies that would otherwise be burned off during oil drilling. This strategy enables the negotiation of an inexpensive source of electricity. As a result, its data centers are typically far from business centers (e.g., in areas like West Texas). Crusoe Energy has made a strategic bet that the lower cost of electricity will make up for the high air conditioning costs of its West Texas locations. It also makes a bet that there will be sufficient growth in demand for data centers to perform actions related to Bitcoin mining. This service does not require proximity to urban users or algorithm development for AI, making proximity essentially unnecessary. Further, it has designed its data centers to accommodate these specialized applications.[27]

This research has another somewhat different implication for third-party providers. Instead of viewing cloud providers as competitors, third-party data center managers should recognize the strategic value of their location for cloud providers. They are not just service providers but also integral players in shaping the industry's future. Just as in Böttger et al. (2018), with a large number of CDNs housed within data centers across various regions, there is a significant potential for data centers to profit from partnering with cloud providers. These providers may increasingly seek space in favorable locations to establish distributed architectures with a local presence in major cities. This trend could give many data center managers opportunities to negotiate colocation agreements with major cloud providers, enhancing their profitability.

6.2.2. Cloud Providers. The differences between the location determinants of significant cloud suppliers and third-party firms underscore the urgency and significance of location decisions for cloud providers. The importance of proximity is a limiting factor for cloud providers who seek to concentrate their data centers in remote, low-cost locations. We also see limited evidence of agglomeration among the cloud providers. AWS does locate some of its data centers near one another—for example, in Ashburn, VA. However, AWS has adopted geographically tailored services for its cloud infrastructure applications (Byrne et al. 2018), indicating that localization also plays a role in its design choices.

The demand for distributed cloud architectures could significantly shape the industry's future.[28] To achieve an architecture sensitive to local demand, cloud providers would either build their data centers in many locations, acquire third-party data centers in many locations, or negotiate colocation agreements in third-party data centers. Again, as noted above, we might expect many colocation agreements to be the easiest way to become close to potential customers to relieve concerns about network congestion.

6.2.3. Buyer Perspective. Business developers who need proximity to large data centers, such as those with enormous data sets, will experience better performance in medium to large cities, where they can be assured of access to nearby data centers. Our findings align with the view that firms operating in clusters of advanced firms in major urban areas receive better technical services (Alcácer and Delgado 2016, Delgado et al. 2016), while firms that operate businesses in small and medium cities will be most affected by the geographic choices of vendors. They must either build for their own use and not gain the benefits of scale economies from sharing infrastructure with others, or they must choose to find services from distant locations, giving up the benefits of proximity (Forman et al. 2008). This has implications for operational costs and service quality, particularly for services with potential latency issues and specialized needs.

6.3. Public Policy Implications

Numerous government programs seek to build out infrastructure in nondense areas. For example, the $42 billion Broadband, Equity, Access, and Development (BEAD) program targets households and businesses in low-density locations where private firms are unwilling to build due to high location-specific costs (Rosston and Wallsten 2022). It has led to the construction of high-speed lines in many small cities, most of which have a population of under 25,000. The BEAD program affects data center demand, potentially increasing usage from these areas to the closest data centers. While this improves the speed of service for business over the alternatives (typically satellite), we remain skeptical that this will motivate more additional data centers to enter markets in rural locations. Our results indicate that demand in small and medium-sized cities with such broadband before the pandemic was insufficient to generate any new openings in the past in cities with a population of less than 160,000. Moreover, most large cloud providers already have many rural locations from which to choose. Hence, we conclude that the BEAD program is unlikely to motivate a data center to be located nearby unless inexpensive electricity justifies it.

Local and state business development policies have also tried to use abatements and other tax policies to encourage data center entry (CBRE 2013, 2015). Some of these actions are premised on aspirations to generate an agglomeration of technology activity. Our results suggest that the cost-benefit calculation should focus solely on building the data center alone, and it is an error to try to encourage agglomeration.

Since the pandemic, local business development policies have had to account for "Work-from-Home" (Bloom et al. 2024) and the intriguing concept of "Work-from-Anywhere" (Choudhury et al. 2021). These new work trends have altered the geographic features of data traffic during work hours, making data centers in downtown areas less attractive but generally moving traffic to suburban locations in the same county and region. While we do not expect this to significantly impact the data center location in the short term, there is potential for it to do so in the long run. The potential widespread adoption of work-from-anywhere could have more dramatic effects on the location of cloud infrastructure. It would need to accommodate many places far from downtown, a demand that seems especially suited to cloud provision (Byrne et al. 2018).

As the public cloud diffuses across a global market, policy initiatives such as GDPR could also shape data center locations outside the United States (Kannan et al. 2021). These regulations pressure firms to use local facilities for computing and storage to prevent privacy-violating events (Zhuo et al. 2021). For example, the GDPR could motivate the building of data centers within European boundaries, thereby shifting some demand for computation and storage away from U.S. data centers (Leka 2025). That could reduce the growth in demand for data centers within the U.S. boundaries.

6.4. Further Research

This research addresses two views that animate open policy questions about digital infrastructure supply. An "optimistic" outlook anticipates experimentation in a few places, usually large cities and technology hubs, followed by more diffusion to more users, regions, and applications. It interprets the state of digital infrastructure at a point in time as temporary, transient, and amid broader 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, increased productivity from the colocation of many related activities, and 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 because cloud providers typically locate their data centers in a limited number of locations.

One central open question is whether such patterns will persist after the providers adjust to the postpandemic demands for their services and the rise of artificial intelligence-related computing needs (Potts et al. 2024). 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. However, it focused on only one aspect of the strategy firms pursue, and there are many issues that remain uncovered. Additional studies could focus on the build and acquisition strategies of the largest third-party providers. There is room for careful analysis of the strategies of firms that meet specific industry needs, such as finance and healthcare. Enough time has also passed to observe the consequences of building cloud infrastructure in less-dense locations and evaluate the implications for the locations that subsidized their entry. Finally, many cloud providers have undertaken significant investments outside the United States. All the strategic issues behind those investments are open questions.

7. Conclusion

This study highlights a divergence in the economic geography of digital infrastructure. Third-party data centers exhibit a clear urban bias, driven by a strategic need to be physically close to localized demand. On the other hand, cloud providers show a heightened sensitivity to supply-side costs, leading them to concentrate in a few low-cost locations. As the digital economy continues to consolidate in the cloud, these distinct location strategies—and the choices made by a few large providers—will profoundly shape which regions capture the economic benefits of the next wave of digitization.

Acknowledgments

The authors thank the senior editor, Robert Seamans, for insightful suggestions. Thanks to seminar audiences, Victor Bennett, Tim DeStefano, Alphonso Gambardella, Igal Hendel, Aaron Kulick, Michael Rechtin, Jonathan Timmins, and anonymous reviewers for comments and suggestions. Thanks to Patrick Clapp, Paul Gregory, Maggie Kelleher, and Christine Snively for excellent research assistance. Any errors are our own.

Endnotes

[1] For example, if a big data application uses two data centers, Amazon keeps them within a hundred miles 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.

[2] An 8MW data center in a moderate climate approximately splits electricity between servers and air conditioning, for example.

[3] That capacity measure accommodates different uses for the servers, whether for computation or storage. As a general rule, an 8 MW data center would typically house approximately 50,000 servers.

[4] See https://www.infrastructure.aws/, accessed December 2020.

[5] 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).

[6] These providers are Alphabet, Amazon, Meta, and Microsoft.

[7] A third-party data center is considered “Downtown” if it is within 5 kilometers of the nearest downtown area. “Suburban” centers are between 5 and 30 kilometers from downtown. “Footloose” centers are more than 30 kilometers away. The cloud providers own the “Cloud” data centers.

[8] The first calculation is the number of counties in MSAs over the total number of counties (1,173/3,150). The second calculation compares the number of counties in MSAs with at least one data center against all counties with at least one data center (218/26).

[9] Micropolitan areas have less than 50k population.

[10] Of the 26 counties with at least one cloud data center, 19 of them are in MSAs (73%).

[11] 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).

[12] In almost all cases, square footage corresponds with total capacity. The rare exceptions are retrofitted buildings, such as the RR Donnelly building, located at 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.

[13] Replacement value is a standard concept for asset valuation, which involves valuing an asset at the reconstruction costs required to replace it from scratch.

[14] 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.

[15] We are somewhat doubtful of the interpretation of supply based on industry interviews. Most third-party data centers employ small workforces, and shortages of inputs other than skilled labor are mentioned more frequently.

[16] Population and density are also highly correlated in ZIP code data.

[17] We also test the prices for commercial customers. Our results remain unchanged when we use alternative measures.

[18] For counties that did not provide an estimate of the hourly wage for network engineers or construction laborers, we used averages from each MSA or state.

[19] Tax abatement and incentive programs can apply to property, sales, and personal taxes. CBRE (2013).

[20] During the dot-com boom, public maps of the long-distance lines for the U.S. internet network were widely available, and firms published these openly to attract customers. This practice ceased after September 11. Nonetheless, even by that point, all the major urban areas in the United States had ample internet line connectivity.

[21] 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).

[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 United States, 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, especially on the East Coast of the United States.

[24] That location was home to MAE-East, one of the earliest data-exchange points on the commercial internet, which was a converted exchange point from NSFNET, and today contains several Amazon data centers. As it turns out, the firms that locate in Loudoun operate some of the very largest data centers in the United States. 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; however, we have no insight into their strategies for these locations. It is a topic for further study.

[26] Fluctuations in weather may also impact energy costs for data center facilities. In Table A.9 of the Online Appendix, we add additional measures for variation during the winter and summer to the baseline model and find similar results.

[27] See transcript at https://www.acquired.fm/episodes/saving-the-planet-with-better-ai-data-centers-with-crusoe-ceo-chase-lochmiller, accessed October 2024.

[28] https://www.gartner.com/smarterwithgartner/the-cios-guide-to-distributed-cloud, accessed October 2024.

References

Alcácer J, Chung W (2014) Location strategies for agglomeration economies. *Strategic Management J.* 35(12):1749–1761.

Alcácer J, Delgado M (2016) Spatial organization of firms and location choices through the value chain. *Management Sci.* 62(11):3213–3234.

Bennett VM, Hall TA (2020) Software availability and entry. *Strategic Management J.* 41(5):950–962.

Bloom N, Han R, Liang J (2024) Hybrid working from home improves retention without damaging performance. *Nature* 630(8018):920–925.

Blum BS, Goldfarb A (2006) Does the internet defy the law of gravity? *J. Internat. Econom.* 70(2):384–405.

Böttger T, Cuadrado F, Tyson G, Castro I, Uhlig S (2018) Open connect everywhere: A glimpse at the internet ecosystem through the lens of the Netflix CDN. *ACM SIGCOMM Comput. Comm. Rev.* 48(1):28–34.

Bresnahan T, Gambardella A (1998) The division of inventive labor and the extent of the market. Helpman E, ed. *General Purpose Technologies and Economic Growth* (MIT Press, Cambridge, MA), 253–281.

Byrne D, Corrado C, Sichel DE (2018) The rise of cloud computing: Minding your P’s, Q’s and K’s. NBER Working Paper No. 25188, National Bureau of Economic Research, Cambridge, MA.

Cardona M, Kretschmer T, Strobel T (2013) ICT and productivity: Conclusions from the empirical literature. *Inform. Econom. Policy* 25(3):109–125.

Cassiman B, Ricart JE, Valentini G (2022) Commitment and competitive advantage in a digital world. *Strategy Sci.* 7(2):130–137.

CBRE (2013) Impact of taxes & incentives on data center locations. Report, CBRE Research, New York.

CBRE (2015) Site selection strategies for enterprise data centers. Report, CBRE Research, New York.

Certo ST, Busenbark JR, Woo H, Semadeni M (2016) Sample selection bias and Heckman models in strategic management research. *Strategic Management J.* 37(13):2639–2657.

Choudhury P, Foroughi C, Larson B (2021) Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management J.* 42(4):655–683.

Coyle D, Nguyen D (2018) Cloud computing and national accounting. ESCoE Discussion Paper 2018–19, ESCoE.

Delgado M, Porter ME, Stern S (2010) Clusters and entrepreneurship. *J. Econom. Geogr.* 10(4):495–518.

Delgado M, Porter ME, Stern S (2016) Defining clusters of related industries. *J. Econom. Geogr.* 16(1):1–38.

DeStefano T, Kneller R, Timmis J (2018) Broadband infrastructure, ICT use and firm performance: Evidence for UK firms. *J. Econom. Behav. Organ.* 155(1):110–139.

DeStefano T, Kneller R, Timmis J (2023) Cloud computing and firm growth. *Rev. Econom. Statist.* 1–47.

Forman C, Ghose A, Goldfarb A (2009) Competition between local and electronic markets: How the benefit of buying online depends on where you live. *Management Sci.* 55(1):47–57.

Forman C, Goldfarb A, Greenstein S (2008) Understanding the inputs into innovation: Do cities substitute for internal firm resources? *J. Econom. Management Strategy* 17(2):295–316.

Gervais A, Jensen JB (2019) The tradability of services: Geographic concentration and trade costs. *J. Internat. Econom*. 118(1): 331–350.

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

Greenstein S (2021) Digital infrastructure. Edward LG, James MP, eds. *Economic Analysis and Infrastructure Investment* (University of Chicago Press, Chicago), 409–452.

Greenstein S, Forman C, Goldfarb A (2018) How geography shapes—and is shaped by—the internet. Clark GL, Feldman M, Gertler MS, Wójcik D, eds. *The New Oxford Handbook of Economic Geography* (Oxford University Press, Oxford, UK), 269–285.

Jin W, McElheran K (2024) Economies before scale: IT strategy and performance dynamics of young US businesses. *Management Sci.* 71(9):7449–7473.

Kalyani A, Bloom N, Carvalho M, Hassan T, Lerner J, Tahoun A (2025) The diffusion of new technologies. *Quart. J. Econom.* 140(2): 1299–1365.

Kannan A, LaRiviere J, McAfee RP (2021) Characterizing the usage intensity of public cloud. *ACM Trans. Econom. Comput. (TEAC)* 9(3):1–18.

Le Bas C, Miribel F (2005) The agglomeration economies associated with information technology activities: An empirical study of the US economy. *Indust. Corporate Change* 14(2): 343–363.

Leka E (2025) Exploring digital sovereignty through data flows: Empirical evidence from the backbone of the internet. Report, International Telecommunications Society, Edinburgh, UK.

Potts J, Torrance A, Harhoff D, von Hippel E (2024) Profiting from data commons: Theory, evidence, and strategy implications. *Strategy Sci.* 9(1):1–17.

Puga D (1999) The rise and fall of regional inequalities. *Eur. Econom. Rev.* 43(2):303–334.

Rosston GL, Wallsten SJ (2022) Maximizing BEAD's broadband reach. Technology Policy Institute, Washington, DC.

Sinai T, Waldfogel J (2004) Geography and the internet: Is the internet a substitute or a complement for cities? *J. Urban Econom.* 56(1):1–24.

Tambe P (2014) Big data investment, skills, and firm value. *Management Sci.* 60(6):1452–1469.

Wang S, LaRiviere J, Kannan A (2019) Spatial competition and missing data: An application to cloud computing.

Wooldridge JM (2010) *Econometric Analysis of Cross Section and Panel Data* (MIT Press, Cambridge, MA).

Xiao M, Orazem PF (2011) Does the fourth entrant make any difference?: Entry and competition in the early US broadband market. *Internat. J. Indust. Organ.* 29(5):547–561.

Zhuo R, Huffaker B, Greenstein S (2021) The impact of the general data protection regulation on internet interconnection. *Telecomm. Policy* 45(2):102083.

Tommy Pan Fang is an Assistant Professor of Strategic Management at the Jones Graduate School of Business at Rice University. Professor Pan Fang's research examines how digitization reshapes competition and strategy, with work extending to platform ecosystems, the economic geography of technology, and entrepreneurship. He holds a PhD from Harvard University and a B.S. from the University of Pennsylvania.

Shane Greenstein is the Martin Marshall Professor of Business Administration at Harvard Business School. An economist by training, his research spans boundaries, extending to strategy, regulation, history, marketing, information systems, and organization design. His book, *How the Internet Became Commercial,* won the 2016 Schumpeter Prize. He holds a PhD from Stanford University and a B.A. from UC Berkeley. He previously taught at Northwestern University and the University of Illinois.