International Trade in Data on the Subsea Internet Cable Network*

Jihye Jeon† and Marc Rysman‡

July 2025

Abstract

This paper studies data flows on subsea internet cables as a form of international trade in data. Utilizing a dataset from the cable industry, we estimate a model of country-to-country data exchange. We show that trade in data is growing faster and is more geographically concentrated than trade in goods. While the underlying demand for trade in data is becoming more dispersed, the concentration of cable capacity is causing data flows to be more concentrated. We calculate the elasticity of trade in data to features of the cable network, highlighting the critical role of cable infrastructure for international data flows.

*We thank Haoran Pan and Tim Lee for excellent research assistance. We also thank Savannah Kelly, Lilian Lea, and Paul Brodsky from TeleGeography for providing valuable insights about the industry and the data.

†Boston University. Email: jjeon@bu.edu

‡Boston University. Email: mrysman@bu.edu.

1 Introduction

The internet allows for global flows of data that are growing both in volume and significance. International data flows shape and support many aspects of modern economic activity, including the traditional trade of goods. The central piece of communication infrastructure for these data flows is the network of subsea fiber optic cables that carry internet data. More than 99% of international data traffic travels through these subsea cables, which remain the most efficient way to send information across the ocean. (Mauldin, 2023).

In this project, we study the evolution of international data flows, leveraging new data on the subsea cable network. We characterize flows on the network as *international trade in data*, and we contrast the patterns of trade in data with trade in goods, which has long been a focus of economic research. Studying trade in data provides a new perspective on globalization and the sources of economic growth. There is very little prior literature on this market, especially compared to the long-standing and vast literature on international trade in goods. We contrast the growth of trade in goods with trade in data and examine how geographically concentrated trade in data is relative to trade in goods, and how central specific regions are in these trade networks. Our model allows us to decompose changes in these statistics over the last two decades into the change due to secular trends in internet usage and the amount due to construction of the cable network.

Our paper is the first to use the data on bandwidth usage on submarine cables to infer country-to-country data flows and characterize them as trade in data. Before going forward, we briefly discuss the analogy between trade in data and trade in goods. Fiber optic cables carry internet communication in the form of packets of data that may be physically tracked. Consumers pay local carriers for internet access, who contract with subsea cable owners to carry their traffic. In the sense of packets and fees moving between countries, trade in data is similar to trade in goods. An important difference is that the market for internet access does not typically charge consumers based on the value of individual packets, regardless of content or where the packets originate. We do not attempt to establish a causal relationship between trade in data and goods or evaluate the benefits created by trade in data. We view that as outside the scope of our paper. Rather, we seek to describe trade in data and how it contrasts with more familiar trade in goods, which we view as valuable in its own right and a first step in the larger understanding of these processes.

Our approach is based on Jeon and Rysman (2025). That paper models both trade on the subsea network and construction of the network, and focuses on the efficiency of investment incentives. This paper models only trade on the network, which allows us to consider richer and more varied demand specifications. This paper also focuses on the international trade perspective, particularly the contrast with trade in goods. Our data is primarily drawn from TeleGeography, a proprietary data company that covers internet equipment and services, with particular expertise in the subsea cable industry. TeleGeography tracks a measure of cable usage by countries and also between regions (roughly, continents) from 2002 to 2020. We also observe landing points and construction dates of all cables, and typically, more information such as bandwidth capacity, length of the cable, and owners. We supplement the data with various sources, including trade data from CEPII.

Throughout the paper, we refer to *bandwidth usage* at the region-pair level as a measure of cable usage through the two regions, which includes traffic between endpoints beyond the two regions. In contrast, we use *trade in data* and *data flows* interchangeably to refer to endpoint-to-endpoint usage, which may not use the cables directly connecting those regions. As in Jeon and Rysman (2025), we observe *bandwidth usage* between regions in the data, but not *data flows* or *trade in data* between beginning and endpoints. This problem is similar to Allen and Arkolakis (2022), who observe vehicle traffic between geographic points but not where the vehicles start and end travel. Similarly, our structural model allows us to infer country-to-country trade in data from observed flows on the network. Our model incorporates demand for bilateral trade in data in the spirit of a gravity model, along with a stylized model of routers choosing to allocate traffic to available paths on the cable network.

We find that GDP and population are important drivers of international trade in data, similar to trade in goods. We also find that countries with a trade agreement have higher trade in data. Our results show that data tend to flow on cable paths with higher capacity, towards the shorter paths, and on paths with more cables among those available. The importance of distance in trade in data has implications for our results. For instance, regions-pairs without direct connections must trade via circuitous routes, which degrades their trading benefits.

Our model yields a measure of trade in data, which we compare with trade in goods. We find that trade in data is substantially more geographically concentrated, but falling in concentration at a rapid pace. North America, Europe, and East Asia are the most central in the network of both trade in goods and trade in data. However, East Asia is relatively less central to the network of data flows, and regions such as South Asia and Sub-Saharan Africa are also far more peripheral in the network of data flows.

Naively treating observed bandwidth usage as if it were endpoint-to-endpoint data flows would understate concentration by 23% for the 2018 to 2020 period. Moreover, we find that endowing the world with the cable network of 2020 while fixing other factors at the observed level in 2012 would increase the concentration of data flows by 19.8%. Conversely, endowing the world with all non-cable-related variables at the 2020 level while leaving the cable network as observed in 2012 would decrease the concentration of data flows by 7.3%. These findings indicate that while the underlying demand for trade in data is becoming less concentrated, the cable network remains concentrated, which leads to the concentration of data flows.

We quantify how data flows change in response to increasing the capacity of available paths between countries. We find that increasing the capacity of the direct path of each country pair by one percent would increase the total data flows across the world by 0.37 percent, while increasing the capacity of all paths would result in a 0.57 percent increase. This finding highlights the importance of the cable infrastructure, particularly the importance of indirect paths. We also find that the elasticity of data flows is particularly high in markets with low cable provisions, which indicates that trade in data is more constrained in those markets.

Related Literature

Our paper relates to the literature on spatial equilibrium models and the role of transportation networks. For example, Fréchette, Lizzeri and Salz (2019) study the NYC taxi market in a dynamic model of search and matching and Brancaccio, Kalouptsidi and Papgeorgiou (2020) study the role of transportation for world trade focusing on the dry bulk shipping industry. Other papers that study the impact of transportation networks through spatial models include Fajgelbaum and Schaal (2020) and Allen and Arkolakis (2022) (see Redding and Rossi-Hansberg, 2017, for a survey). Relative to these papers, we provide a new model suited to address communication over a network.

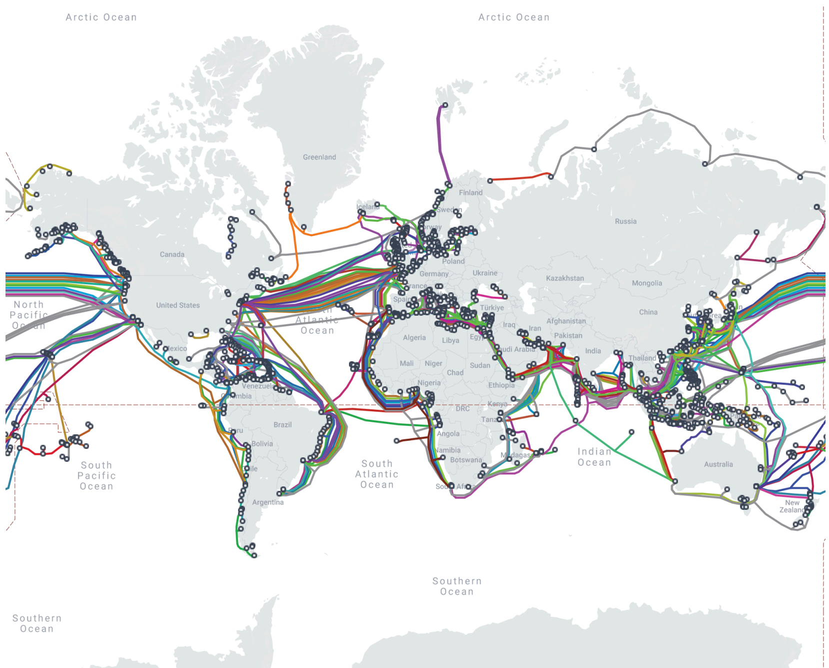
There is a vast literature on global trade flows, in particular on the gravity model (for reviews, see Anderson, 2011, Head and Mayer, 2014). In contrast, there is limited research that measures or analyzes data or information flows with a few exceptions. Keller and Yeaple (2014) characterize information flows within multinationals. Somewhat closer to us, Mueller and Grindal (2019) use data on IP addresses of top websites of each country to measure internet data exchanges between regions and correlate these measures with the trade in goods and services. Krings, Calabrese, Ratti and Blondel (2009) use data from a Belgian mobile phone operator to estimate a gravity model of telephone communications. Our paper is the first to use data on bandwidth usage on submarine cables to infer country-to-country data flows, providing a more comprehensive picture of international data flows. Our structural model allows us to shed light on the drivers of the observed patterns in data flows and provide direct comparisons to trade in goods.

Beyond Jeon and Rysman (2025), there is a small literature in economics on subsea cables. Hjort and Poulsen (2019) show that the arrival of a cable in Africa led to measurable improvement in economic activity for affected populations. Caoui and Steck (2025) study the role of cable breakdowns and the value of redundancy in the market for cables. Cariolle, Houngbonon, Silue and Strusani (2024) evaluate the relationship between cable usage and pricing. Eichengreen, Lafarguette, Mehl and Minesso (2023) study the effect of a new cable arriving in Singapore on trading on financial exchanges. Fang, Tang, Yu and Zhang (2025) study the effect of a cable between China and other countries on Chinese firms' exports. These papers establish the importance of cable landings on local market outcomes, whereas we focus on traffic on the cables themselves.

2 Data

We use the same data sources as in Jeon and Rysman (2025). The main data we use in this paper are from TeleGeography (2002-2020), a telecommunications market research company that provides detailed information on subsea cables and international bandwidth usage. TeleGeography provides a comprehensive picture of the subsea cable network based on information collected using various methods, including confidential surveys, interviews with telecommunications company executives and engineering staff, and other network discovery tools such as aerial or satellite photographs. We are not aware of previous research in economics making use of these data, other than the concurrent papers by Caoui and Steck (2025) and Cariolle et al. (2024). TeleGeography's data provide the characteristics of active and planned subsea cables, including ready-for-service year, cable length, construction cost, ownership structure, landing points, and various capacity measures. Figure 1 provides a map from TeleGeography of commercial subsea cables in 2023.

Figure 1: Map of subsea cables in 2023



Notes: This map shows undersea telecommunications cables as of October 5, 2023. Source: TeleGeography, Submarine Cable Map, www.submarinecablemap.com

TeleGeography collects data on subsea cables, not terrestrial cables, and we are not aware of systematic data on terrestrial cables. As a result, we focus our analysis on pairs of countries that require subsea routes to communicate. In general, this means pairs of countries that do not have a terrestrial route between them (or the route is very limited, as we describe below). Thus, we define regions as sets of countries such that communication between countries in different regions must traverse subsea cables for at least part of the way. Countries in the same region typically have a land route between them, and a significant share of communication goes over terrestrial cables only.[1]

[1]It is sometimes cheaper to lay subsea cables than terrestrial cables so countries with land routes between them may still use subsea cables. For instance, subsea cables along the coast of Europe

We divide the world into the following seven regions: East Asia, Europe, North America, Oceania, South America, South Asia, and Sub-Saharan Africa. TeleGeography has a finer definition of regions (13 regions) in their original data. After discussions with TeleGeography, we combine certain regions into one. For example, we group Central Asia & Caucasus, Eastern Europe, Western Europe, Northern Africa, and the Middle East into one region of Europe based on the observation that there are many viable terrestrial connections across these regions. By contrast, there is almost no terrestrial communication through the Himalayan mountain range, so it is reasonable to keep South Asia and East Asia as separate regions. TeleGeography provides an exact mapping of countries to regions.

There are several concepts of cable capacity. The construction of a cable determines its *potential capacity*.[2] Cable owners sell capacity, which then becomes *purchased capacity*. Purchasers mostly consist of local carriers, content providers, large enterprises (e.g., governments, large companies), and educational institutions (e.g., CERN), with local carriers and content providers taking the great majority. Cable owners and their direct customers always transact in terms of capacity, for example, a given number of wavelengths on a fiber optic cable for a given set of years. After purchasing, customers (such as local carriers and content providers) choose how much to activate, which requires further investment. Our data set includes this as *used capacity*, or equivalently *bandwidth usage*. TeleGeography constructs an equivalent to bandwidth usage for content providers that own and use their own cable. After bandwidth usage is established, it is up to the local carrier how much internet traffic to transmit. If traffic levels exceed bandwidth usage for any given time period, consumers experience service with low quality.

We take potential capacity as our measure of how much capacity a cable provides. We take bandwidth usage as our measure of cable usage between two regions, the dependent variable in our model. We do not observe traffic, but purchasers of capacity closely manage bandwidth usage relative to traffic, so bandwidth usage should be a good measure of traffic. Note that bandwidth usage is not directional. TeleGeography reports a single number for bandwidth usage going in both directions. Our understanding is that this reflects how bandwidth usage is actually transacted. That

connect several different countries. However, we ignore these cables for this paper to focus on communication between countries that must use subsea cables.

[2] It is possible to enhance potential capacity over time through further investment, typically based on technological innovation, although we do not observe this process.

is, in practice, cable customers purchase capacity and convert it to bandwidth usage in both directions in the same amounts.

TeleGeography reports bandwidth usage at the level of the annual region-pair, not the country-pair.[3] We observe bandwidth usage between each of the seven regions. If there is no cable between two regions, we observe a bandwidth usage for that region pair of zero.

Importantly, bandwidth usage between two regions may reflect traffic that is from further endpoints. For example, there was no major cable between Europe and South America as of 2020.[4] Thus, traffic between these regions flowed through intermediate regions, such as Europe to North America to South America. In this case, traffic between Europe and South America would show up as bandwidth usage on these intermediate routes.

TeleGeography also reports international bandwidth usage at the country level (not at the country pair level), that is, the country's total bandwidth usage to other countries. Bandwidth usage for a given country reflects traffic that traverses through the country, as well as traffic starting or ending in that country. This is one variable from TeleGeography that includes bandwidth usage going over terrestrial cables, which makes it difficult to integrate into our model. Still, for island countries, all international usage must be on subsea cables. We pick seven island countries that have substantial international bandwidth usage: Singapore, Japan, Taiwan, Australia, South Korea, the Philippines, and New Zealand.[5] As described below, we use our model to predict bandwidth usage at the country level for island countries.

We also utilize data from the Centre d'Etudes Prospectives et d'Informations Internationales (2002-2020) (hereafter CEPII), as discussed in Conte and Mayer (2022). This data set is standard in the gravity equation literature and includes country-level and country-pair-level information such as GDP, population, geographical distance, proxies for cultural proximity, trade flows, and information on trade agreements and international relationships. We drop countries that ever had a GDP of less than $1

[3]In its own reports, TeleGeography often reports data at higher levels of aggregation and in fact refers to what we call "regions" as "subregions." Fortunately, TeleGeography was able to provide data on used capacity at TeleGeography's subregion-pair level.

[4]The only viable cable in 2020 was Atlantis-2, which had a small potential capacity of 0.16 Tbps and was decommissioned in 2022.

[5]In our framework, the United Kingdom is not an island country. Significant bandwidth runs through the Channel Tunnel and is not classified as subsea, and does not appear in our data set. South Korea is not physically an island, but because no South Korean communication runs through North Korea, its only land connection, South Korea is like an island for our purposes.

billion and countries with missing GDP information, arriving at a sample of 161 countries.

Table 1: Summary statistics

	Mean	SD	5th	95th	N
Market characteristics (21 unique markets)					
Total potential capacity (in Tbps)	73.10	133.20	0.00	397.15	399
Number of cables	3.64	4.55	0.00	14.00	399
Number of owners	29.07	23.98	0.00	66.00	399
Country characteristics (161 unique countries)					
GDP (in billion current USD)	405.47	1,561.70	3.05	1,796.19	3,059
Population (in millions)	42.61	147.29	0.29	143.82	3,059
Country-pair characteristics (9,383 unique country pairs)					
Common language	0.09	0.29			178,277
Trade agreement	0.10	0.30			177,549
GDPR	0.11	0.31			178,277

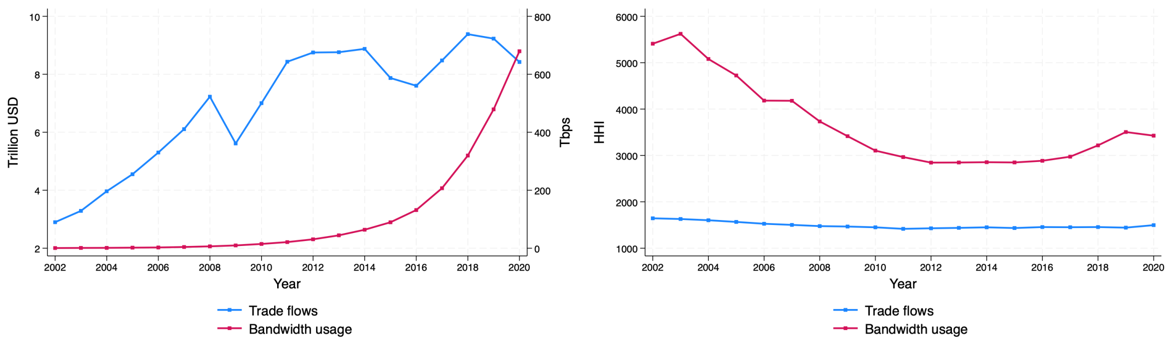
Notes: This table summarizes the market, countries, and country pairs included in our analysis.

We define region-pairs as markets. With seven regions, we have 21 markets. Table 1 provides information on the final sample included in our analysis. By the end of our sample, 15 of 21 markets have at least one cable. On average, a market is connected by 3.6 cables with 73.1 Tbps and 29.1 owners. Markets vary widely in their characteristics, with the largest capacity market, connecting North America and Europe, having 21 cables with a total capacity of 1,118 Tbps in 2020.

We consider several country-pair characteristics that may affect data flows. In our dataset, 9% of country pairs share a common language and 10% have a trade agreement. We also consider the European Union's General Data Protection Regulation (GDPR), which requires data on European citizens to be kept in Europe and so may be expected to reduce data flows between European and non-European countries. We construct an indicator for the GDPR that is set to one for the years 2018 to 2020 whenever one but not both of the countries is in Europe, with the exception of the United Kingdom in 2020 due to the country leaving the EU. Country pairs are under the GDPR for 11% of our sample.

As a reminder, we refer to bandwidth usage as the region-pair bandwidth usage, a market-level variable that does not reflect endpoint-to-endpoint usage, and trade in data or data flows as endpoint-to-endpoint usage. We start by comparing bandwidth

Figure 2: The level and concentration of interregional trade flows and bandwidth usage



(a) Level (b) HHI

Notes: Panel (a) shows the total volume of interregional trade flows and bandwidth usage from 2002 to 2020. The HHI for interregional trade flows (bandwidth usage) in Panel (b) is computed as the sum of the squared share of trade flows (bandwidth usage) over markets (region pairs).

usage in our raw data to international trade in goods. While this comparison is not perfect because of the endpoint issue, the comparison does not require any modeling. Figure 2 shows the overall trend in international trade in goods and bandwidth usage. Panel (a) shows that the total bandwidth usage grew by a factor of 1000 from 2002 to 2020, while the volume of interregional trade in goods grew less than threefold in the same period.

For this figure, we compute the global share of trade in goods in each market and sum squared shares across markets to get the concentration of trade in goods. We do so again for bandwidth usage. For instance, for bandwidth usage, let $\tilde{D}_{mt}$ be the observed level of bandwidth usage in market m and period t and $\bar{\tilde{D}}_t = \sum_m \tilde{D}_{mt}$. Then, concentration is $\sum_m (\tilde{D}_{mt}/\bar{\tilde{D}}_t)^2$. From 2002 to 2020, the HHI for trade flows was substantially lower and relatively constant around 1,500, while the HHI for data flows fell rapidly from approximately 6,000 to below 3,000. The uptick in the concentration of data flows starting around 2016 is surprising, and we return to it below in the context of the results of our structural estimation.

Appendix Table A1 reports measures of concentration at the country level, which shows a similar pattern. In general, bandwidth usage is substantially more concentrated than trade in goods. However, the concentration of bandwidth usage changed rapidly from 2002 to 2020, with the HHI falling from 1,076 to 719. In 2002, only 14 countries accounted for 90% of bandwidth usage, while by 2020 the number rose to 24 countries.

3 A Model of Data Flows and Cable Usage

We present our model in this section. Key points for our model from our data are that bandwidth usage is observed at the region-pair level, not the country-pair level or by individual cable, we do not observe endpoints of bandwidth usage, and we observe only subsea cable networks and not terrestrial networks.

In our model, we study a population of countries indexed by $c = 1, \ldots, C$, and time is indexed by $t = 1, \ldots, T$. Each pair of countries has a joint demand for trade in data between each other that is a function of their demographic characteristics and the level of connectivity with each other provided by the cable network. Let $\overline{M}$ be the maximum amount of data that any pair of countries can trade. Let x_{ckt} represent time trends and country and country-pair characteristics such as the population and GDP of the two countries c and k in period t. We think of x_{ckt} as capturing secular trends in demand for data flows. We represent the quality of connectivity between the countries as v_{ckt}, further described below. The total data traded between the two countries c and k in period t is:

$$d_{ckt} = \frac{\exp(x_{ckt}\theta^d + v_{ckt})}{1 + \exp\left(x_{ckt}\theta^d + v_{ckt}\right)}\overline{M} \quad (1)$$

where the parameters θ^d are to be estimated. Thus, some share of potential demand $\bar{M} - d_{ckt}$ is not realized. This could represent packets that are lost in internet transit or information that consumers choose never to search for in the first place because consumers are aware of internet quality or because the cost of data flows is not worthwhile given the consumer's income, language, and other issues.

We aggregate countries into *regions* $g = 1, \ldots, G$. To match our data set, we assign countries to regions so that communication between countries in different regions must traverse subsea cables. We index pairs of regions by $m = 1, \ldots, M$ and refer to pairs of regions as *markets*. Note that markets are non-directional, so North America-Europe is the same market as Europe-North America.

We also make predictions on total data flows for each of the seven island countries. To do so, we divide regions into subregions, with one subregion for each island country (these subregions contain a single country) and one subregion for the remaining countries. Regions with no island countries contain a single subregion that contains all of the countries in the region. We index *subregions* $r = 1, \ldots, R$. Each country is a member of one subregion, and each subregion is a member of one region.

We assume all countries in the same subregion have the same level of internet connectivity to each other and to the subsea cable network, as if subregions have uniform terrestrial networks. In estimation, we include market fixed effects, which account for the quality of the terrestrial network in each region as well as important pair characteristics, such as region-pair distance.[6]

Trade between a pair of countries must travel on cables through a sequence of subregions. We index the existing *p*aths as $p = 1, \ldots, P_{ckt}$. Each path is a set of subregion pairs for which there is an active cable. In practice, we restrict ourselves to paths that go through four subregions at most.[7] The *direct path* is the path with just the two endpoints on it.

Trade in data d_{ckt} is made up of a continuum of packets of data. Internet routers choose which path each packet will follow according to router software. We capture router decision-making with a reduced-form approximation; routers choose paths according to path *attractiveness* δ_{pckt} that depends on cable features such as path distance and the total capacity associated with each path. The share of data going from c to k on path p at time t is:

$$s_{pckt} = \frac{\exp\left(\delta_{pckt}\right)}{\sum_{z=1}^{P_{ckt}} \exp\left(\delta_{zckt}\right)}. \tag{2}$$

Implicitly, it is as if the router draws a logit epsilon for each packet. We parameterize δ_{pckt} as:

$$\delta_{pckt} = Z_{pckt}\theta^{\delta},$$

where Z_{pckt} includes a constant and path characteristics such as the length, the potential capacity, and the number of cables on path p.

Because internet cables are bidirectional, we have that $P_{ckt} = P_{kct}$ and $\delta_{pckt} =$

[6]Because the market fixed effects capture both cable characteristics (the terrestrial network) and region characteristics such as distance, it is not clear whether it should go in x_{ckt} or v_{ckt}. We place the market fixed effects in x_{ckt} so that v_{ckt} represents the quality of just the subsea cable network.

[7]For example, possible paths between North America and Europe include {North America-Europe; North America-East Asia-Europe; North America-South America-Sub-Saharan Africa-Europe; etc}.

δ_{pkct}. As new cables are constructed, δ_{pckt} may evolve, and the construction of a cable in a new market can change the set of available paths, so P_{ckt} changes over time. All countries in the same subregion have access to the same paths. For example, the US and Spain have the same paths between them as Canada and France.

We are now ready to specify the quality of the connection between two countries. We use the inclusive value of the logit choice among cables:

$$v_{ckt} = \ln\left(\sum_{p=1}^{P_{ckt}} \exp\left(\delta_{pckt}\right)\right).$$

Note that the denominator of s_{pckt} is $\exp(v_{ckht})$. Let $\hat{d}_{pckt}$ be the amount of trade in data between c and k on path p in period t:

$$\hat{d}_{pckt} = s_{pckt} d_{ckt} = \frac{\exp\left(\delta_{pckt} + x_{ckt}\theta^d\right)}{1 + \sum_{z=1}^{P_{ckt}} \exp\left(\delta_{zckt} + x_{ckt}\theta^d\right)} \overline{M}.$$

Thus, data flows increase with the number of paths between two countries (higher P_{ckt}) and the increased quality of those paths (higher δ_{pckt}, such as because of higher capacity). Note that our model does not capture congestion explicitly. In practice, instances of congestion often last for only a short period of time so our data set, which is annual, is not well placed to capture it. But to the extent that congestion leads trade in data to follow high-capacity paths, we address the main feature of congestion by including capacity in Z_{pckt}.

With $\hat{d}_{pckt}$ defined, we can sum up to the values observed in our data set. Recall that a path is defined as a sequence of subregions, but we observe bandwidth usage at the region-pair level, not at the subregion-pair level. Therefore, we sum over predicted bandwidth usage for paths passing through each subregion pair to get the total predicted bandwidth usage between each subregion pair, and then we sum over subregion pairs to get predicted bandwidth usage between each region pair. For example, consider bandwidth usage between the United States and the United Kingdom going through the path consisting of North America-Japan-Europe. Then, we include bandwidth usage on this path in the following two markets: North America-East Asia and East Asia-Europe. Formally, we define $\mathcal{M}_{pckt}$ to be the set of markets on path p between country pair ck.[8] In the above example, $\mathcal{M}_{pckt}$ for path North America-

[8]A path is assigned to a market if two subregions that are adjacent in the path are in different regions. For example, Japan-East Asia-Europe is assigned to the East Asia-Europe market, because

Japan-Europe would include North America-East Asia and East Asia-Europe. Then, we sum $\hat{d}_{pckt}$ over paths that include market m to get the total quantity of data traded between c and k in m:

$$\tilde{d}_{mckt} = \sum_{p=1}^{P_{ckt}} \hat{d}_{pckt} \mathbb{1}\{m \in \mathcal{M}_{pckt}\}.$$

Next, we sum up over country pairs to compute the level of data flows appearing in each market, which is what we observe in our data set. Thus, our prediction for bandwidth usage in market m in period t is:

$$D_{mt} = \sum_{c=1}^{C-1} \sum_{k=c+1}^{C} \tilde{d}_{mckt}. \tag{3}$$

We can define total bandwidth usage at the subregion level in a similar way. We observe country-level bandwidth usage, so these values are useful for island countries that rely on subsea cables for all their international bandwidth usage. We first compute total data flows between country pair ck that goes through subregion r as follows:

$$\check{d}_{rckt} = \sum_{p=1}^{P_{ckt}} \hat{d}_{pckt} \mathbb{1}\{r \in \mathcal{R}_{pckt}\}$$

where $\mathcal{R}_{pckt}$ is the set of subregions on path p between country pair ck. For instance, $\mathcal{R}_{pckt}$ for the North America-Japan-Europe path includes North America, Japan, and Europe, since Japan is its own subregion. Then, we sum up these values over all country pairs to obtain bandwidth usage in subregion r in period t as:

$$Q_{rt} = \sum_{c=1}^{C-1} \sum_{k=c+1}^{C} \check{d}_{rckt}. \tag{4}$$

Japan and East Asia are in the same region. By the same logic, Japan-East Asia is not assigned to any market.

4 Estimation

4.1 Estimation Procedure

We first cover a few more issues in data construction. We must construct path characteristics Z_{pckt}. We sum over cables connecting each subregion to construct total capacity (that is, total potential capacity) and the number of cables between subregions. To do so, we treat any cable that has a landing point in two subregions as connecting those subregions. Some cables connect more than two subregions. In that case, the cable contributes to the capacity between each subregion, but not between indirect paths between subregions. For example, a cable that connects subregions A, B, and C adds capacity to the paths $A - B$, $B - C$, and $A - C$, but not $A - B - C$, or else that would be double-counting the contribution of that cable to get from A to C. In this sense, the capacity on a path $A - B - C$ may be different from the capacity on paths $A - B$ and $B - C$.

Furthermore, if the capacities in the segments (i.e., subregion pairs) of an indirect path are different, which is almost always the case, we take the minimum of capacity across segments to be the capacity for the path. That is, high capacity in one segment cannot substitute for low capacity in another segment. This reflects real constraints in the communication network.[9] We construct the distance of a path as the sum of distances across segments, where the distance of a segment is computed as the average distance between countries in the subregion-pair. For the number of cables, we take the capacity-weighted average across segments. This reflects that, to the extent that consumers care about competition or redundancy, they care about it on all segments. We drop intraregional cables entirely, so, for instance, cables that connect European countries or connect islands in Indonesia are not part of our analysis. Further discussion of data construction appears in the appendix of Jeon and Rysman (2025).

Turning to our moments, our model makes predictions of D_{mt} and Q_{rt}. We denote the predicted bandwidth usage for market m and time t given parameter θ as $D_{mt}(\theta)$ where $\theta = [\theta^d, \theta^\delta]$. We similarly use $Q_{rt}(\theta)$ to denote the subregion-level bandwidth usage. Let $\tilde{D}_{mt}$ and $\tilde{Q}_{rt}$ denote the observed levels in the data. Our es-

[9] Both of these features (the facts that capacity on paths $A-B$ and $B-C$ may differ from $A-B-C$ and that we use the minimum capacity on a path for the capacity of the path) distinguish our model from Allen and Arkolakis (2022). As a result, their multiplicative framework does not apply to our setting.

timation procedure matches the predicted bandwidth usage to the observed levels using the generalized method of moments (GMM). We write the differences between the predicted and observed levels and their first differences as follows.

$$\xi_{mt}^{D}(\theta) = \tilde{D}_{mt} - D_{mt}(\theta) \qquad \xi_{rt}^{Q}(\theta) = \tilde{Q}_{rt} - Q_{rt}(\theta)$$
$$\Delta\xi_{mt}^{D}(\theta) = \xi_{mt}^{D}(\theta) - \xi_{mt-1}^{D}(\theta) \qquad \Delta\xi_{rt}^{Q}(\theta) = \xi_{rt}^{Q}(\theta) - \xi_{rt-1}^{Q}(\theta).$$

For constructing $\xi_{rt}^{Q}(\theta)$ and $\Delta\xi_{rt}^{Q}(\theta)$, we include subregions that contain a single island only and drop other subregions.

Let $\boldsymbol{y}_{mt}^{D}$ be a vector of instrumental variables. We assume that at the true parameters θ^0, $E[\xi_{mt}^{D}(\theta^0)|\boldsymbol{y}_{mt}^{D}] = 0$. Analogously, we assume $E[\xi_{rt}^{Q}(\theta^0)|\boldsymbol{y}_{rt}^{Q}] = E[\Delta\xi_{mt}^{D}(\theta^0)|\boldsymbol{y}_{mt}^{D,\Delta}] = E[\Delta\xi_{rt}^{Q}(\theta^0)|\boldsymbol{y}_{rt}^{Q,\Delta}] = 0$. Let $\boldsymbol{\xi}^{D}(\theta)$ be the $M \times T$ vector of elements $\xi_{mt}^{D}(\theta)$, and analogously for $\boldsymbol{\xi}^{Q}(\theta)$, $\boldsymbol{\xi}^{D,\Delta}(\theta)$ and $\boldsymbol{\xi}^{Q,\Delta}(\theta)$. Similarly, let $\boldsymbol{Y}^{D}$, $\boldsymbol{Y}^{Q}$, $\boldsymbol{Y}^{D,\Delta}$, and $\boldsymbol{Y}^{Q,D}$ be the matrices of instruments $\boldsymbol{y}_{mt}^{D}$ and so on. Then we have moments:

$$\boldsymbol{m}^{D}(\theta) = \boldsymbol{Y}^{D\prime}\boldsymbol{\xi}^{D}(\theta) \qquad \boldsymbol{m}^{Q}(\theta) = \boldsymbol{Y}^{Q\prime}\boldsymbol{\xi}^{Q}(\theta)$$
$$\boldsymbol{m}^{D,\Delta}(\theta) = \boldsymbol{Y}^{D,\Delta\prime}\boldsymbol{\xi}^{D,\Delta}(\theta) \qquad \boldsymbol{m}^{Q,\Delta}(\theta) = \boldsymbol{Y}^{Q,\Delta\prime}\boldsymbol{\xi}^{Q,\Delta}(\theta). \tag{5}$$

Let $\boldsymbol{\Gamma}(\theta)$ be the vector of stacked moments from Equation (5). The GMM estimator is then given by:

$$\hat{\theta} = \arg\min_{\theta} \boldsymbol{\Gamma}(\theta)'\mathbf{W}\boldsymbol{\Gamma}(\theta),$$

where $\mathbf{W}$ is a positive definite weighting matrix that we select through two-stage GMM estimation.

We provide a heuristic discussion of the econometric identification of our model. We cannot allow both x_{ckt} and δ_{pckt} to vary non-parametrically. Consider a single period with one country per region in G regions, each region-pair with a cable connection between them. In this case, we would have $G \times (G-1)/2$ observations of bandwidth usage. We would have the same number of country-match qualities $x_{ckt}\theta^d$ to calculate, plus an equivalent number of path qualities δ_{pckt} for direct paths, plus an even higher number of indirect path qualities. Naturally, we require some restrictions.

We make reasonable functional form assumptions. We project country-match quality onto region-pair fixed effects that are constant over time and a finite set of variables capturing country-to-country match quality (e.g., GDP and population of the two involved countries). Furthermore, moments drawn from island countries contribute to precision over country-level variables. Cable quality parameters in δ_{pckt} are determined by the level of bandwidth usage observed in markets conditional on country-to-country demand.

We include year fixed effects, market fixed effects, and a time trend in our instruments $\mathbf{y}^{D}_{mt}$, $\mathbf{y}^{D,\Delta}_{mt}$, $\boldsymbol{y}^{Q}_{rt}$, and $\boldsymbol{y}^{Q,\Delta}_{rt}$. That is, we take averages of $\xi^{D}_{mt}(\theta)$ across markets and time, and also interact with time to form moments, and similarly for the other error terms. We assume that at the time of investment, firms cannot predict $\Delta\xi^{D}_{mt}(\theta)$ for future values of t.[10] This allows to form additional instruments $\mathbf{y}^{D,\Delta}$ based on lagged cable characteristics. We include the one-period lagged values of capacity and the number of cables for market m in $\mathbf{y}^{D,\Delta}_{mt}$. Jeon and Rysman (2025) include a more detailed discussion about instrument choice and explore alternative choices.

4.2 Estimation Results

Table 2 presents our model estimates for the various specifications we explore. Panel A shows the demand coefficients, θ^{d}. Focusing on our main specification shown in Column (1), the estimates suggest that a one-percent increase in the sum of the GDP of two countries involved in trade would result in an approximately 1.26 percent increase in trade in data, and a one-percent increase in the sum of the population would result in a -0.30 percent decrease in trade in data. Thus, an increase in both GDP and population (an increase in GDP per capita) would lead to an increase in data flows.

We include dummies for three separate time periods: 2002-2009, 2010-2017, and 2018-2020 with the first as the omitted category.[11] We find small negative coefficients for the two later periods, but also a strong time trend coefficient, so overall, demand is growing over time. Note that we do not include distance between countries as an explanatory variable, which would be standard in a gravity regression. Instead, we include market fixed effects, which capture distance and we found valuable in matching the data.

[10] A similar approach was used in Lee (2013) in the context of video games.

[11] The choice of these periods is driven by the patterns in investment shown in Figure A1. The level of investment is relatively low from 2002 to 2009, increases slightly from 2010 to 2017, and then is substantially higher in the 2018-2020 period. Moreover, 2010 marks the first investment by content providers, and the rate of investment by content providers is distinctly higher in 2018-2020.

Panel B presents the path coefficients, θ^{δ}. Again focusing on Column (1), we find that, consistent with our expectations, path length has a negative effect, while path capacity and the number of cables on a path have positive effects on the amount of data flows allocated to the path. One way to interpret the number of cables is as a measure of market structure among cable owners on the path, in which case the positive effect would reflect the enhanced competition when there are more cables. Another interpretation is about redundancy, as carriers sometimes value having more options on a path in case one cable malfunctions, as in Caoui and Steck (2025).

The coefficients in Panel B determine the share of data flows between two countries allocated to each available path. For example, the relative magnitude of the coefficient on path length and capacity determines how much usage goes over the shortest path (the direct path between two regions if a direct path exists) versus more circuitous paths that may have higher capacity. These coefficients also jointly determine the importance of cable quality for the demand for data.

We explore including additional variables in our main specification. We consider including in x_{ckt} whether the two countries share a common language, have a trade agreement, and whether one of the countries is under the GDPR. We include these variables one at a time and altogether. We find that sharing a common language between two countries leads to a 1.5 percent increase in trade in data, as shown in Column (2). The GDPR indicator (when one but not both countries are under the GDPR) has a negative effect and would result in a 0.84 percent decrease. Having a trade agreement would result in a 3.3 percent increase in trade in data. When we put all three variables (common language, GDPR, and trade agreement) in the same specification in Column (5), the coefficients on common language and GDPR lose statistical significance, but the trade agreement maintains a significant and positive effect. Thus, all three variables have the expected signs, and there is some evidence for their statistical significance, but the importance of a trade agreement is our most robust result.

4.3 Understanding trade in data

We now perform several calculations to better understand trade in data between regions. To calculate regional trade, we first compute data flows between each country pair based on Equation (1). Then, we sum these estimates of data flows across all country pairs in each market to obtain predicted trade in data between regions. That

Table 2: Demand Estimates

	(1)	(2)	(3)	(4)	(5)
Demand for data (x_{ckt})					
Constant	-182.32	-153.57	-177.40	-145.44	-185.31
	(71.33)	(55.04)	(54.18)	(81.73)	(31.68)
GDP (in logs)	1.26	1.37	1.64	1.00	2.69
	(0.12)	(0.08)	(0.05)	(0.04)	(0.10)
Population (in logs)	-0.30	-0.86	-0.35	-0.91	-1.73
	(0.11)	(0.06)	(0.05)	(0.02)	(0.10)
Language		1.52			0.90
		(0.17)			(0.63)
GDPR			-0.84		0.22
			(0.11)		(0.20)
Trade agreement				3.27	4.59
				(0.07)	(0.21)
Time trend	0.36	0.32	0.36	0.33	0.37
	(0.01)	(0.01)	(0.01)	(0.004)	(0.01)
2010-2017	-0.47	-0.50	-0.57	-0.46	-0.82
	(0.03)	(0.02)	(0.02)	(0.01)	(0.04)
2018-2020	-0.80	-0.86	-0.71	-0.74	-1.31
	(0.06)	(0.04)	(0.05)	(0.02)	(0.07)
Cable usage (Z_{pckt})					
Constant	126.15	105.77	101.42	117.92	97.02
	(70.21)	(55.78)	(55.23)	(81.11)	(32.09)
Path length (in logs)	-1.82	-1.89	-1.75	-2.39	-2.05
	(0.09)	(0.05)	(0.07)	(0.06)	(0.10)
Potential capacity (in logs)	1.16	1.14	1.50	1.07	1.27
	(0.04)	(0.01)	(0.03)	(0.02)	(0.03)
Number of cables	0.12	0.15	0.14	0.07	0.14
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)

Notes: This table reports estimates for θ^D, the parameters that govern how country-pair characteristics (x_{ckt}) affect demand for data flows between the two countries and estimates for θ^δ, the parameters that govern how cable features of a path (Z_{pckt}) affect the share of the data served by that path. The variable 'GDP' is computed as the sum of logged GDP for two countries in the country pair, and the variable 'Population' is constructed similarly. The vector x_{ckt} also includes a linear time trend, indicators for the 2010-2017 and 2018-2020 periods, and market fixed effects. The 2002-2009 period is the omitted category. The path length is computed as the sum of cable lengths over markets involved in the path. We take the minimum capacity over markets as the potential capacity of a path and the capacity-weighted average as the number of cables. The sample includes 179,778 country-pair-year observations and 1,710,964 country-pair-path-year observations, as well as 399 market-year and 133 island-year observations of used bandwidth. The estimation uses 188 moments in total.

is, suppose market m refers to region-pair gf. Let $\mathcal{C}_g$ be the set of countries in region g. We define:

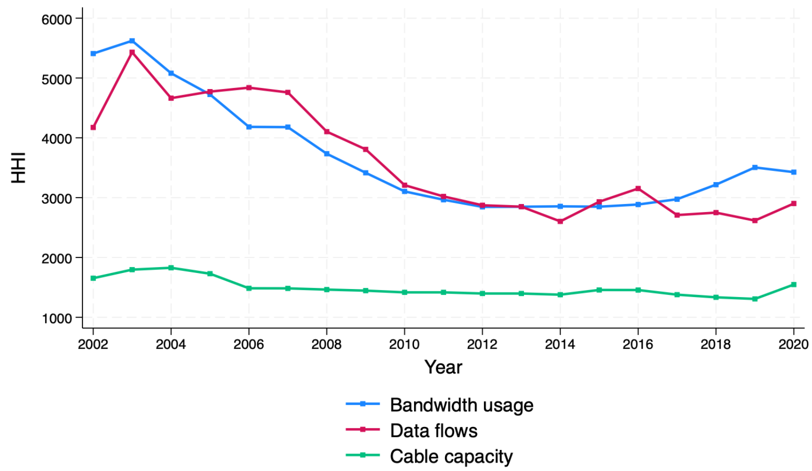
$$D^*_{mt} = \sum_{c \in \mathcal{C}_f} \sum_{k \in \mathcal{C}_g} d_{ckt} \tag{6}$$

The statistic D^*_{mt} reflects endpoint-to-endpoint exchanges between regions, unlike $D_{mt}(\theta)$ (our model prediction of bandwidth usage) or $\tilde{D}_{mt}$ (bandwidth usage observed in the data). The total level of trade in data in t is $\bar{D}^*_t = \sum_m D^*_{mt}$ and the geographic concentration (HHI) of trade in data is $\sum_m (D^*_{mt}/\bar{D}^*_t)^2$.

Figure 3 plots the concentration of raw bandwidth usage from Figure 2 along with the concentration of trade in data. We also include the concentration of cable capacity. Data flows are slightly more concentrated compared to raw bandwidth usage from 2006 to 2009, but become less concentrated than bandwidth usage from 2017 to 2020. This reversal can be explained by the fact that, despite the growth in demand for internet usage in markets that traditionally have small demand, cable investment continues to be concentrated in large markets. We earlier referred to the uptick in concentration of raw bandwidth usage as surprising but in this figure, we show that the uptick disappears when looking at the concentration of trade in data. That is, the uptick is an artifact of raw bandwidth usage and does not represent endpoint-to-endpoint usage. Raw bandwidth usage is concentrated in larger markets because underlying capacity is. As shown in Figure 3, the HHI of cable capacity is relatively constant throughout our sample period despite the reduction in the concentration of bandwidth usage and data flows, and even shows a slight uptick in 2020.

In order to better understand the concentration of data flows and the connectivity of regions, we visualize the network of data flows in Figure 4 using the widely used algorithm of Fruchterman and Reingold (1991). We use D^*_{mt} for the edge values in the algorithm and also for the width of the edges in the figure. The nodes that have a stronger connection (e.g., higher data flows) are more closely positioned. In 2002-2009, three regions–North America, East Asia, and Europe–are clustered in the center with especially strong connections between North America and Europe and between North America and East Asia. South America is also close to the middle via its connection with North America, with Oceania, Sub-Saharan Africa, and South Asia on the periphery. In 2018-2020, the network is much less clustered. North America,

Figure 3: The concentration of trade in data and raw bandwidth usage



Notes: This figure shows the geographic HHI of raw bandwidth usage, data flows, and cable capacity. The HHI is computed as the sum of the squared share of trade flows (bandwidth usage or cable capacity) over markets (region pairs).

East Asia, and Europe are still in the center, but relatively further away from each other. The connection between North America and South America becomes stronger as well.[12]

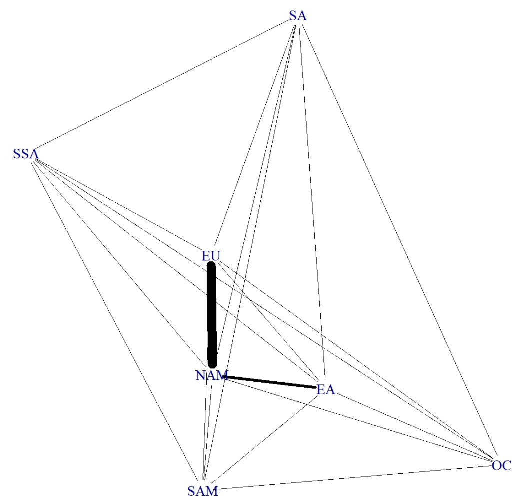
Because the data flows between the top regions are so much higher than the rest, we cannot see the difference in data flows between the smaller regions in Figure 4. The other lines all appear to be of the same width. However, the placement of the regions is informative about which regions they are most closely connected with. Focusing on panel (b), the later period, South America is most closely connected to North America, Sub-Saharan Africa is most closely connected to Europe followed by East Asia, and Oceania is most closely connected to East Asia followed by North America. Interestingly, South Asia is more closely connected to Europe and North America than East Asia.

5 Comparing Trade in Data to Trade in Goods

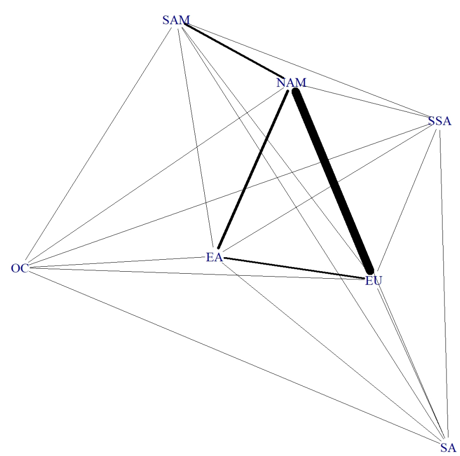
In this section, we draw comparisons between trade in goods and trade in data. We first contrast measures of concentration at the global and regional levels using HHI

[12]To compute the edge values, we sum D^*_{mt} over years, for instance for years $t = 2002, \ldots, 2009$.

Figure 4: The network graph of data flows by period



(a) 2002-2009



(b) 2018-2020

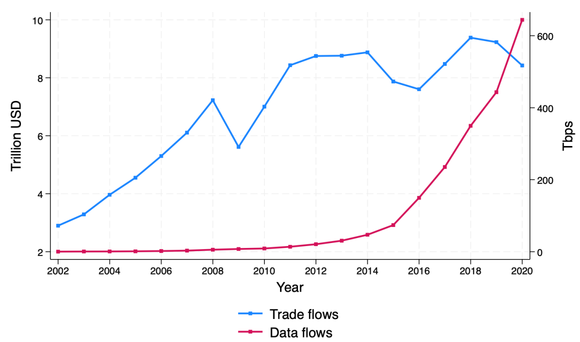
Notes: This plot visualizes the network of data flows between regions using the Fruchterman and Reingold algorithm. Panel (a) is from 2002 to 2009 and panel (b) is from 2018 to 2020. Abbreviations: EA: East Asia; EU: Europe; OC: Oceania; NAM: North America; SA: South Asia; SAM: South America; SSA: Sub-Saharan Africa.

and eigenvector centrality. Then, motivated by literature on trade frictions, we study a comparative static with respect to cable infrastructure to understand the role of infrastructure on trade in data.

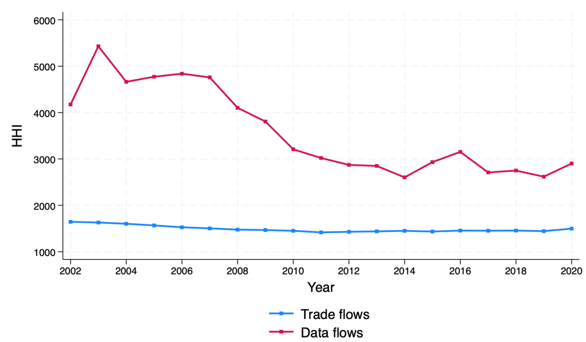
5.1 Concentration in trade

Figure 5 shows the trends in the total level and concentration of trade in goods and trade in data. Similar to raw bandwidth usage that we used in Figure 2, we find that trade in data is growing and becoming geographically dispersed at a much more rapid pace compared to trade in goods.

Figure 5: The level and concentration of trade in goods and data



(a) Level



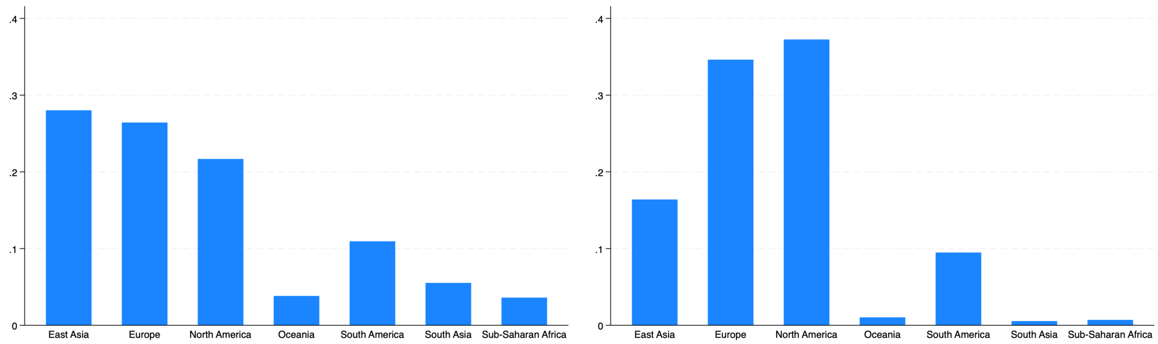
(b) HHI

Notes: Panel (a) shows the total volume of interregional trade flows and data flows from 2002 to 2020. The HHI for interregional trade flows (data flows) in Panel (b) is computed as the sum of the squared share of trade flows (data flows) over markets.

We plot the eigenvector centrality of the seven regions based on trade flows and data flows using our 2018-2020 estimates in Figure 6. There are several noteworthy differences in the centrality of regions between trade in goods and trade in data. The differences in centrality between the most central and least central regions are much higher in data flows than trade flows. East Asia, South Asia, and Sub-Saharan Africa are much more central to the network of trade flows than that of data flows.

To make a comparison of the geographical patterns of trade in goods and trade in data at the market level, we list the top 10 markets for trade in goods in Panel A of Table 3 and for trade in data in Panel B. Focusing on the later period, the largest market for trade in goods is East Asia-Europe, while it ranks fourth in terms of trade

Figure 6: The centrality of the trade network



(a) Trade flows

(b) Data flows

Notes: This figure plots the eigenvector centrality of trade flows and data flows in 2018-2020. The centrality measures are normalized such that they sum to one in each panel.

in data. Europe-North America is ranked first for trade in data and dominates other markets accounting for over half of all interregional data flows, while it is ranked third for trade flows accounting for only 13% of total trade in goods.

5.2 The Role of Cable Infrastructure

Having studied the geographical patterns in trade in data, we seek to understand factors that drive these patterns. To do so, we compute the HHI of data flows over markets based on our model estimates for year 2012 as described in Section 4.3. We compare it to three alternative cases. In the first case, we change the quality of cable connections (v_{ckt}) to the level in 2020 while holding variables in x_{ckt}, such as GDP and population, at the observed level in the data. In the second, we change x_{ckt} to the 2020 level, while holding cable quality at the observed level. In the third, we change both x_{ckt} and v_{ckt} to the level in 2020, simply recovering d_{ckt} in 2020. As shown in Figure 7, the concentration increases by 19.8% when given the 2020 cable network and falls by 7.3% when given the 2020 level of demographics. These results suggest that secular trends in demand for trade in data and the cable network work in opposite directions on concentration, with the evolution of secular trends leading to a more disperse network and the cable network leading to a more concentrated network.

Lastly, we use our estimates to study trade frictions in our context and the role

Table 3: Top 10 markets for interregional trade in goods and trade in data

Panel A: Top 10 markets for trade in goods

2002-2009			2018-2020		
Market	Volume ($B)	Share (%)	Market	Volume ($B)	Share (%)
East Asia - Europe	1294.41	26.59	East Asia - Europe	2524.82	28.01
East Asia - North America	919.50	18.89	East Asia - North America	1546.60	17.16
Europe - North America	807.52	16.59	Europe - North America	1210.25	13.43
North America - South America	495.88	10.18	North America - South America	832.14	9.23
Europe - South America	211.05	4.33	East Asia - South America	561.40	6.23
East Asia - South America	190.61	3.91	East Asia - Oceania	397.36	4.41
East Asia - Oceania	166.33	3.42	Europe - South Asia	373.98	4.15
Europe - South Asia	165.53	3.40	Europe - South America	341.34	3.79
Europe - Sub-Saharan Africa	153.20	3.15	East Asia - South Asia	309.20	3.43
East Asia - South Asia	101.69	2.09	Europe - Sub-Saharan Africa	259.90	2.88

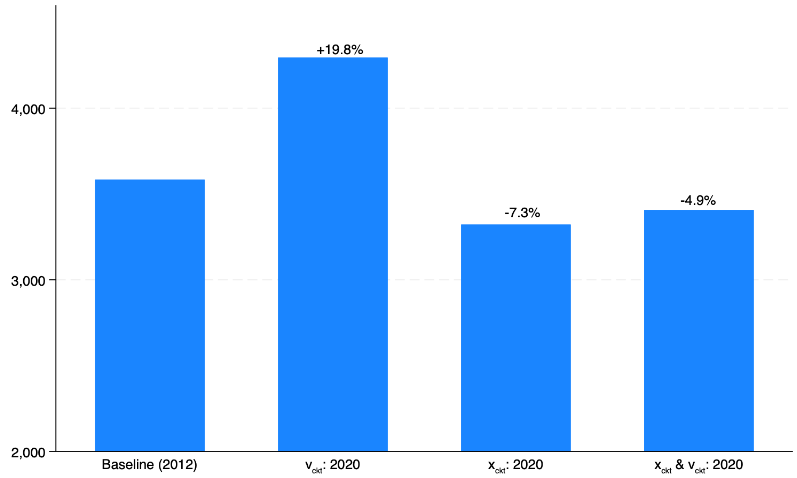
Panel B: Top 10 markets for trade in data

2002-2009			2018-2020		
Market	Volume (Gbps)	Share (%)	Market	Volume (Gbps)	Share (%)
Europe - North America	1,533.20	66.97	Europe - North America	32,028.28	49.53
East Asia - North America	546.33	23.86	East Asia - North America	17,800.72	27.53
North America - South America	143.38	6.26	North America - South America	8,314.54	12.86
East Asia - Europe	31.00	1.35	East Asia - Europe	3,401.98	5.26
East Asia - Oceania	20.05	0.88	Europe - Sub-Saharan Africa	894.68	1.38
East Asia - South America	9.23	0.40	East Asia - South America	889.30	1.38
Europe - Sub-Saharan Africa	5.12	0.22	East Asia - Oceania	880.63	1.36
Europe - South Asia	1.10	0.05	Europe - South Asia	436.77	0.68
South America - Sub-Saharan Africa	0.04	0.00	South Asia - Sub-Saharan Africa	6.63	0.01
South Asia - Sub-Saharan Africa	0.03	0.00	South America - Sub-Saharan Africa	5.44	0.01

Notes: Panel A reports the top 10 markets in terms of the average yearly volume of trade in goods and Panel B in terms of trade in data. We exclude intraregional trade when computing the trade flows in goods and data.

of the cable infrastructure by quantifying the elasticity of trade in data with respect to path capacity. Although not a perfect analogy, it is interesting to compare our es-

Figure 7: The decomposition of concentration in data flows



Notes: The first bar of the plot is the baseline and shows the HHI of data flows computed based on our model estimates for year 2012, holding country-pair characteristics (x_{ckt}) and quality of connection (v_{ckt}) at the observed levels. The HHI is computed as the sum of the squared share of data flows over markets (region pairs). The second to fourth bars plot HHI under the following three counterfactuals: (i) holding v_{ckt} at the 2020 level and x_{ckt} at the observed level in 2012; (ii) holding x_{ckt} at the 2020 level and v_{ckt} at the observed level in 2012; and (iii) holding x_{ckt} and v_{ckt} at the 2020 levels. The number on top of each bar shows the percent difference from the baseline.

timates to the elasticity of trade in goods to trade frictions found in other papers, a long-term focus of the literature on international trade and transportation networks. We compute the elasticity of trade in data with respect to path capacity using the estimates in Column (1) of Table 2 by quantifying the change in data flows that would follow a one-percent increase in the capacity of paths between each country pair. Increasing the path capacity of a path has two effects. It increases the overall cable quality term (v_{ckt}), driving up data flows. It also redirects data flows among the available paths. We interpret markets with higher elasticity as being more constrained by cable availability or facing higher trade frictions.

We calculate the elasticity of D^*_{mt}, the trade flows in country-pairs directly connected by market m, to a one-percent change in capacity on the paths between those countries. We calculate elasticity in two ways. First, we calculate the elasticity to an increase in capacity only on the direct path, and second, to an increase in capacity on all paths connecting those countries. In Table 2, the result for the direct path appears in Column 1 and ranges from 0 to 1.1%. It is zero for six markets that do not

Table 4: The elasticity of trade in demand to path capacity

	(1) Direct path	(2) All paths
Effect on D^*_{mt}		
Europe - North America	0.45	0.47
East Asia - North America	0.07	0.38
North America - South America	0.62	0.62
Europe - South Asia	1.01	1.13
East Asia - South Asia	0.43	1.16
East Asia - Europe	0.17	1.03
East Asia - Oceania	0.12	0.74
Europe - Sub-Saharan Africa	1.08	1.15
North America - Oceania	0.11	1.16
South America - Sub-Saharan Africa	0.24	1.16
South Asia - Sub-Saharan Africa	0.13	1.16
Europe - Oceania	0.01	1.16
Oceania - South Asia	0.02	1.16
East Asia - Sub-Saharan Africa	0.00	1.16
Europe - South America	0.00	1.16
East Asia - South America	0.00	1.05
North America - South Asia	0.00	1.16
North America - Sub-Saharan Africa	0.00	1.16
Oceania - South America	0.00	1.16
Oceania - Sub-Saharan Africa	0.00	1.16
South America - South Asia	0.00	1.16
Effect on $\sum_m D^*_{mt}$	0.37	0.56

Notes: This table reports the elasticity of data flows with respect to path capacity using the 2020 data. Column (1) shows the percent change in the data flows between country pairs in market m (D^*_{mt}) and in the total data flows between all country pairs ($\sum_m D^*_{mt}$) from increasing the capacity of the direct path for market m by one percent. Column (2) shows results from similar calculations for increasing the capacity of all available paths. The markets are sorted based on the 2020 capacity level in descending order.

have direct cables and close to zero for markets that have little direct capacity and rely almost exclusively on indirect paths, such as East Asia-Sub-Saharan Africa and Europe-South America.

In Column 2, we see the effect in market m of changing the capacity of all paths. In our setup, the differences across markets are driven by the quantity of trade flows, but the comparison between the two columns is instructive.[13] The elasticity to all

[13]For country pairs with very low trade in data, the elasticity of quantity to v_{ckt} is 1, so the elasticity

paths is always greater than the elasticity to just the direct path. The two columns are close to each other for cases in which the country pairs are heavily reliant on the direct path, such as North America to South America, and different for cases in which they use indirect paths, such as East Asia to Europe.

We can also calculate the elasticity of world data flows. The world elasticity to changing all of the direct paths is 0.37, whereas the world elasticity to changing all paths is 0.56, about 50% higher.[14] This difference highlights the importance of indirect paths in our study. The elasticities we obtain for the majority of markets are in line with various measures of the elasticity of trade flows to trade costs or distance computed in other contexts, and slightly lower for the high-capacity markets. For example, Brancaccio et al. (2020) recover the elasticity of trade with respect to shipping prices of -1, and Overman, Redding and Venables (2003) state that the elasticity of trade to the distance between two countries tends to fall in the range of -0.9 to -1.5. Our results overall indicate that the subsea cable infrastructure has a substantial impact on trade in data, and places particularly strong constraints on countries with limited cable availability.

6 Conclusion

Despite the dramatic growth in international flows of data and the critical role it plays in supporting the global economy, there has been little research in this area. Our paper seeks to understand international trade through a new lens using data on the usage of the subsea cable network and compare trade in data with trade in traditional goods. Our main empirical challenge is that our observations of bandwidth usage between regions include traffic that is simply traveling through those regions. The bandwidth usage measure at the country level suffers from a similar issue. This motivates us to develop a model of country-to-country demand for data and how the data flows are allocated on the cable network. We estimate the model using data on the subsea cable industry and find that country-level variables, such as GDP and population, as well as country-pair variables such as the presence of a trade agreement, are strong predictors of demand. Moreover, the capacity, distance, and

with respect to capacity the coefficient on capacity from Table 2, which is 1.16. Thus, we see several rows in Column 2 with this value.

[14]We find similar results when changing the distance of paths, instead of the capacity.

number of cables of the cable path are important in determining the quality of cable connections and cable usage.

Our model yields predictions about the endpoint-to-endpoint data flows between countries. We show that trade in data is much more concentrated than trade in goods. Regions like East Asia are relatively less central to the network of data flows compared to the network of trade in goods, while regions such as South Asia and Sub-Saharan Africa are also far more peripheral. We find that secular trends in demand for trade in data and the cable network work in opposite directions on concentration: while the underlying demand is becoming less concentrated, the continued concentration in cable capacity is leading to a more concentrated network. We show that trade in data is more constrained in markets with limited cable provision and that indirect cable routes play an important role in supporting trade in data in these markets.

References

Allen, Treb and Costas Arkolakis, "The Welfare Effects of Transportation InfrastructureImprovements," *Review of Economic Studies*, 2022, *89*, 2911–2957.

Anderson, James E, "The gravity model," *Annu. Rev. Econ.*, 2011, *3* (1), 133–160.

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papgeorgiou, "Geography, Transportation, and Endogenous Trade Costs," *Econometrica*, 2020, *88*, 657–691.

Caoui, El Hadi and Andrew Steck, "Network Diversity, Market Entry, and the Global Internet Backbone," *SSRN Electronic Journal*, 2025.

Cariolle, Joël, Georges Vivien Houngbonon, Tarna Silue, and Davide Strusani, "The Impact of Submarine Cables on Internet Access Price, and the Role of Competition and Regulation," Technical Report, The World Bank 2024.

Centre d'Etudes Prospectives et d'Informations Internationales, "Gravity Database," https://www.cepii.fr/cepii/en/bdd_modele/bdd_modele_item.asp?id=8 2002-2020. Accessed: 2022-03-22.

Eichengreen, Barry, Romain Lafarguette, Arnaud Mehl, and Massimo Ferrari Minesso, "Technology and the Geography of the Foreign Exchange Market," *Journal of International Money and Finance*, 2023, *131*, 102802.

Fajgelbaum, Pablo D. and Edouard Schaal, "Optimal Transport Networks in Spatial Equilibrium," *Econometrica*, 2020, *88* (4), 1411–1452.

Fang, Chao, Xuepeng Tang, Linhui Yu, and Junsen Zhang, "Connecting to Trade: Submarine Optical Cables and Chinese Firms' Exports," *Available at SSRN 5254120*, 2025.

Fréchette, Guillaume R., Alessandro Lizzeri, and Tobias Salz, "Frictions in a Competitive, Regulated Market: Evidence from Taxis," *American Economic Review*, 2019, *109*, 2954–2992.

Fruchterman, T.M.J. and E.M. Reingold, "Graph drawing by force-directed placement," *Journal of Software: Practice and Experience*, 1991, *21*, 1129–1164.

Head, Keith and Thierry Mayer, "Gravity equations: Workhorse, toolkit, and cookbook," in "Handbook of international economics," Vol. 4, Elsevier, 2014, pp. 131–195.

Hjort, Jonas and Jonas Poulsen, "The Arrival of Fast Internet and Employment in Africa," *American Economic Review*, 2019, *109*, 1032–1079.

Jeon, Jihye and Marc Rysman, "Investment and Usage of the Subsea Internet Cable Network," *Available at SSRN 4925783*, 2025.

Keller, Wolfgang and Stephen Ross Yeaple, "The Gravity of Knowledge," *American Economic Review*, 2014, *103*, 1414–1444.

Krings, G, F Calabrese, C Ratti, and V Blondel, "A gravity model for inter-city telephone communication networks," *Journal of Statistical Mechanics: Theory and Experiment*, 2009, *7003.*

Lee, Robin, "Vertical Integration and Exclusivity in Platform and Two-Sided Markets," *American Economic Review*, 2013, *103*, 2960–3000.

M., P. Cotterlaz Conte and T. Mayer, "The CEPII Gravity database," CEPII Working Paper N°2022-05 July 2022.

Mauldin, Alan, "Over 99% of Intercontinental Data Traffic?," https://blog.telegeography.com/2023-mythbusting-part-3 May 2023.

Mueller, Milton and Karl Grindal, "Data flows and the digital economy: information as a mobile factor of production," *Digital Policy, Regulation and Governance*, 2019, *21* (1), 71–87.

Overman, Henry G, Stephen Redding, and Anthony J Venables, "The economic geography of trade, production, and income: a survey of empirics," *Handbook of international trade*, 2003, pp. 350–387.

Redding, Stephen J and Esteban Rossi-Hansberg, "Quantitative spatial economics," *Annual Review of Economics*, 2017, *9* (1), 21–58.

TeleGeography, "Global Bandwidth Research Service," https://www.telegeography.com/transport-networks 2002-2020. Accessed: 2020-05-28.

A Additional Figures and Tables

Table A1: Trade flows and international bandwidth usage

Panel A: Trade flows from 2002 to 2020

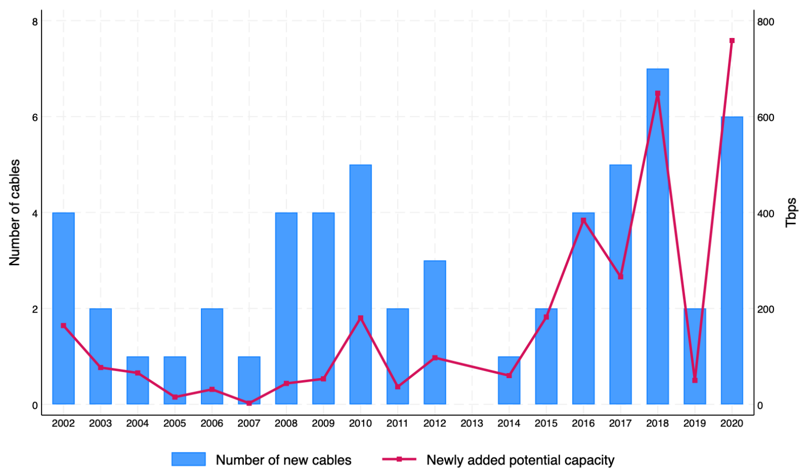
Year	Total trade volume (in trillion USD)	Number of countries making up 50% of trade flow	Number of countries making up 90% of trade flow	HHI
2002	6.51	8	36	506.36
2003	7.63	8	37	479.73
2004	9.25	8	38	462.95
2005	10.45	9	39	450.19
2006	12.00	9	40	436.86
2007	13.89	9	42	422.30
2008	16.07	10	44	399.92
2009	12.36	9	43	413.82
2010	15.01	10	43	417.51
2011	17.82	10	44	404.08
2012	18.11	10	45	404.27
2013	18.33	10	45	408.41
2014	18.46	10	45	420.48
2015	16.19	9	42	447.33
2016	15.81	9	41	450.55
2017	17.55	10	42	442.43
2018	19.38	10	42	440.28
2019	18.90	10	42	440.82
2020	17.43	10	41	461.20

Panel B: International bandwidth usage from 2002 to 2020

Year	Total international bandwidth usage	Number of countries making up 50% of usage	Number of countries making up 90% of usage	HHI
2002	1.49	4	14	1076.32
2003	2.41	3	14	1141.85
2004	3.41	4	15	1066.20
2005	5.00	4	17	1041.90
2006	7.00	4	20	978.47
2007	11.49	4	20	970.06
2008	19.22	4	25	860.42
2009	30.54	4	26	798.88
2010	46.68	5	28	777.18
2011	69.99	5	28	753.24
2012	101.05	5	29	739.92
2013	145.44	5	29	719.61
2014	212.08	5	28	720.43
2015	299.61	5	28	721.82
2016	444.09	5	28	718.06
2017	667.58	5	27	730.33
2018	997.90	5	26	736.50
2019	1,459.39	5	24	767.32
2020	2,074.88	5	24	719.16

Notes: Panel A reports the level and concentration measures of country-level trade flows. Panel B reports the level and concentration measures of country-level bandwidth usage. All figures include intraregional trade. The HHI is computed as the sum of the squared share of trade flows (bandwidth usage) over countries.

Figure A1: Subsea cable investment



Notes: This figure shows the total number and capacity of new cables by year.