International Trade in Data: Investment and Usage of the Subsea Internet Cable Network*

Jihye Jeon† and Marc Rysman‡

November 2023

Abstract

This paper studies the construction and use of undersea internet cables, a critical piece of communication infrastructure. We view traffic on these cables as *international trade in data* and we contrast patterns of trade in data with international trade in goods. We propose a model in which country-to-country trade in data, similar to a gravity equation model, traverses the observed cable network leading to interregional flows observed in our data. We estimate the model via GMM and use the results to decompose growth in global internet usage into growth in demographic variables and improvements in the cable network. We find that improvements in the cable network are an important contributor on par with growth in demand and time trends. For the supply side, we estimate a model of cable construction in which consortiums of investors join together for profitable construction opportunities recognizing the impact of their investment in any one market on global internet flows. Our results highlight the importance of business stealing in leading private incentives for investors to deviate from globally optimal choices.

*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

Communication technology is critical to modern trade and economic activity. A central piece of communication infrastructure 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. Clark (2018, pg. 24) writes about fiber optics: "The economic implications of its development and deployment are perhaps the single most important factor in the growth and success of the internet." Growth has indeed been impressive; global internet traffic grew by a factor of 1000 between 2002 and 2020, far surpassing the growth of traditional trade of goods and services.

We study the development, growth, and usage of the subsea internet cable network to shed light on this understudied, yet important market. We leverage new data on the subsea cable network to study two main research questions. First, 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. We develop an analog of a gravity model for trade in data that accounts for both country-to-country demand to exchange data and the quality of the network that connects countries. We decompose the relative contribution of these two factors in the evolution of trade in data.

Second, we study investment decisions to build new cables. A new cable in one region can affect traffic flows globally. The networked nature of this product means that investors' private incentives can deviate substantially from the social optimum. In addition, the subsea internet cable market is subject to a very low level of economic regulation. As a result, the network we observe is unlikely to exhibit global efficiency. We develop a model that highlights the international implications of local investments and the role for inefficient business stealing incentives in investment decisions. Furthermore, whereas cables have traditionally been owned by telecommunication carriers such as AT&T and British Telecom, recent construction has been dominated by content providers such as Google and Microsoft. We allow parameters in the objective function of owners to differ between telecoms and content providers to study how their incentives differ.

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. We observe landing points and construction dates of all cables, and typically more information such as bandwidth capacity, length of the cable, and owners and ownership shares. Critically, we observe used bandwidth, which is a measure of usage of a cable that we further describe below. We observe used bandwidth between some markets and also at the level of the individual countries. Our current data set goes from 2002 to 2020. We supplement the data with various sources, such as data on internet penetration from the International Telecommunication Union

and trade data from CEPII.

We face an empirical challenge in that we observe a measure of how many data flow between two regions, but we do not observe endpoint-to-endpoint usage. That is, data flowing between two countries may have different starting and endpoint countries, and data traveling between two endpoint countries often take multiple routes as determined by various factors, such as cable availability, capacity, and length. To address this issue, we develop a structural model that maps country-to-country demand for data into observed data flows. We model bilateral data demand through a gravity equation in which demand is a function of characteristics such as the GDP and population of the two countries. The cable network determines the set of available paths that data may take between two countries and the features of those paths, which determines the quality of the connection. We model the allocation of demand for data across these different potential paths. In this way, the existing cable network determines which cable paths data travel on and thus, the observed usage of the cables.

Our model allows us to infer country-to-country trade in data. We compare and contrast it with traditional trade in goods. We characterize the concentration of trade in data in individual countries and regions, how and where it has grown, and how the construction of new cables has affected that. We draw measures from the network statistics literature, particularly eigenvector centrality, to evaluate the centrality of individual countries in data flows and contrast that with the centrality in trade in goods. We decompose changes in these various measures of trade in data into changes in demographics and changes in the quality of the cable network.

On the supply side, we estimate a model of cable construction using a moment inequalities approach. In each period, opportunities to build new cables in markets appear exogenously, and investors choose whether to join consortiums that can take advantage of these opportunities. Investor decision-making depends on demand and the state of rival cable ownership and construction. Firms profit based on their ownership share of cables and the amount of data that traverses those cables. Our model highlights the business-stealing nature of cable construction. In our model, new cables on busy routes, which typically have substantial previous investment, often have high private value relative to social value as new owners steal business from existing cables. In contrast, cables on less busy routes, which typically have lower investment levels, are the reverse: the cable may have high social value but relatively low demand.

In our results, we first document that the usage of the internet bandwidth is more highly concentrated geographically compared to trade in traditional goods. We then estimate our demand model, drawing moments from the variation in usage across continents and years. We also exploit "island countries," for which all communication must traverse subsea cables. 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 4 percent increase in trade in data. Further, the path length has a negative effect and the path capacity has a positive effect on the amount of data flows allocated to the path.

Next, we document the slow arrival of cables in certain regions in contrast to repeated investments in enormous quantities in others. In the supply model, we use the demand model to pin down quantities with and without investment. Combined with construction costs, which we observe in our data, we estimate markups (that is, variable profit) utilizing moment inequalities. We find a 95% confidence interval for the variable profit between 2.8 and 4.4 million dollars from an increase in data flows served by a firm by one 1 Tbps. We do not find that the markups are significantly different for content providers.

Taking the estimates together, our counterfactuals consider a de novo entrant adding a new cable in each market to evaluate the efficiency in the cable network. We compute the change in the combined profits of incumbent firms, decomposing it into the change in the market where the investment happens and the change in the rest of the world. It is ambiguous how a new cable construction would affect incumbent profits. First, the new cable construction would shift market shares from incumbents to the entrant in the market with the investment, decreasing the incumbent profit in that market. Second, the increased capacity would attract more data flows into cables in that market, shifting down the incumbent profit in the other markets. Lastly, the increased capacity would have a market expansion effect, some of which would be served by cables in that market and some by cables in the other markets, raising incumbent profits in all markets.

We find that the effect of a new cable indeed varies widely across markets. In 2010, building a new cable would have a negative effect on incumbent profits for many of the markets that involve North America, which already possess large capacity. This suggests that substantial business stealing incentives exist in these markets. By contrast, incumbents would experience profit increases in many of the markets involving Sub-Saharan Africa, for example, suggesting that a new cable would have substantial market expansion effects in these markets.

We quantify the social value of a new cable as the total increase in worldwide data flows and the private value as the total increase in the data flows served by the entrant. We find that markets such as Sub-Saharan Africa-East Asia and Sub-Saharan Africa-South America have high social values relative to private values in 2010. This is striking given that there was only one cable connecting Sub-Saharan Africa and South America as of 2010 and no cable connecting Sub-Saharan Africa and East Asia, and implies a misallocation of investments. Markets with high capacity in 2010 that continued to receive large investments after 2010 such as North America-South America and North America-East Asia have low social gains from a new cable.

Related Literature

There is little economic research on subsea cables market. Hjort and Poulsen (2019) show that the arrival of subsea cables to Africa improved internet service for consumers living along the relevant terrestrial network and led to increased productivity and high-skilled employment. Their paper highlights the economic impact of internet access driven by the expansion of the

subsea cable network, but does not explore when or why cables are constructed, whereas our work focuses on the market for subsea cables itself. The closest paper that we are aware of is the concurrent paper by Caoui and Steck (2023), which focuses on bilateral incentives to build subsea cables and the role of the number of cables. The main difference from their paper is that our model takes into account the fact that all regions are globally connected, which means that demand for communication in one market can be transmitted through cables in other markets and new construction in one market can affect data flows in others. This is a key feature not only in this setting as well as other settings such as transportation markets (e.g. shipping ports, airports, and railroads), power and energy industries, and other telecommunication markets (e.g. data centers). More broadly, our paper relates to papers that study infrastructure that supports the internet such as Greenstein (2015).

Our paper also connects to papers on the interconnection between the internet and trade. Steinwender (2018) studies how advances in telecommunications technology affect trade. Blum and Goldfarb (2006) examines whether the law of gravity holds in the case of digital goods sold over the internet.

Our approach is related to equilibrium models of transportation and cargo networks, such as Fréchette et al. (2019) and Brancaccio et al. (2020). Our research is also related to empirical research on the construction of networks in retail, such as Jia (2008), Holmes (2011), and Houde et al. (2022), and in airlines, such as Aguirregabiria and Ho (2010) and Li et al. (2021). Our research also relates to papers on investment in trade-related infrastructure such as Jeon (2022) and Brancaccio et al. (2020) and in internet services markets such as Rysman (2016).

Lastly, we contribute to a growing empirical literature that employs moment inequalities, recently reviewed by Kline et al. (2021) and Canay et al. (2023). Our paper is especially closely related to papers that estimate an endogenous entry or product choice model using inequaliites such as Eizenberg (2014), Wollmann (2018), Dickstein and Morales (2018), and Fan and Yang (2022).

The rest of the paper proceeds as follows. Section 2 provides background on the subsea cables market, information about our data, and motivating facts. Section 3 describes the model of demand for international data, cable usage, and cable construction. Section 4 discusses our estimation strategy and results. Section 5 discusses our analysis based on demand-side estimates. Section 6 presents results from counterfactual exercises. Section 7 concludes.

2 Subsea Cable Networks and Data

2.1 Subsea Cable Networks

As of 2021, there are over 400 active subsea cables spanning 1.5 million kilometers in length, connecting 182 countries across all continents except Antarctica. Industry experts estimate that the undersea telecommunication cable network carries about 95% of intercontinental global in-

ternet traffic, and 99% of transoceanic digital communications such as voice and data including trillions in international financial transactions daily (CRS, 2022). The core of a cable consists of strands of glass fibers through which lasers can propel data, which is surrounded by additional layers that are meant to provide protection and transfer power. A cable is buried on the ocean floor by specialized cable-laying ships through a long and complex process.

Typically, cables are owned by a consortium of owners, including telecommunication companies, governments, and other businesses. Historically, telecommunication providers such as AT&T, British Telecom, and NTT (of Japan) were the main owners of the cables. Starting around 2017, an overwhelming share of investment has been by tech firms or content providers such as Google and Microsoft that are seeking to meet the growing demand for data exchanges and more control over ownership and usage.

The lack of economic regulation is a striking feature of the undersea internet cable market. There is some government intervention in regard to national security, but that has little impact on the features of the market we consider. For example, in the United States, all subsea cable landings must be approved by the Federal Communication Commission (FCC). Historically, this power was used to ensure reciprocal acceptance of cables owned by US entities in foreign countries. However, the FCC has not used its powers for this purpose in several decades. Rather, the FCC refers applications to Team Telecom, an ad hoc committee of interested federal agencies.[1] The Department of Homeland Security often takes the lead in these instances, particularly monitoring landings for their potential for terrorist attacks, as well as the involvement of Chinese owners and equipment makers. We regard this form of regulation as non-economic. While these national security considerations may affect the way a landing point is constructed, the exact location of a landing point within a broader region, and the choice of equipment, these are not features we study. Thus, we model the market as unregulated.[2]

2.2 Data

The main data we use in this paper are from Telegeography, 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

[1]The Trump Administration turned Team Telecom into a formal committee called the Committee for the Assessment of Foreign Participation in the United States Telecommunications Services Sector in 2020. Mergers in which both merging firms have holdings of subsea internet cables, such as the merger of Level3 and Global Crossing, involve economic analysis of this market by the FCC. However, we are not aware of merging parties facing issues with their subsea holdings.

[2]Much of the information in this paragraph is drawn from an interview of a knowledgeable economist at the FCC.

in economics making use of these data, other than the concurrent paper by Caoui and Steck (2023). 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. We do not observe terrestrial cables. Figure 1 provides a map from Telegeography of commercial subsea cables in 2023.

We observe several measures of capacity. *Potential capacity* is the capacity of data that a cable could potentially carry. Activating that capacity is costly, and *lit capacity* is a measure of how much of potential capacity is ready for service. Lit capacity is what is available to cable owners to sell. Primary purchasers are local carriers and content providers, although some large enterprises (e.g., governments, large companies) and educational institutions (e.g. CERN) also take a share. *Purchased capacity* is the amount sold to customers. After purchasing, customers choose how much to activate, which requires further investment. We observe this as well, as *used capacity*. Telegeography constructs an equivalent to used capacity for content providers that own and use their own cable. Note that whereas a retail consumer may purchase a certain amount of data communication per month, cable owners and their direct customers always transact in terms of capacity, for example, five wavelengths on a fiber optic cable for five years.[3] Traffic is the amount of data actually transmitted over the cable, but we do not have measures of this. We take used capacity as our measure of usage. In general, purchasers of capacity devote substantial engineering resources to tightly manage the relationship between purchased capacity, used capacity, traffic, and realized quality of communication, so used capacity is likely a good approximation of traffic.

Telegeography offers bandwidth price data at the city-pair-level. We do not use this data in our analysis for several reasons. These city-to-city prices emerge from a market downstream from cables and typically cannot be linked to the use of any particular cable or cables. It is unclear how much traffic between cities pays the city-to-city rate because it represents a specialized product. Furthermore, traffic between two cities that does not have those cities as endpoints would not normally pay this price. In addition, the pricing data reflect the price of the public internet (e.g., prices charged to internet service providers), but the bandwidth usage data includes the quantity used by content providers on their own cables, which do not pay these fees. Lastly, to the extent traffic between cities uses terrestrial networks (which must be the case if cities are inland or do not have a cable landing point), city-to-city prices reflect those costs as well.

Figure 2 shows the remarkable growth in the number of cables worldwide and total potential capacity from 1987 to 2020. Note that the detailed usage data cover only from 2002 to 2020, so our analysis will focus on this period. There is a slowdown in investment starting around 2002, but it accelerates again around 2011. It is often important to distinguish between

[3] A strand of fiber optic wire may carry 100 wavelengths of light. A fiber optic cable typically carries from 4 to 24 wires during our sample period.

the number of cables in an area and the capacity of those cables. For example, Indonesia is a country of islands with many subsea cables connecting many relatively small islands and is a world leader in terms of the number of cables landing on its shores but is only 29th in the world in terms of total capacity of its landings in 2020.

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 communication between countries that primarily use subsea routes. We define *regions* 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 large share of communication goes over terrestrial cables only.[4] We focus our analysis on interregional communication, which must traverse subsea cables.

We divide the world into the following 7 regions: East Asia, Europe, North America, Oceania, South America, South Asia, and Sub-Saharan Africa. Telegegraphy 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, and Western Europe into one region of Europe based on the observation that there are many viable terrestrial connections across the four regions. By contrast, there is almost no terrestrial communication through the Himalaya 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.

Whereas Telegeopraphy reports potential and lit bandwidth at the level of the cable, it reports used bandwidth annually at the level of the region pair.[5] So for instance, we observe the level of used bandwidth between North America and Europe. Used bandwidth is not directional. We observe a single used bandwidth number for the region pair, the total of used bandwidth going in both directions.[6] We observe used bandwidth between each of the seven regions. However, if there is no cable between two regions, used bandwidth for that region pair is naturally zero. Importantly, used bandwidth reflects bandwidth between the two regions, although used bandwidth may carry traffic that is from further endpoints. For example, there is no cable between Africa and East Asia as of 2023, only a cable that is planned to connect these regions by 2024. This means that carriers route traffic between these regions through intermediate regions, such as Sub-Saharan Africa to South Asia to East Asia. In this case, traffic between Sub-Saharan Africa and East Asia would show up as used bandwidth on these inter-

[4]Even countries with land routes between them may use subsea cables. Building subsea cables is often cheaper than building terrestrial cables and so is an attractive option even when terrestrial cables are feasible. For instance, there are a number of subsea cables running along the coast of Europe (such as between Italy and Greece).

[5]Telegeography often reports data at higher levels of aggregation and in fact refers to what we call "regions" as "sub-regions." Fortunately, Telegeophgraphy was able to provide used bandwidth data at Telegeography's subregion level.

[6]Our understanding is that used bandwidth is almost always sold symmetrically. That is, wavelengths sold in one direction come with the same number of wavelengths in the other direction.

mediate routes.

Telegeography also reports international bandwidth usage for each country, that is, the total used bandwidth to other countries. This variable is reported at the level of the country, not the country pair. Similar to the interregional used bandwidth numbers, used bandwidth for a given country may reflect traffic that traverses through the country, not starting or ending there. Our model makes predictions about total used bandwidth of subsea cables so this variable is potentially useful for us. But this is one of the few Telegeography variables that includes terrestrial cables, which makes it difficult to integrate into our model. However, for island countries, all international usage is on subsea cables. We focus on island countries with substantial international bandwidth usage, of which we identify seven countries: Singapore, Japan, Taiwan, Australia, South Korea, the Phillippines, and New Zealand.[7] Section 4.1 describes how including islands as separate regions provides additional moments in our estimation.[8]

We consider only cables that connect separate subregions. So for instance, we drop cables that connect Italy to Greece or connect among Indonesian islands. Panel A of Table 1 provides information on the cables included in our analysis. There are a total of 105 unique cables, with an average potential capacity of 34,439 Gbps and an average length of 10,498 km.

We define *markets* as pairs of regions (e.g. North America-Europe, North America-South America). Lastly, we define *path* as a set of markets for which there is a cable that connects two regions in the market. For example, possible paths between North America and Europe include {North America-Europe; North America-South America-Europe; North America-South America-Sub-Saharan Africa-Europe; etc}. We refer to the path with just the two endpoints on it as the *direct path*. So, in the previous list, the first element is the direct path and the rest are indirect paths.

An important input into our model is the amount of capacity in each of the paths between two regions in a given year. We construct that as follows. If a cable in our data set has landing points in two markets, we denote that cable as providing capacity in the direct path between those markets. For instance, if a cable has Northern France and New Jersey as its only landing points (the Apollo cable), we denote it as connecting the regions of North America and Europe. To find the capacity on the direct path between two regions, we add up all of the capacity for the cables in that path (that is, the capacity for the cables in that market, since direct

[7] 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.

[8] Even if we did observe country-to-country used bandwidth, it would not be useful for us in our approach. That is because we would not know the beginning and endpoints of the usage and our model does not make predictions about that. For instance, suppose we observed used bandwidth between France and the United States in addition to used bandwidth between Europe and North America. We would know how much of that usage originated elsewhere in Europe rather than France, and because we do not model terrestrial networks (for which we lack data), our model does not make predictions about France-to-US usage. As shown below, our model makes predictions about how much data France and the US communicate (that is, with France and the US as endpoints), how much goes between the US and Europe (potentially with other endpoints) but not how goes between the US and France.

paths correspond to markets). Similarly, the capacity of the Apollo cable would contribute to any indirect path involving Europe and North America. For instance, one of the paths from Southern Africa to North America is South Africa to Europe to North America. The Apollo capacity is included as part of the capacity on the second leg of this path.

Some cables have multiple landing points. If all of the landing points fall in only two regions, we treat the cable as providing capacity on the direct path (as if it had only two landing points). If a cable has landing points in multiple regions, we denote it as providing direct connections between each of the regions. For instance, a cable that goes from East Asia to South Asia to Europe (such as the SeaWeMe cables) provides capacity on the direct path between each pair of regions. Furthermore, when a cable contributes to the direct path between two regions, we do not also record it as contributing to an indirect path. For example, SeaWeMe contributes capacity to the direct path between Europe and South Asia, the direct path between South Asia and East Asia, but not the indirect path Europe-to-South-Asia-to-East Asia. That would be double counting its capacity because SeaWeMe already contributes to the direct path between Europe to East Asia. In this sense, whether capacity appears in a market depends on which pair of regions we are considering data traveling between. SeaWeMe contributes capacity in the Europe-South Asia market if we are considering data sent from France to India but not France to China.

We construct other measures of cable features similarly. For instance, we construct variables such as the distance of a path and the number of cables on a path. We construct path distance by summing the market-level cable lengths for each market in a path. We construct path capacity by taking the minimum capacity of market-level capacities (alternatively, maximum or weighted average) of each market on a path. The number of cables in the path is computed as the capacity-weighted average across markets connected by the path. We take this as a measure of the level of competition on the path.

We supplement Telegeography's data with gravity data from Centre d'Etudes Prospectives et d'Informations Internationales (hereafter CEPII). It provides information used to estimate gravity equations for traditional trade 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 lower than $1 billion and countries with missing GDP information, arriving at a sample of 161 countries.

There are 21 unique markets summarized in Panel C of Table 1. On average, a market is connected by 4.1 cables and 3.1 direct cables, which connect only two regions in the market. 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 by 2020.

2.3 Motivating Facts

Comparing Internet Usage and Trade in Goods

Table 2 shows the overall trend in international bandwidth usage and trade in goods from 2002 to 2020. Total international bandwidth usage grew by a factor of 1000 from 2002 to 2020, while total trade flows grew threefold. In addition, we observe two important patterns in the geographic concentration: (i) trade in data is more geographically concentrated relative to trade in goods, and (ii) the network of data trade is becoming less concentrated at a far more rapid pace than the network of goods trade. In 2002, only 14 countries accounted for 90% of bandwidth usage, while 36 countries accounted for 90% of goods trade. By 2020, the number grew to 24 countries for bandwidth usage and 41 countries for trade flows. We also compute the HHI of goods trade and bandwidth usage. The HHI is relatively stable over this period for the trade flows, whereas it falls substantially from 1076 to 719 for bandwidth usage.

Table 3 shows in which markets trade and bandwidth usage are concentrated. The East Asia-Europe market accounts for the largest share of trade in goods accounting for 23% in 2002 and 28% in 2020. In terms of bandwidth usage, the top market (Europe - North America) accounts for a much larger share: 71% in 2002 and 52% in 2020. Table 4 shows a similar list but for top countries. For trade in goods, the share of the United States fell slightly from 14.7% to 10.7% from 2002 to 2020, while that of China rose sharply from 5.9% to 13%, rising to the first place. For bandwidth usage, the U.S. is on top throughout this period, but there is more movement compared to trade flows. Countries like Ireland, Singapore, Spain, China, and Japan entered the list by 2020, replacing countries like Canada and Denmark.

Investment in Subsea Cables

We show the overall trend in the new construction of cables in panel A of Figure 3. It shows that although the number of new cables in the recent years from 2016 to 2020 is similar to the level in the previous active building periods, the newly added capacity in this period trumps that of previous periods.

As previously mentioned, there is a potential misallocation of cable capacity considering the high social value of connecting under-served regions. For example, we observe large new investments in the Europe-North America market especially in 2016 to 2020 despite the market's already large capacity, while the first South America-Africa cable with substantial capacity does not appear until 2018. Panel B of Figure 3 shows that new investments continue to be concentrated on existing routes that already have high capacity, such as Europe-North America and East Asia-North America. There are also markets that are recently attracting more investment, such as North America-South America and East Asia-Oceania.

We also highlight the recent growth in subsea cable investments by major content providers. Historically, cable construction was dominated by telecommunication carriers, such as AT&T,

British Telecom, and NTT (of Japan), who used cables to serve the public internet. However, the last five years have witnessed a dramatic increase in investment by major content providers. Anecdotally, they are interested in connecting their data centers but telecommunication carriers are unwilling to build such specialized cables. Content providers may lease some bandwidth on their cables to carriers and the cables often have telecom coinvestors, but these cables are oriented, at least in part, towards the private communication needs of content providers.

We observe four content providers with substantial ownership of subsea cables: Google, Facebook, Amazon, and Microsoft. Panel C Figure 3 shows that content provider investments are growing rapidly and account for overwhelming shares in markets such as Europe-North America in the 2018-2020 period.

3 A Model of Data Flows and Cable Investment

3.1 Data Flows and Cable Usage

In this section, we present an estimable model of how country-to-country demand for data maps into flows between regions that we observe in our data. In our model, countries are indexed by $c = 1, \ldots, C$ and each country pair ck has demand for data from each other. Time is indexed by t. Data are measured in Gigabits per second. Every pair of countries c and k has a potential demand for data from each other of $\overline{M}$. The share of data transmitted between countries is a function of their demand for data, which we capture with demographic and country characteristics, and the quality of the connection between the two countries, which is a function of cable characteristics. The total data transmitted between the two countries is:

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

where x_{ckt} captures time trends and country and country-pair characteristics such as the population and GDP of the two countries The parameters θ^d are to be estimated. The variable v_{ckt} captures the quality of the connection between the two countries, further described below. The share of potential demand that is not realized could be packets that are lost in internet transit. It could also be information that consumers choose never to search for in the first place because consumers are aware of internet quality or because the cost of exchanging data is not worthwhile given the consumer's income, distance, language, and other issues.

We group countries into subregions $r = 1, \ldots, R$ and subregions into regions $g = 1, \ldots, G$. The subregions form a partition of the countries, and the regions form a partition of the subregions. As described above, we choose regions so that communication between countries in different regions must traverse subsea cables. We use subregions to accommodate the seven island countries for which we utilize total international bandwidth data. We divide regions that contain island countries into subregions, with one subregion for each island country (these sub-

regions 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. Our model makes predictions at the subregion level that we add up to form region-to-region predictions of usage, which is the level we observe usage. We denote pairs of subregions as *markets* indexed by $m = 1, \ldots, M$. For example, North America-Europe is one market and North America-East Asia is another market. Markets are non-directional so North America-Europe is the same market as Europe-North America.

The network of cables dictates a set of paths that an internet packet might travel over. Between any two countries c and k, there exists P_{ckt} *paths* indexed by $p = 1, \ldots, P_{ckt}$. Each path is a set of markets for which there is an active cable. For instance, three possible paths from North America to Europe include {North America-Europe; North America-South America $\rightarrow$ South America-Europe; North America-Sub-Saharan Africa$\rightarrow$ Sub-Saharan Africa-Europe}. Internet cables are bidirectional so $P_{ckt} = P_{kct}$ and any sequence of markets that connect countries c to k also connects k to c. The set of paths can evolve as new cables are constructed in markets for the first time. Pairs of countries in the same respective sub-regions have the same set of paths available to them. For example, the US and Spain have the same paths between them as Canada and France.

The way in which data travels over the internet is a complex function of a number of variables. In general, routing behavior depends on traffic, so internet routers will automatically direct packets of information to the least crowded routes. Even packets from the same overall stream, for instance, a video or email message, may take very different routes to their destination. We use a simple reduced-form approximation for this process.

Each path is characterized by an *attractiveness* δ_{pckt}. Attractiveness δ_{pckt} depends on cable features such as path distance and bandwidth associated with each path as well as the level of competition between cable owners in the markets that make up the path.[9] The share of data going from c to k on path p at time t is:

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

We parameterize the attractiveness of the path δ_{pckt} to be a linear function of path characteristics as follows:

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

where Z_{pkt} includes a constant and path characteristics such as the length and the potential capacity of path p. We also use HHI for ownership share on that path. We construct poten-

[9]Local carriers and content providers negotiate with cable owners to obtain passage. Breakdowns in negotiations mean the cable will not accept that traffic, akin to an engineering constraint on how packets can be routed. In this sense, more competitive ownership structures can lead to more traffic on a path.

tial capacity by summing over potential capacity for cables that have landing points in both subregions containing countries c and k.

We let the quality of the connection between two countries equal the inclusive value of logit choice among cables:

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

Thus, $s_{pckt} = \exp\left(\delta_{pckt}\right) / \exp(v_{ckt})$. The amount of data traveling on a given path between two countries is:

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

In our model, more paths between two countries (higher P_{ckt}) and the increased quality of those paths (higher δ_{pckt}, such as because of higher capacity) lead those countries to communicate more data (more share drawn from the outside option). In this way, the model captures that the quality of connectivity affects the quantity of data transmission. We do not model network congestion explicitly because congestion, in the form of reduced consumer service, is typically realized only for short periods of time and we have annual data on usage, but the model captures congestion in the sense that better connectivity leads to more network use.

Let $\mathcal{M}_{pckt}$ be the set of markets associated with path p between country pair ck. The total demand between countries c and k traveling on market m in period t is:

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

Thus, total data in market m in period t is:

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

We denote total usage for country c in period t to be Q_{ct}. In our model, this equals:

$$Q_{ct} = \sum_{k=1}^{C} d_{ckt} \left(\sum_{p=1}^{P_{ck}} s_{pckt}\right).$$

3.2 Cable Investment

In our model, firms $j = 1, \ldots, J$ simultaneously choose whether to invest in a new cable in each market $m = 1, \ldots, M$. Firms profit from their ownership share of cables and the amount of data that traverses those cables. Each market is characterized by an amount of capacity available on that market K_{mt}. The amount of capacity owned by firm j is denoted by k_{mjt}.

In each period, a new line may be constructed in a market. Firms observe the capacity of the potential cable, $\tilde{k}_{mt}$. Then, each firm makes a binary choice $a_{jmt} \in \{0,1\}$ in each market whether to build the available line. Let $\boldsymbol{a}_{jt} = (a_{jmt})_m$ denote a vector of firm j's investment choices in all markets. If multiple firms choose to build in a given market, they become joint owners of the line, evenly splitting ownership of the line. Owners split the cost of the line $\omega_{mt}(\tilde{k}_{mt})$ evenly. Each owner also receives a firm-specific investment cost shock (ν_{jmt}), facing the total cost of $\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}}\omega_{mt}(\tilde{k}_{mt}) + \nu_{jmt}$ if it decides to invest in market mt. In our framework based on Pakes et al. (2015), the value of ν_{jmt} can be common knowledge across firms or privately observed.

The profit for firm j is given by

$$\bar{\pi}_{jt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) = \sum_{m \in \mathcal{M}} \left\{ r_m(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \right\}.$$

where $r_m(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt})$ is the revenue in market m from choosing $\boldsymbol{a}_{jt}$ (i.e. the total profit before paying the investment cost) when the competitors choose $\boldsymbol{a}_{-jt}$ and X_{jt} is the set of all exogenous variables that affect the profit. Firms have measurement error over elements of $r(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt})$, which is represented by the additive term ε_{ajmt}. Thus, letting $\mathcal{J}_{jt}$ be the set of information firm j uses in forming expectations about the revenue to choosing $\boldsymbol{a}_{jt}$, firm j perceives the profit from investment in mt as:

$$\begin{aligned}
\pi_{jt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) = \; & E\left[\sum_{m \in \mathcal{M}} \left\{ r_m\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \right\} \Big| \mathcal{J}_{jt} \right] \\
= \; & \sum_{m \in \mathcal{M}} \left\{ r_m\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) + \varepsilon_{ajmt} - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \right\}.
\end{aligned}$$

Firms play a Nash equilibrium, simultaneously solving:

$$\max_{\boldsymbol{a}_{jt}} \pi_{jt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right).$$

Let $\tilde{\boldsymbol{a}}_{jt}^{m,1}$ be the vector $\boldsymbol{a}_{jt}$ where the mth element is set to 1. We similarly define $\tilde{\boldsymbol{a}}_{jt}^{m,0}$. That is, $\tilde{\boldsymbol{a}}_{jt}^{m,1} = (a_{j1t}, ..., a_{jm-1t}, 1, a_{jm+1t}, ..., a_{jMt})$ and $\tilde{\boldsymbol{a}}_{jt}^{m,0} = (a_{j1t}, ..., a_{jm-1t}, 0, a_{jm+1t}, ..., a_{jMt})$. The incremental change in the revenue in market z to j from investing in market m is:

$$\Delta r_{mz}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) = r_z(\tilde{\boldsymbol{a}}_{jt}^{m,1}, \boldsymbol{a}_{-jt}, X_{jt}) - r_z(\tilde{\boldsymbol{a}}_{jt}^{m,0}, \boldsymbol{a}_{-jt}, X_{jt}).$$

Also, we define $\Delta\varepsilon_{jmt} = \varepsilon_{1jmt} - \varepsilon_{0jmt}$. Revealed preference tells us that if firm j invests in

market m, t, it must be that:

$$\sum_{z \in \mathcal{M}} \Delta r_{mz}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) + \Delta\varepsilon_{jmt} - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \geq 0 \tag{4}$$

We take the expectation of equation (4) conditional on the firm's information set $\mathcal{J}_{jt}$ to obtain

$$E\left[\sum_{z \in \mathcal{M}} \Delta r_{mz}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \Big| a_{jmt} = 1, \mathcal{J}_{jt} \right] \geq 0$$

where $E[\Delta\varepsilon_{jmt}|\mathcal{J}_{jt}, a_{jmt}] = 0$ follows from the rational expectations assumption. Suppose that $Z_{jmt} \subsetneq \mathcal{J}_{jt}$ is the set of variables in the firm's information set that are exogenous in the sense that they do not respond to changes in a_{jt} and a_{-jt}. Then, we can apply the law of iterated expectations to derive:

$$E\left[\sum_{z \in \mathcal{M}} \Delta r_{mz}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \Big| a_{jmt} = 1, Z_{jmt} \right] \geq 0. \tag{5}$$

Similarly, we can define a separate moment for the case of no investment:

$$E\left[\sum_{z \in \mathcal{M}} \Delta r_{mz}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) + a_{jmt}\nu_{jmt} \right) \Big| a_{jmt} = 0, Z_{jmt} \right] \leq 0. \tag{6}$$

It remains to specify the revenue function $r_m()$ and the investment cost function $\omega_{mt}()$. We specify the revenue function as:

$$r_m(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) = \gamma D_{mt} \frac{k_{jmt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right)}{K_{mt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right)}$$

where γ is the parameter of our interest. The total data transmitted in market m, denoted $D_{mt}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt})$, depends not only on the total capacity in market m, but also on the capacities in all other markets, as well as interregional demand for data flows across all markets. The variables K_{mt} and k_{jmt} are determined based on investment choices as follows:

$$K_{mt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) = K_{mt-1} + \tilde{k}_{mt}\mathbf{1}\left\{ \sum_{j=1}^{J} a_{jmt} \geq 1 \right\}$$

$$k_{jmt}\left(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}\right) = k_{jmt-1} + \frac{a_{jmt}}{a_{jmt} + \sum_{-j} a_{-jmt}} \tilde{k}_{mt}$$

We specify the investment cost function as:

$$\omega_{mt}(\tilde{k}_{mt}) = \psi_0 + \psi_1 \ln\left(\tilde{k}_{mt}\right) + \psi_2\left(\tilde{d}_{mt}\right) + \rho_{mt} \tag{7}$$

where $\tilde{d}_{mt}$ is the distance of the line.

Equations (5) and (6) still cannot be applied to the data as they depend on unobserved structural errors (ν_{jmt}), which causes the well-known selection issue. That is, conditional on choices, these errors are not mean zero. To move forward, we assume $\nu_{jmt} = 0$ for all j, m, t, obtaining the following inequality from equation (5) :

$$E\left[\sum_{z\in\mathcal{M}} \Delta r_{mz}(\boldsymbol{a}_{jt}, _{-jt}, X_{jt}) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}}\omega_{mt}(\tilde{k}_{mt})\right)\Big| a_{jmt} = 1, Z_{jmt}\right] \geq 0. \tag{8}$$

We can define the analogous inequality for the case of no investment.

4 Estimation

4.1 Data Flows and Cable Usage

In estimation, we would like to match D_{mt} to the observed used bandwidth in the data. A challenge is that markets m are defined to be pairs of subregions but we observe used bandwidth between pairs of regions. We need to sum over markets to obtain a region-to-region variable for estimation. Let j index pairs of regions, and let $\mathcal{S}_j$ be the set of markets contained in the region pair. For instance, the region *East Asia* has three subregions, *Japan*, *Korea* and *Rest of East Asia*. The region *North America* has only one subregion. For j=*East Asia-North America*, $\mathcal{S}_j$ contain three markets, *Japan-North America*, *Korea-North America* and *Rest of East Asia-North America*.

We define $\tilde{D}_{jt} = \sum_{m\in\mathcal{S}_j} D_{mt}$. We denote the predicted data flows for region-pair j and time t given parameter θ as $\tilde{D}_{jt}(\theta)$ where $\theta = [\theta^d, \theta^\delta]$ includes parameters that govern demand between countries (θ^d) and parameters that govern how demand gets allocated across the cable network (θ^δ). Let $\overline{D}_{j.}(\theta)$ is the average over time of $\tilde{D}_{jt}(\theta)$, and $\overline{D}_{.t}(\theta)$ is the average over j. Similarly, we denote the predicted international data flow for country c and time t given parameter θ as $\tilde{Q}_{ct}(\theta)$. We let $\overline{Q}_{.t}(\theta)$ denote the average of $\tilde{Q}_{ct}(\theta)$ over island countries and $\overline{Q}_{.t}(\theta)$ the average over time.

Given the parameter values, we compute the data flows for each market and time period based on equations (3). We construct the set of moments by stacking the average predicted data flows at the market level and year level:

$$\Gamma(\theta) = \left[\overline{Q}_{c.}(\theta), \overline{Q}_{.t}(\theta), \overline{D}_{j.}(\theta), \overline{D}_{.t}(\theta)\right].$$

Let Z be the set of instruments such that $E\left[Z'\left(\Gamma^d - \Gamma(\theta)\right)\right] = 0$ where Γ^d is the set of moments based on the observed bandwidth usage in the data. We search for the parameter vector that minimizes the weighted distance between the predicted and observed moments given as

$$f(\theta) = \left(\Gamma^d - \Gamma(\theta)\right)' Z' W^{-1} Z \left(\Gamma^d - \Gamma(\theta)\right)$$

where W is the consistent estimate of $E[Z'\left(\Gamma^d - \Gamma(\theta)\right)\left(\Gamma^d - \Gamma(\theta)\right)' Z]$. We use constants and time trends as instruments and set $\overline{M} = 1000$.

Table 10 presents the demand-side estimates. 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 4 percent increase in trade in data, and a one-percent increase in the sum of the population would result in a 2.4 percent decrease in trade in data. Thus, a one-percent increase in both GDP and population (an increase in GDP per capita) leads to an increase in trade in data. We estimate high fixed effects for markets such as East Asia-North America, East Asia-South America, and low fixed effects for markets such as Europe-South Asia and North America-South Asia. Further, consistent with our expectations, the path length has a negative effect, while the path capacity and number of cables 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.

4.2 Cable Investment

We derive the following inequalities by interacting inequalities in (8) with positive valued functions of Z_{jmt}, denoted as $h(Z_{jmt})$.

$$\frac{1}{N^{a_{jmt}}=1} \sum_{\{j,m,t|a_{jmt}=1\}} \left(\sum_{z \in \mathcal{M}} \Delta r_{mz}(\boldsymbol{a}_{jt}, \boldsymbol{a}_{-jt}, X_{jt}) - \left(\frac{a_{jmt}}{\sum_{k=1}^{J} a_{kmt}} \omega_{mt}(\tilde{k}_{mt}) \right) \right) h(Z_{jmt}) \geq 0$$

where $N^{a_{jmt}} = 1$ is the number of observations (firm-market-year combinations) with investment. We can define an analogous inequality for the case of no investment.

We focus on recovering the revenue function while calibrating the cost of investment using the consturction data for each cable.[10] We estimate equation (7) by regressing the construction cost on a costant, length, and capacity of the cable using only cables that connect one market (see Appendix Table A1 for detailed results). We use predicted values from this regression as the investment cost.

[10]The usage of data on observed transaction costs is not new in papers that study investment in capital-intensive industries. For example, see .

Another step that is necessary to apply the inequalities is defining counterfactual outcomes. That is, given that we observe investment by a firm in a certain market, what would happen to the market if the firm decides to not invest, holding the strategies of its rivals constant? In the case of investment, we assume that the cable of the same capacity observed in the data if there is at least one other firm that invested in the market and year. Otherwise, we assume that there is no new capacity added. In the case of no investment, if we observe a cable built in that year and market, we assume that there will be a cable of the same capacity in the counterfactual scenario. If there is no cable built, we use the predicted capacity from a regression of capacity on time and market fixed effects as the counterfactual capacity. The results from this capacity regression is presented in Appendix Table A2.

We use the method by Cox and Shi (2023) as our inference procedure (see in Appendix B for detailed steps). Table 11 shows the estimates of the revenue parameter, γ. When using the full sample of observations, we recover a 95% confidence interval of (0.0028, 0.0044). Given that the construction costs are measured in million dollars and the capacities in Gbps, the estimates imply that an increase in a firm's demand ($\sum_{m\in\mathcal{M}} D_{mt}\frac{k_{jmt}}{K_{mt}}$) by 1 Tbps would lead to a profit increase ranging from 2.8 to 4.4 million dollars. We estimate the revenue parameter for non-content providers and content providers separately. We obtain a confidence interval of (0.0034,0.0044) for non-content providers and (0.0003, 0.0055) for content providers, suggesting that there is no statistically significant difference in the revenue function between the two groups of firms.

We highlight the network nature of the market by examining how much of the data transmissions on cables in a given market can be attributed to actual endpoint-to-endpoint usage. For example, Table 8 shows that 55% of the usage in cables in the Europe-North America market in 2020 can be attributed to data exchanges between Europe and North America, while the rest is due to usage in other markets such as Europe-East Asia and Australia-Europe.

5 The Gravity of Data Flows

A limitation in comparing observed bandwidth usage to trade flows in goods as in Section 2.3 is that bandwidth usage does not capture endpoint-to-endpoint usage, because data traveling between two countries can take multiple routes determined by various factors. Our structural model allows us to map observed data flows to country-to-country data exchanges. That is, given our estimates of the demand parameters, we infer country-to-country demand for data based on equation (1), which we compare directly with international trade flows in goods.[11]

[11] Our model does not capture intraregional demand for data. Our calculation of country-level trade in data, therefore, excludes this portion of trade. As a result, we expect the total international data flows for countries with a large amount of intraregional trade (for example, European countries) would be higher than our predictions. We also exclude intraregional trade when computing trade in goods in Panel A of Tables 5 to 7 to make a clear comparison.

As shown in Panel B of Table 6, trade in data is heavily concentrated in the top two markets (East Asia-North America and Europe-North America) in 2002. The concentration in data trade falls rapidly throughout our sample period, even faster than the concentration of bandwidth usage does. Similarly, whereas the list of top countries and their shares remain relatively stable for trade in goods in this period, the concentration falls much faster for trade in data and this decline is also more drastic than for bandwidth usage. Several Asian countries also entered the list by 2020 including China and Singapore.

We compute the centrality in the global networks of trade for goods and data, respectively, based on eigenvector centrality. Table 9 reports which countries are most central in these networks. We compute eigenvector centrality separately for the goods and data networks, and separately for each year. In 2020, the list of top countries is similar for trade in goods and trade in data, while in 2002 there are more European countries and less Asian countries for trade in data. We measure the concentration of centrality by computing HHI (sum of squared eigenvector values) in each year and find that the HHI for trade in data is much higher than the HHI for trade in goods in 2002 but falls below it by 2020 as shown in Figure 4. That is, centrality is highly concentrated for data relative to goods at the start of our data in 2002 but is more dispersed for data than trade by the end of the same, reflecting the spread of trade in data across a wide set of countries.

We study which factors have contributed to the growth in trade in data, decomposing the effects into that arising from (a) changes in x_{ckt} including demographic characteristics and time trends and (b) changes in v_{ckt}, or the quality of cable connections. To do so, we compute the total globel data flows based on our model from 2002 to 2020, which is shown as the blue line in Figure 5. To isolate the effect of the changes in the quality of the cable connection, We compute the total data flows with x_{ckt} fixed at the 2020 level and setting v_{ckt} to its observed level for each year. Similarly, to isolate the effect of changes in x_{ckt}, we compute the total data flows with v_{ckt} fixed at the 2020 level while setting x_{ckt} to its observed level each year. Results appear in Figure 5. Because the two effects interact, the sum of each isolated effect is often less than the total effect. We find that in 2002, for example, enhancing the cable network to the 2020 level while fixing everything else at the 2002 level would not lead to a substantial increase in data transmissions. Endowing the world with the 2020 level of x_{ckt} while fixing the cable network at the 2002 level, however, would bring the data flows up to approximately 45% of the 2020 level. Starting from 2014, however, the demographic characteristics (including the time trend) and the quality of cable connections have comparable contributions to the growth in data flows.

6 Inefficiencies in the Subsea Cable Network

The high level of geographical concentration in the cable network suggests a potential misallocation of cables to markets that receive relatively low social values from these cables. We characterize these inefficiencies in the cable network and quantify the business stealing effects based on our estimates. For each market, we consider constructing a new cable by a de novo entrant with no cable. We assume that the capacity of the new cable is given in the same way we construct counterfactual outcomes in the estimation based on the capacity regression described in Section 4.2. We also assume that all other firms, or incumbents, take actions observed in the data.

For each counterfactual scenario, we compute the effect of the new cable on incumbent revenues (the sum of revenues over all firms in our data). Unlike in other settings with independent markets, an investment in this setting affects incumbents' revenues not only in the market where the investment happens but also in all other markets due to the network nature of the market. Our model allows us to decompose the effect of a new cable into these two effects.

The direction of these effects is theoretically ambiguous. A new cable would transfer market shares from incumbents to the entrant in the investment market. Furthermore, an increase in the capacity in that market would divert data flows away from the other markets, lowering profits in these markets as well. At the same time, the increased capacity would increase demand for data flows in the investment market, resulting in a market expansion effect. Some of the increased demand would be served by cables in other markets, potentially increasing demand everywhere.

We compute changes in the incumbent revenue for the market in which the investment happens and all other markets separately (Note that the change in the revenue is equal to the change in the profit for these counterfactuals as incumbents experience no change in the costs). Panel A of Table 12 shows the results computed based on the lower bound of γ. We focus on the year 2010 and three markets that involve North America and three markets that involve Sub-Saharan Africa. In contrast to the North American markets that have received repeated, large investments throughout the sample period, Sub-Saharan Africa has been generally considered to be underserved. For example, Sub-Saharan Africa-East Asia has no cable connection as of 2023, and Sub-Saharan Africa-South America received the first substantial investment in 2018.

Our estimates suggest that the new investment would lower the incumbent revenue in two of the markets involving North America. Much of the negative effects are coming from the market where the investment happens, suggesting a substantial transfer from the incumbents to the entrant in this market. For the North America-Europe and North America-South America markets, incumbents even enjoy some positive revenue gains in the rest of the world due to the market expansion and spillovers. By contrast, in all three of the Sub-Saharan markets,

incumbents receive positive and large gains from the investment. Except for the Sub-Saharan Africa-Europe market, the investment has a positive effect on the incumbent revenue both in the investment market and the rest of the world.

We measure business stealing as part of the entrant revenue that is a transfer from the incumbent revenue. We quantify this as the ratio of the negative value of the total change in the incumbent revenue (in the third column) to the entrant revenue (in the fourth column). As discussed above, there are strong business stealing effects in the North American markets. The business stealing effect is negative in the Sub-Saharan African markets as the incumbents capture increased revenues from increased demand.

We examine how social gains from a new cable entry correlate with private gains across markets in Figure 6 for the six selected markets in the year 2010. The social gain is computed as the sum of the change in data flows (D_{mt}) across all markets, and the private gain is computed as part of the data flows captured by the entrant. We find that markets like Sub-Saharan Africa-East Asia and Sub-Saharan and South America have high social gains relative to private gains compared to other markets. The three North American markets have relatively low social values from the new cable. Our findings, therefore, imply that there is a substantial misallocation of investment and that targeted subsidies could lead to a more efficient network.

7 Conclusion

Undersea internet cables are a critical piece of communication infrastructure, carrying more than 99% of international data traffic and underpinning modern global commerce. There is little economic research on this market despite massive attention to the digitization and e-commerce that it enables. Our paper provides two main contributions.

First, we treat global traffic on the cables as *international trade in data*, and we compare the evolution of trade in data with trade in goods. We utilize a new data set on cable construction and use to estimate a model of country-to-country trade in data, akin to a gravity equation model well-known in the literature on international trade. We find that the concentration of trade in data among wealthy countries started at a much higher level than trade in goods but declined at a much faster rate. We also decompose the growth of trade in data into an amount due to increases in demand and an amount due to investment in the network. We find that cable construction was a substantial contributor, on par with increases in demand.

Second, we develop and estimate a model of cable construction. Our model accounts for how constructing a cable between any two locations can affect global traffic flows according to the logic of internet transmission. In addition, our model endogenizes the formation of consortiums of investors, which are frequently observed in practice. We also allow for content providers, such as Google and Microsoft, to have different monetization than telecommunication carriers, such as AT&T and British Telecom, as the recent growth of content-provider

investment has been an important development in the industry.

Our model of investment allows us to contrast private investment incentives with socially optimal investment incentives and we find that they deviate substantially. The importance of business stealing is consistent with our descriptive evidence, where we document repeated investment in busy routes and practically non-existent investment in routes with lower demand. Importantly, the global internet backbone industry is characterized by a very low level of economic regulation. Our results highlight how appropriate subsidies could lead the industry towards a more efficient global network.

Overall, we provide a new model of undersea internet cable construction and use and estimate the model on new data on this important market. Our results emphasize the importance of cable construction in global data flows and the likelihood that the market delivers inefficient outcomes. Future research on adjacent markets, such as the growth of data centers, appears valuable.

References

Aguirregabiria, V. and Ho, C.-Y. (2010). A dynamic game of airline network competition: Hub-and-spoke networks and entry deterrence. *International Journal of Industrial Organization*, 28:377–382.

Andrews, D. W. K. and Barwick, P. J. (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica : journal of the Econometric Society*, 80(6):2805–2826.

Blum, B. S. and Goldfarb, A. (2006). Does the internet defy the law of gravity? *Journal of International Economics*, 70(2):384–405.

Brancaccio, G., Kalouptsidi, M., and Papgeorgiou, T. (2020). Geography, transportation, and endogenous trade costs. *Econometrica : journal of the Econometric Society*, 88:657–691.

Canay, I. A., Illanes, G., and Velez, A. (2023). A User's Guide for Inference in Models Defined by Moment Inequalities.

Caoui, E. H. and Steck, A. (2023). Network diversity, market entry, and the global internet backbone. *SSRN Electronic Journal*.

Clark, D. D. (2018). *Designing an Internet*. MIT Press, Cambridge, MA.

Cox, G. and Shi, X. (2023). Simple adaptive size-exact testing for full-vector and subvector inference in moment inequality models. *Review of Economic Studies*, 90:201–228.

CRS (2022). Undersea Telecommunication Cables: Technology Overview and Issues for Congress. Technical report.

Dickstein, M. J. and Morales, E. (2018). What do Exporters Know?*. *The Quarterly Journal of Economics*, 133(4):1753–1801.

Eizenberg, A. (2014). Upstream innovation and product variety in the United Sstates home PC market. *The Review of Economic Studies*, 81:1003–1045.

Fan, Y. and Yang, C. (2022). Estimating Discrete Games with Many Firms and Many Decisions: An Application to Merger and Product Variety.

Fréchette, G. R., Lizzeri, A., and Salz, T. (2019). Frictions in a competitive, regulated market: Evidence from taxis. *American Economic Review*, 109:2954–2992.

Greenstein, S. (2015). *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton University Press.

Hjort, J. and Poulsen, J. (2019). The arrival of fast internet and employment in africa. *American Economic Review*, 109:1032–1079.

Holmes, T. J. (2011). The diffusion of Wal-Mart and economies of density. *Econometrica : journal of the Econometric Society*, 79:253–302.

Houde, J.-F., Newberry, P., and Seim, K. (2022). Nexus tax laws and economies of density in E-commerce: A study of Amazon's fulfillment center network. *Econometrica : journal of the Econometric Society*, In Press.

Jeon, J. (2022). Learning and investment under demand uncertainty in container shipping. *RAND Journal of Economics*, 53:226–259.

Jia, P. (2008). What happens when Wal-Mart comes to town: An empirical analysis of the discount retailing industry. *Econometrica : journal of the Econometric Society*, 76:1263–1316.

Kline, B., Pakes, A., and Tamer, E. (2021). Chapter 5 - Moment inequalities and partial identification in industrial organization7373We are grateful for feedback from a set of reviewers that improved the chapter. We also thank Aureo de Paula and Christopher D. Walker for insightful comments on the chapter. In Ho, K., Hortaçsu, A., and Lizzeri, A., editors, *Handbook of Industrial Organization*, volume 4 of *Handbook of Industrial Organization, Volume 4*, pages 345–431. Elsevier.

Li, S., Mazur, J., Park, Y., Roberts, J., Sweeting, A., and Zhang, J. (2021). Repositioning and market power after airline mergers. Unpublished manuscript, University of Maryland.

Pakes, A., Porter, J., Ho, K., and Ishii, J. (2015). Moment inequalities and their application. *Econometrica : journal of the Econometric Society*, 83:315–334.

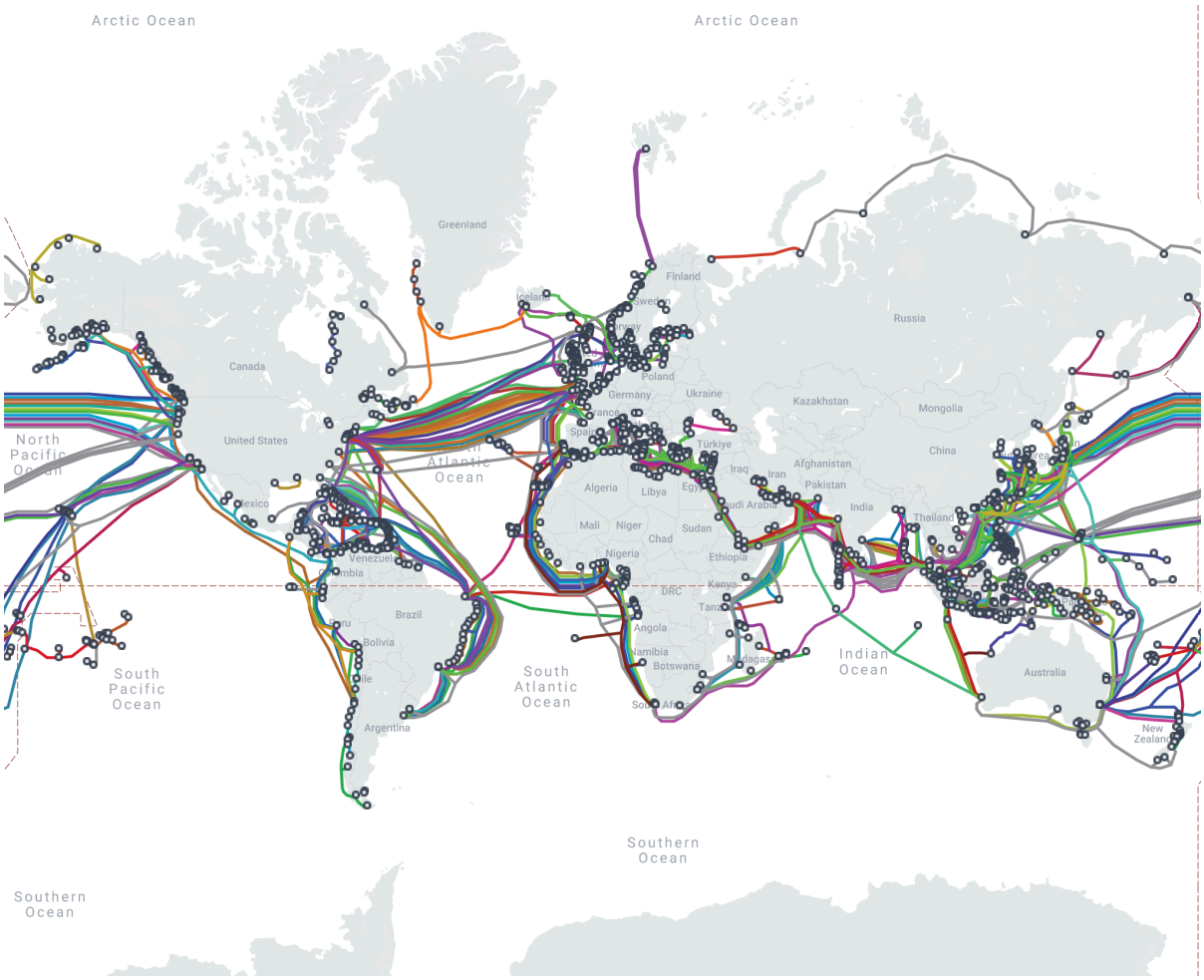
Rysman, M. (2016). Empirics of business data services. Appendix B of *Business Data Services Federal Notice of Proposed Rulemaking*, FCC 16-54.

Steinwender, C. (2018). Real Effects of Information Frictions: When the States and the Kingdom Became United. *American Economic Review*, 108(3):657–696.

Wollmann, T. G. (2018). Trucks without Bailouts: Equilibrium Product Characteristics for Commercial Vehicles. *American Economic Review*, 108(6):1364–1406.

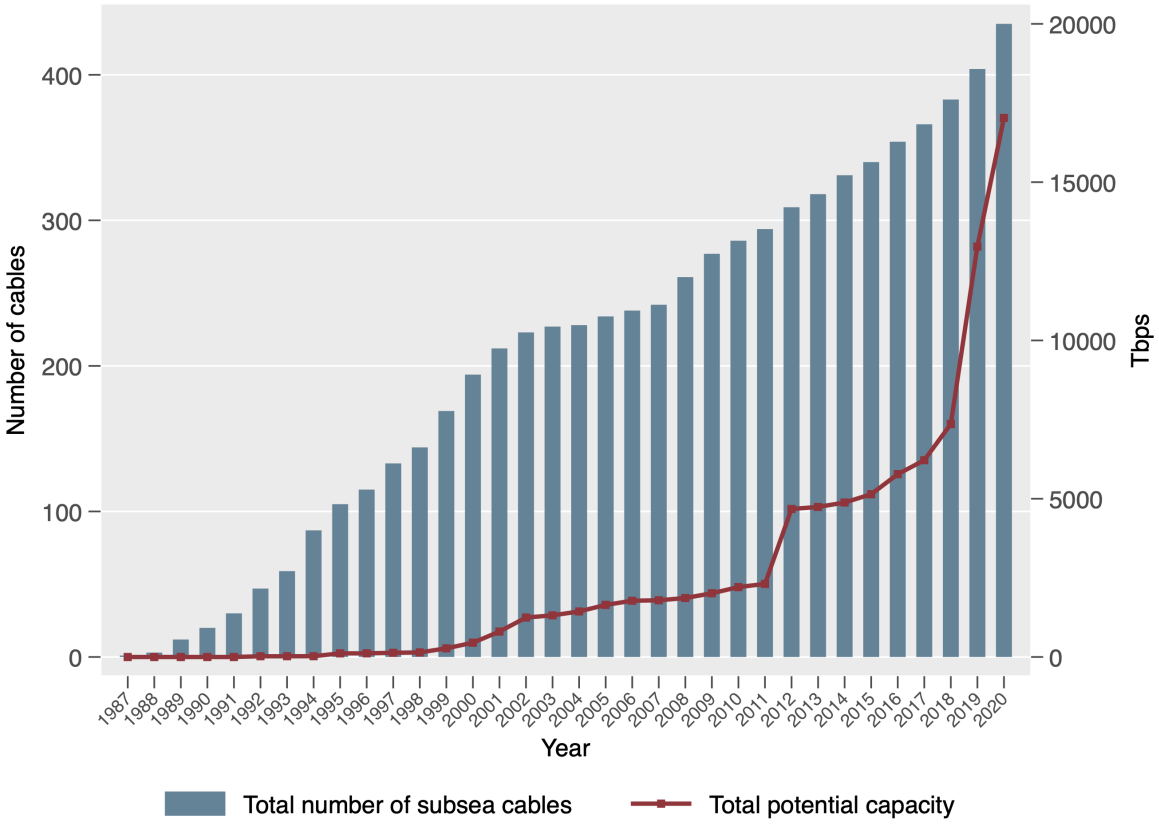
Tables and Figures

Figure 1: Map of subsea cables in 2023



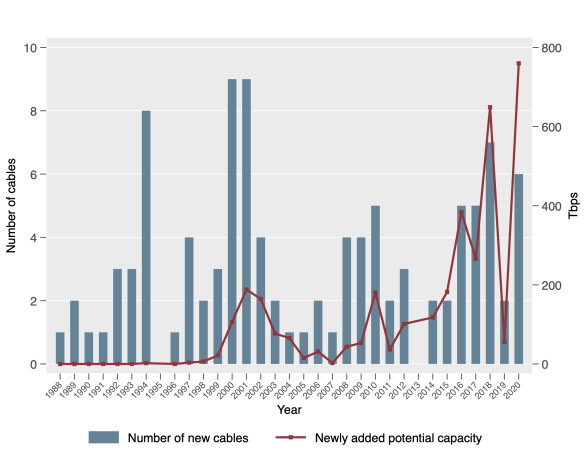
Notes: This map shows undersea telecommunications cables as of October 5, 2023. Source: TeleGeography, Submarine Cable Map, www.submarinecablemap.com

Figure 2: Total number and capacity of subsea cables

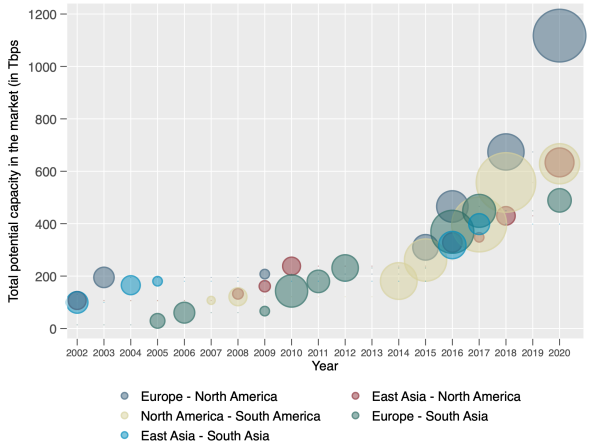


Notes: This figure displays the total number of subsea cables and the total potential capacity from 1987 to 2020 for all cables included in the data provided by Telegeography.

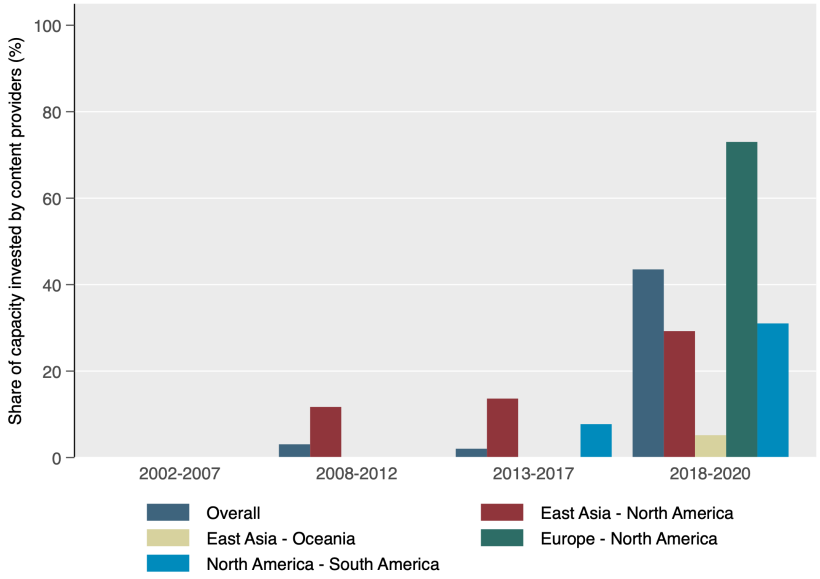
Figure 3: Investment in subsea cables



(a) New cables



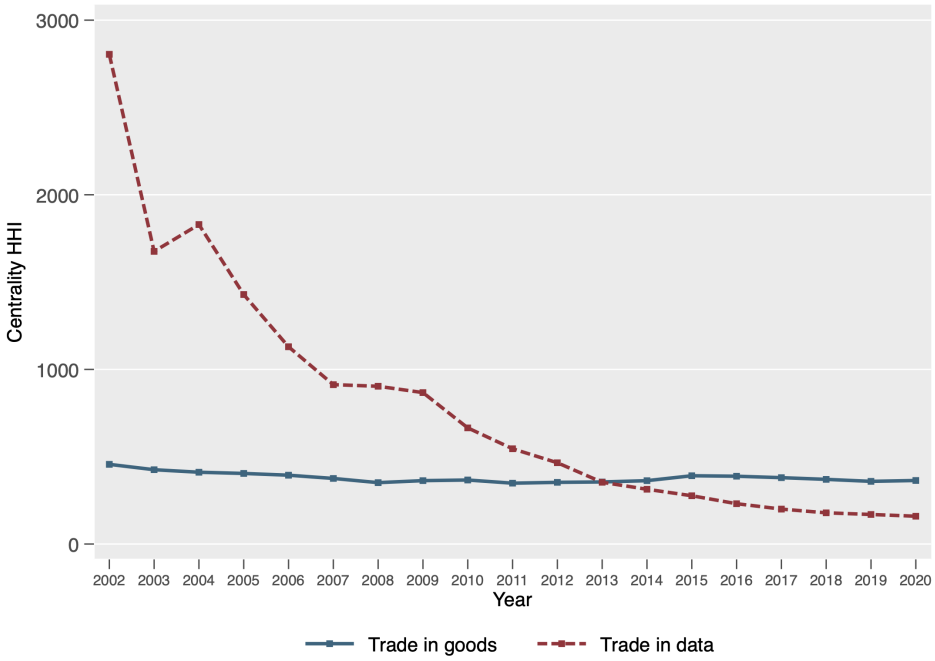
(b) Geographic concentration



(c) The share of newly added capacity by content providers by market

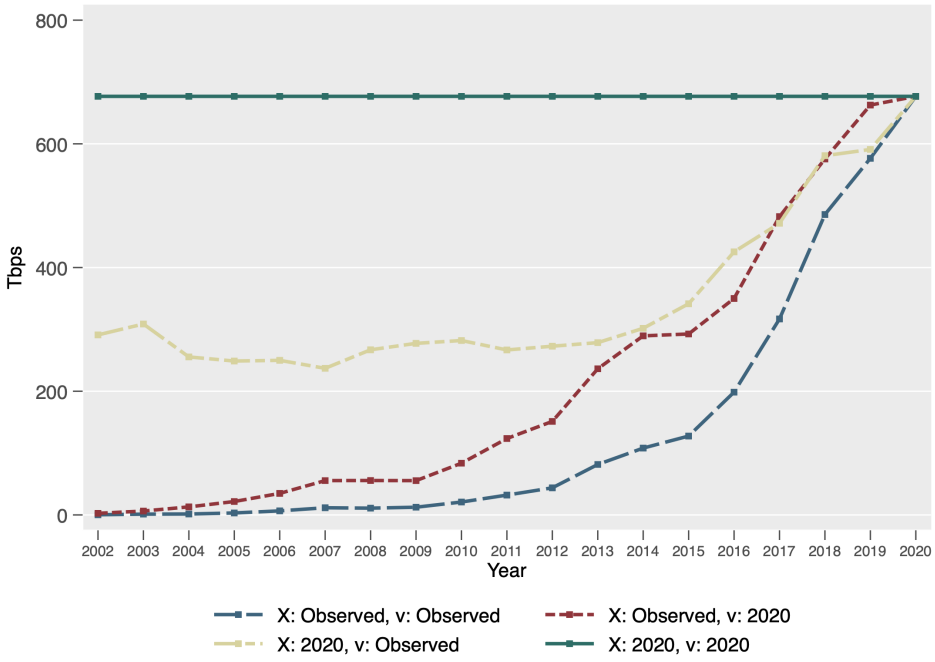
Notes: Panel (a) shows the number of new cables and newly added capacity. Panel (b) shows the size of new investments represented by the size of the bubbles in relation to the total existing capacity for major markets. Panel (c) shows the share of capacity invested by content providers for the overall industry and for selected markets.

Figure 4: The centrality of trade in goods and data



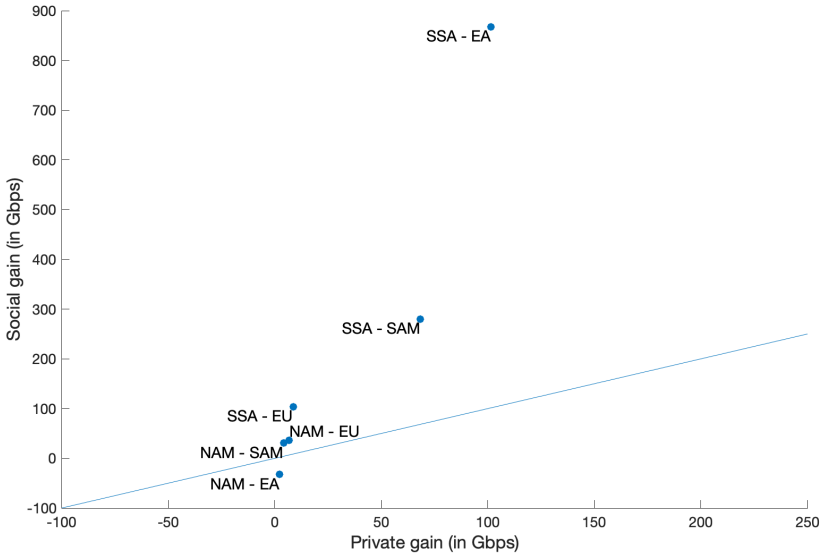
Notes: This figure shows the HHI of centrality in trade in goods and trade in data based on the eigenvector centrality measure computed at the country level.

Figure 5: The decomposition of trade in data



Notes: This figure plots the total data flows computed based on our model estimates, holding country-pair characteristics (x_{ckt}) and quality of connection (v_{ckt}) at the observed levels. It also plots data flows under the following three counterfactuals: (i) holding x_{ckt} and v_{ckt} at the 2020 levels; (ii) holding x_{ckt} at the 2020 level and v_{ckt} at the observed level; and (iii) holding v_{ckt} at the 2020 level and x_{ckt} at the observed level.

Figure 6: Social and private gains from new cable entry



Notes: The social gain is computed as the sum of total data flows (D_{mt}) across all markets. The private gain is the data flows served by the new entrant. The abbreviations stand for the following regions: EA for East Asia, EU for Europe, NAM for North America, SAM for South America, and SSA for Sub-Saharan Africa.

Table 1: Summary statistics

Panel A: Cable characteristics at the cable level (105 unique cables)

	Mean	SD	Min	25th	Median	75th	Max	N
Number of owners	6.27	8.16	1	1	3	7	49	105
Number of markets served	1.44	1.32	1	1	1	1	10	105
Ready-for-service year	2,005.38	9.37	1,988	1,999	2,003	2,015	2,020	105
Cable length (in km)	10,498.13	7,549.58	269	4,852	9,330	14,326	39,000	104
Potential capacity (in Gbps)	34,439.32	49,045.97	0	1,000	15,600	57,600	300,000	102
Construction costs (in million USD)	456.62	397.05	8	200	310	600	1,700	92
Direct cable	0.86	0.35	0	1	1	1	1	105

Panel B: Owner characteristics at the owner-year level (207 unique owners)

	Mean	SD	Min	25th	Median	75th	Max	N
Number of markets served	5.06	4.79	1	1	3	10	16	2889
Number of cables owned	2.88	3.43	1	1	1	3	29	2889
Total potential capacity	8,574.08	26,558.80	0	185	913	4,760	489,456	2889
Content provider	0.01	0.08	0	0	0	0	1	2889

Panel C: Market characteristics at the market-year level (21 unique markets)

	Mean	SD	Min	25th	Median	75th	Max	N
Total potential capacity (in Tbps)	82.34	136.81	0	0	15	118	1,118	399
Number of cables	4.06	4.80	0	1	2	6	21	399
Number of direct cables	3.13	4.73	0	0	1	4	21	304
Distance between regions (in thousand km)	10.27	3.69	4	8	10	14	16	399
Average length of cables (in thousand km)	20.63	10.54	7	11	19	26	39	304
Average length of direct cables (in thousand km)	9.76	6.24	1	6	8	13	30	161
Owner-level HHI of potential capacity	1,539.83	1,995.61	192	223	983	1,724	10,000	304
Number of owners	36.66	27.99	0	3	45	55	113	399

Panel D: Country characteristics at the country-year level (161 unique countries)

	Mean	SD	Min	25th	Median	75th	Max	N
GDP (in billion current USD)	405.47	1,561.70	1	11	38	221	21,433	3059
Population (in millions)	42.61	147.29	0	4	10	30	1,411	3059
Internet user %	40.05	31.14	0	10	35	69	100	3059
Number of mobile subscriptions (in millions)	33.23	118.06	0	2	7	21	1,746	3059

Table 2: 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	13.01	8	36	506.36
2003	15.25	8	37	479.73
2004	18.51	8	38	462.95
2005	20.90	9	39	450.19
2006	23.99	9	40	436.86
2007	27.79	9	42	422.30
2008	32.15	10	44	399.92
2009	24.72	9	43	413.82
2010	30.02	10	43	417.51
2011	35.65	10	44	404.08
2012	36.23	10	45	404.27
2013	36.66	10	45	408.41
2014	36.92	10	45	420.48
2015	32.38	9	42	447.33
2016	31.61	9	41	450.55
2017	35.11	10	42	442.43
2018	38.77	10	42	440.28
2019	37.81	10	42	440.82
2020	34.87	10	41	461.20

Panel B: International bandwidth usage from 2002 to 2020

Year	Total international bandwidth usage (Tbps)	Number of countries making up 50% of usage	Number of countries making up 90% of usage	HHI
2002	2.97	4	14	1076.32
2003	4.83	3	14	1141.85
2004	6.83	4	15	1066.20
2005	10.00	4	17	1041.90
2006	14.00	4	20	978.47
2007	22.97	4	20	970.06
2008	38.45	4	25	860.42
2009	61.08	4	26	798.88
2010	93.37	5	28	777.18
2011	139.98	5	28	753.24
2012	202.10	5	29	739.92
2013	290.89	5	29	719.61
2014	424.16	5	28	720.43
2015	599.21	5	28	721.82
2016	888.17	5	28	718.06
2017	1,335.16	5	27	730.33
2018	1,995.80	5	26	736.50
2019	2,918.79	5	24	767.32
2020	4,149.75	5	24	719.16

Table 3: Top 10 markets for trade and international bandwidth usage in 2002 and 2020

Panel A: Top 10 markets for trade in 2002 and 2020

Market	Trade Volume (billion USD)	Share (%)	Market	Trade Volume (billion USD)	Share (%)
East Asia - Europe	664.34	22.93	East Asia - Europe	2395.51	28.43
East Asia - North America	656.94	22.67	East Asia - North America	1484.25	17.62
Europe - North America	571.93	19.74	Europe - North America	1140.34	13.53
North America - South America	371.10	12.81	North America - South America	747.97	8.88
Europe - South America	116.47	4.02	East Asia - South America	542.03	6.43
East Asia - Oceania	82.52	2.85	East Asia - Oceania	377.72	4.48
Europe - Sub-Saharan Africa	80.95	2.79	Europe - South Asia	310.83	3.69
East Asia - South America	80.05	2.76	Europe - South America	307.92	3.65
Europe - South Asia	59.88	2.07	East Asia - South Asia	271.78	3.23
Europe - Oceania	40.86	1.41	Europe - Sub-Saharan Africa	242.65	2.88

Panel B: Top 10 markets for international bandwidth usage in 2002 and 2020

Market	Usage (Gbps)	Share (%)	Market	Usage (Gbps)	Share (%)
Europe - North America	368.34	70.70	Europe - North America	355,365.52	52.30
East Asia - North America	99.12	19.02	East Asia - North America	151,537.48	22.30
North America - South America	34.11	6.55	North America - South America	86,637.99	12.75
North America - Oceania	12.60	2.42	Europe - South Asia	24,333.07	3.58
East Asia - Oceania	2.20	0.42	East Asia - South Asia	22,279.32	3.28
East Asia - South Asia	1.63	0.31	Europe - Sub-Saharan Africa	12,175.97	1.79
East Asia - Europe	1.35	0.26	East Asia - Europe	10,811.18	1.59
North America - South Asia	0.57	0.11	North America - Oceania	9,769.82	1.44
North America - Sub-Saharan Africa	0.55	0.11	East Asia - Oceania	4,708.65	0.69
Europe - Sub-Saharan Africa	0.33	0.06	North America - South Asia	1,264.55	0.19

Table 4: Top 10 countries for trade and international bandwidth usage in 2002 and 2020

Panel A: Top 10 countries for trade in 2002 and 2020

Country	Trade Volume (billion USD)	Share (%)	Country	Trade Volume (billion USD)	Share (%)
United States	1913.21	14.70	China	4528.76	12.99
Germany	1127.30	8.66	United States	3716.98	10.66
Japan	788.98	6.06	Germany	2572.27	7.38
China	765.01	5.88	Japan	1336.62	3.83
France	633.83	4.87	Netherlands	1170.58	3.36
United Kingdom	611.13	4.70	France	1097.92	3.15
Canada	490.18	3.77	South Korea	1036.96	2.97
Italy	484.96	3.73	United Kingdom	991.98	2.84
Netherlands	444.88	3.42	Italy	922.70	2.65
Belgium	380.99	2.93	Mexico	846.11	2.43

Panel B: Top 10 countries for international bandwidth usage in 2002 and 2020

Country	Usage (Gbps)	Share (%)	Country	Usage (Gbps)	Share (%)
United States	647.40	21.77	United States	629,310.40	15.15
United Kingdom	505.37	16.99	Germany	463,945.46	11.17
Germany	301.51	10.14	France	438,407.91	10.56
France	263.69	8.87	United Kingdom	432,579.42	10.41
Netherlands	199.28	6.70	Netherlands	276,455.65	6.66
Canada	132.67	4.46	Ireland	157,918.17	3.80
Denmark	108.84	3.66	Singapore	147,822.51	3.56
Belgium	101.87	3.43	Spain	144,199.52	3.47
Sweden	94.09	3.16	China	137,111.18	3.30
Italy	80.58	2.71	Japan	134,852.63	3.25

Table 5: Trade flows and data flows

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	5.80	5	33	949.06
2003	6.58	5	35	880.09
2004	7.93	6	35	836.96
2005	9.10	6	36	804.40
2006	10.60	6	38	769.08
2007	12.22	6	38	726.53
2008	14.45	7	40	666.81
2009	11.23	7	40	688.85
2010	14.01	6	39	686.52
2011	16.87	7	40	648.65
2012	17.50	7	39	650.99
2013	17.53	7	40	662.55
2014	17.76	7	40	683.48
2015	15.74	6	39	744.31
2016	15.21	6	38	759.20
2017	16.95	6	38	739.41
2018	18.77	6	39	734.10
2019	18.46	6	39	727.59
2020	16.85	6	38	761.23

Panel B: Data flows from 2002 to 2020

Year	Total data flows (in Gbps)	Number of countries making up 50% of data flow	Number of countries making up 90% of data flow	HHI
2002	295.62	2	4	3609.49
2003	1,308.72	2	6	2993.41
2004	1,533.61	2	6	3059.24
2005	3,176.07	2	7	2833.77
2006	6,565.46	2	10	2653.67
2007	11,644.16	2	11	2479.56
2008	10,999.54	2	11	2427.24
2009	12,483.56	2	12	2404.79
2010	20,915.17	2	15	2045.28
2011	32,125.52	3	18	1826.62
2012	43,827.77	3	21	1705.40
2013	81,606.36	3	27	1388.87
2014	108,120.08	4	30	1188.78
2015	127,549.52	4	33	1175.32
2016	198,537.81	5	40	914.51
2017	316,914.44	7	46	666.61
2018	485,749.25	9	50	506.40
2019	576,618.31	9	51	458.67
2020	676,812.81	9	55	431.96

Notes: We exclude intraregional trade when computing the trade flows in goods and data.

Table 6: Top 10 markets for trade and data flows in 2002 and 2020

Panel A: Top 10 markets for trade in 2002 and 2020

Market	Trade Volume (billion USD)	Share (%)	Market	Trade Volume (billion USD)	Share (%)
East Asia - Europe	664.34	22.93	East Asia - Europe	2395.51	28.43
East Asia - North America	656.94	22.67	East Asia - North America	1484.25	17.62
Europe - North America	571.93	19.74	Europe - North America	1140.34	13.53
North America - South America	371.10	12.81	North America - South America	747.97	8.88
Europe - South America	116.47	4.02	East Asia - South America	542.03	6.43
East Asia - Oceania	82.52	2.85	East Asia - Oceania	377.72	4.48
Europe - Sub-Saharan Africa	80.95	2.79	Europe - South Asia	310.83	3.69
East Asia - South America	80.05	2.76	Europe - South America	307.92	3.65
Europe - South Asia	59.88	2.07	East Asia - South Asia	271.78	3.23
Europe - Oceania	40.86	1.41	Europe - Sub-Saharan Africa	242.65	2.88

Panel B: Top 10 markets for data flows in 2002 and 2020

Market	Data Flow (Gbps)	Share (%)	Market	Data Flow (Gbps)	Share (%)
East Asia - North America	197.70	66.88	Europe - North America	190,170.74	28.10
Europe - North America	97.67	33.04	East Asia - South America	160,504.85	23.71
East Asia - South America	0.15	0.05	East Asia - Europe	160,052.27	23.65
East Asia - Europe	0.07	0.02	East Asia - North America	51,884.46	7.67
North America - South America	0.01	0.00	North America - South America	47,416.41	7.01
East Asia - Oceania	0.00	0.00	Europe - Oceania	28,350.54	4.19
North America - Oceania	0.00	0.00	East Asia - Oceania	20,116.73	2.97
Europe - Oceania	0.00	0.00	Europe - Sub-Saharan Africa	7,535.95	1.11
Europe - South America	0.00	0.00	North America - Oceania	4,129.57	0.61
North America - Sub-Saharan Africa	0.00	0.00	North America - Sub-Saharan Africa	3,095.20	0.46

Notes: We exclude intraregional trade when computing the trade flows in goods and data.

Table 7: Top 10 countries for trade and data flows in 2002 and 2020

Panel A: Top 10 countries for trade flows in 2002 and 2020

Country	Trade Volume	Share (%)	Country	Trade Volume	Share (%)
United States	1551.95	26.78	United States	3237.25	19.21
Japan	451.76	7.79	China	2737.50	16.25
China	396.04	6.83	Mexico	814.58	4.83
Mexico	327.97	5.66	Germany	669.78	3.97
Germany	265.57	4.58	Japan	620.49	3.68
United Kingdom	211.34	3.65	India	616.40	3.66
South Korea	168.03	2.90	Australia	465.32	2.76
France	132.76	2.29	South Korea	463.76	2.75
Australia	132.50	2.29	United Kingdom	382.66	2.27
Canada	128.92	2.22	Netherlands	345.51	2.05

Panel B: Top 10 countries for data flows in 2002 and 2020

Country	Data Flow (Gbps)	Share (%)	Country	Data Flow (Gbps)	Share (%)
United States	146.97	49.72	United States	95,312.05	14.08
Japan	97.38	32.94	Canada	53,139.44	7.85
Germany	14.10	4.77	Japan	44,291.93	6.54
United Kingdom	12.91	4.37	China	36,909.16	5.45
France	7.69	2.60	Singapore	29,300.72	4.33
Italy	4.81	1.63	South Korea	27,205.98	4.02
Switzerland	2.12	0.72	Australia	21,201.08	3.13
Netherlands	1.78	0.60	Hong Kong	20,472.80	3.02
Norway	1.10	0.37	Taiwan	15,530.52	2.29
Spain	1.00	0.34	Germany	11,101.22	1.64

Notes: We exclude intraregional trade when computing the trade flows in goods and data.

Table 8: Path decomposition (2020)

Market	Endpoint regions		Demand (Gbps)	Share (%)
Europe - North America	Europe	North America	92,207.20	55.21
Europe - North America	Europe	Japan	16,477.34	9.87
Europe - North America	East Asia	Europe	10,054.47	6.02
Europe - North America	Europe	South Korea	9,436.64	5.65
Europe - North America	Australia	Europe	9,116.14	5.46
Europe - North America	East Asia	North America	7,018.03	4.20
Europe - North America	East Asia	South America	5,243.72	3.14
Europe - North America	Europe	Singapore	3,778.14	2.26
Europe - North America	Europe	Taiwan	3,657.53	2.19
Europe - North America	Europe	South America	1,197.90	0.72

Table 9: Top 10 countries for centrality in trade and data flows in 2002 and 2020

Panel A: Top 10 countries for centrality in trade flows in 2002 and 2020

Country	Centrality	Country	Centrality
United States	0.6167	China	0.5604
Germany	0.3947	United States	0.5150
Canada	0.3802	Germany	0.3922
Mexico	0.3160	Mexico	0.2745
China	0.3136	Japan	0.2426
Japan	0.2997	Netherlands	0.2311
United Kingdom	0.2851	Canada	0.2170
France	0.2572	United Kingdom	0.2018
Netherlands	0.1903	South Korea	0.1932
Italy	0.1895	France	0.1884

Panel B: Top 10 countries for centrality in data flows in 2002 and 2020

Country	Centrality	Country	Centrality
United States	0.8037	United States	0.4178
Japan	0.6905	Canada	0.3166
Germany	0.1556	Japan	0.3091
United Kingdom	0.1425	China	0.2961
France	0.0849	Singapore	0.2637
Italy	0.0532	South Korea	0.2324
Switzerland	0.0234	Germany	0.2060
Netherlands	0.0197	Hong Kong	0.2053
Norway	0.0121	United Kingdom	0.1938
Spain	0.0110	France	0.1907

Notes: We compute the eigenvector centrality of each country and year based on trade flows and data flows. Panel A shows the list of the top 10 countries for trade flows in 2002 and 2020 and Panel B shows the list for data flows.

Table 10: Model estimates: Demand for data and cable usage

Variable	Estimates
Demand Characteristics (x_{ckt})	θ^d
Constant	-207.921
	(10.070)
GDP	3.976
	(0.657)
Population	-2.382
	(0.297)
Time trend	0.654
	(0.265)
2008-2012	-1.191
	(1.384)
2013-2017	-0.237
	(1.140)
2018-2020	-0.592
	(1.009)
Market fixed effects	✓
Cable Characteristics (Z_{pckt}) θ^δ	
Constant	67.046
	(7.018)
Path length	-1.208
	(0.969)
Potential capacity	1.549
	(0.411)
Number of cables	0.351
	(0.149)

Notes: This table reports estimates for θ^D, the parameters that govern how country-pair characteristics (x_{ckt}) affect demand for data transmissions between the two countries and 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 specification also includes a linear time trend, indicators for the 2008-2012, 2013-2017, and 2018-2020 periods, and market fixed effects. The 2002-2007 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.

Table 11: Model estimates: Supply of cables

	γ
All firms	(0.0028,0.0044)
Non-content providers	(0.0034,0.0044)
Content providers	(0.0003,0.0055)

Notes: We report the 95% confidence set for the revenue parameter, γ, capturing revenue (measured in millions of US dollars) from an increase in the data flows captured by a firm $\left(D_{mt}\frac{k_{jmt}}{K_{mt}}\right)$ by one Gbps. In the first row, we estimate a common revenue parameter for all firms. In the second and third rows, we estimate separate parameters for non-content providers and content providers.

Table 12: Counterfactuals: Adding a new cable in selected markets in 2010

Panel A: Based on the lower bound of γ

Market	(1)	(2)	(3)	(4)	(5)
	Change in incumbent revenue			Entrant revenue	Business stealing
	Investment market	Other markets	Total		
North America-East Asia	-8.96	-8.65	-17.61	6.06	2.90
North America-Europe	-15.57	11.62	-3.94	18.86	0.21
North America-South America	-4.15	19.80	15.65	11.73	-1.33
Sub-Saharan Africa-East Asia	3.66	862.99	866.65	283.72	-3.05
Sub-Saharan Africa-Europe	-9.14	44.63	35.49	24.08	-1.47
Sub-Saharan Africa-South America	4.56	583.73	588.30	191.48	-3.07

Panel B: Based on the upper bound of γ

Market	(1)	(2)	(3)	(4)	(5)
	Change in incumbent revenue			Entrant revenue	Business stealing
	Investment market	Other markets	Total		
North America-East Asia	-14.08	-13.59	-27.67	9.53	2.90
North America-Europe	-24.47	18.27	-6.20	29.64	0.21
North America-South America	-6.51	31.11	24.60	18.43	-1.33
Sub-Saharan Africa-East Asia	5.75	1356.12	1361.87	445.84	-3.05
Sub-Saharan Africa-Europe	-14.36	70.13	55.77	37.85	-1.47
Sub-Saharan Africa-South America	7.17	917.30	924.47	300.90	-3.07

Notes: The revenues are reported in thousands of US dollars. Business stealing is computed as the ratio of the negative value of the change in the incumbent revenue in column (3) to the entrant revenue in column (4).

A Additional Figures and Tables

Table A1: Construction Cost Regression

Construction Cost	(1)	(2)
Log(Capacity)	0.0656	0.112**
	(-0.0483)	*(-0.0498)
Log(Cable Length)	1.096***	1.075***
	(-0.0668)	(-0.0503)
Time trend		-0.0292**
		(-0.0118)
Constant	9.060***	9.086***
	(-0.74)	(-0.646)
Observations	31	31
R-squared	0.887	0.908

Table A2: Capacity Regression

	Log Capacity
Year fixed effects	Yes
Market fixed effects	Yes
Constant	8.862***
	-0.82
Observations	81
R-squared	0.684

B Inference Procedure

We implement the conditional chi-squared (CC) test propsed by Cox and Shi (2023) to compute confidence sets for the true parameter value μ^* based on a set of unconditional moment inequalities $l = 1, ..., L$. We denote the set of L moments we use in the estimation as

$$\overline{m}(\gamma) = (m_1(\gamma), ..., m_L(\gamma))'$$

with the inequalities given by

$$\overline{m}_l(\gamma) \geq 0, l = 1, ..., L.$$

$\hat{\Sigma}(\theta)$ is an estimator of variance-covariance matrix of the moments. The procedure looks as follows.

Step 1: Define a grid $\mathcal{A}$ that will contain the confidence set.

Step 2: Choose a point in the grid $\gamma \in \mathcal{G}$. For a given significance level $\alpha \in (0, 1)$, the following steps test the null hypothesis given by $H_0 : \gamma^* = \gamma$.

Step 3: Evaluate the quasi-likelihood ratio statistic at γ as

$$T(\gamma) = \min_{\mu:\mu\geq 0} N(\overline{m}(\gamma) - \mu)'\hat{\Sigma}(\gamma)^{-1}(\overline{m}(\gamma) - \mu) \tag{9}$$

where N is the sample size. The vector μ is of dimensions $L1$. When defining the objective function of the minimization problem in equation (9), it occasionally happens that the covariance matrix $\hat{\Sigma}(\theta_p)$ is singular or close to singular. When this is the case, we follow Andrews and Barwick (2012), and substitute the covariance matrix in equation (9) for the following matrix:

$$\tilde{\Sigma}(\gamma) = \hat{\Sigma}(\gamma) + \max\{0.012 - \det(\hat{\Omega}(\gamma), 0\} Diag(\hat{\Sigma}(\gamma) \tag{10}$$

where $Diag(\hat{\Sigma}(\gamma))$ is the $L \times L$ diagonal matrix whose diagonal elements are equal to those of $\hat{\Sigma}(\gamma)$ and $\hat{\Omega}(\gamma)$ is the correlation matrix of the moments evaluated at γ.

Step 4: Count how many values of μ equal 0. We denote this number as $\hat{r}$.

Step 5: Accept/reject θ_p. Include γ in the $(1-\alpha)\%$ confidence set, $\hat{G}^{1-\alpha}$, if $T(\gamma) \leq \chi^2_{\hat{r},1-\alpha}$, where $\chi^2_{\hat{r},1-\alpha}$ is the $100(1-\alpha)\%$ quantile of $\chi^2_{\hat{r}}$, the chi-squared distribution with $\hat{r}$ degrees of freedom.

Step 6: Repeat steps 2 to 5 for every γ in the grid