THREE ESSAYS ON THE COST OF INTERNET COMMUNICATIONS: MEASUREMENT AND IMPLICATIONS FOR INTERNATIONAL ECONOMIC FLOWS

by

ZACHARY KIEFER

A DISSERTATION

Presented to the Department of Economics
and the Division of Graduate Studies of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

June 2022

DISSERTATION APPROVAL PAGE

Student: Zachary Kiefer

Title: Three Essays on the Cost of Internet Communications: Measurement and Implications for International Economic Flows

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Economics by:

Anca Cristea	Co-Chair
Woan Foong Wong	Co-Chair
Bruce Blonigen	Core Member
Reza Rejaie	Institutional Representative

and

Krista Chronister	Vice Provost for Graduate Studies

Original approval signatures are on file with the University of Oregon Division of Graduate Studies.

Degree awarded June 2022

DISSERTATION ABSTRACT

Zachary Kiefer

Doctor of Philosophy

Department of Economics

June 2022

Title: Three Essays on the Cost of Internet Communications: Measurement and Implications for International Economic Flows

I develop new measures of Internet communication costs that hold advantages over similar measures previously used in the economics literature: they are more solidly based on the technical nature of the Internet, easier to compute, and/or more suitable to use in international economics. To do this, I introduce a pair of novel data sources that describe distinct aspects of Internet communication. I further demonstrate that my developed measures possess explanatory power when used to explain patterns of trade in goods and in services, as well as cross-border portfolio investment.

CURRICULUM VITAE

NAME OF AUTHOR: Zachary Kiefer

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, OR, USA
Washington State University, Pullman, WA, USA

DEGREES AWARDED:

Doctor of Philosophy, Economics, 2022, University of Oregon
Bachelor of Science, Economics, 2012, Washington State University

AREAS OF SPECIAL INTEREST:

International Economics
Development Economics
Game Theory

PROFESSIONAL EXPERIENCE:

Graduate Employee, Department of Economics, University of Oregon, 2017-2022

GRANTS, AWARDS AND HONORS:

Outstanding Third-Year Paper, Department of Economics, University of Oregon, 2019

ACKNOWLEDGEMENTS

The author extends his wholehearted thanks to Drs. Anca Cristea, Woan Foong Wong, and Bruce Blonigen of the University of Oregon Department of Economics, to Dr. Reza Rejaie of the Department of Computer Science, and to the staffs of the Oregon Route Views Project and the Center for Applied Internet Data Analysis.

TABLE OF CONTENTS

Chapter Page

Chapter Page

APPENDICES

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

Table Page

CHAPTER I

INTRODUCTION

In the last four decades, an increasing fraction of communication has taken place by way of the Internet, thanks to email, Voice Over Internet Protocol (VoIP) phone systems, and now the widespread use of videoconferencing. As the Internet has matured, it has become an important factor in many areas with relevance to economics: access to the Internet means access to amenities including instant communication, education, financial and commodity markets, and others. This makes Internet access into a plausible reducer of information frictions as barriers to trade and other international economic flows.

Internet access can be measured along multiple margins: the simplest to measure is likely the extensive margin, i.e. the number of people with any degree of Internet access, but there is also the intensive margin (the degree to which people with Internet access use that access) and what might be called the cost and quality margins, measuring the effectiveness of the available Internet services. Each of these margins potentially affects economic variables in distinct ways: for example, expansion of Internet access along the extensive margin may give a wider segment of the population access to human capital development via e-learning, but seems unlikely to result in an expansion of a country's tech industry unless accompanied by improvements along the intensive or quality margins.

Unfortunately, it has proven difficult to measure any of these margins of Internet access, or the costs economic agents face in using the Internet–especially in ways that are familiar and useful to economists. Particularly in the international economics literature, it is most useful to measure communication costs bilaterally, i.e. pertaining to communication between sender-receiver pairs. Historically, it was more straightforward to do this, as the dominant communication technologies were telegraph or telephone. Both technologies were billed on a per-unit, bilateral basis, making it as simple as looking up the published rate for telegrams or phone calls between a pair of locations. The Internet, however, is generally billed on only a unilateral basis, with rates that do not vary depending on where a user intends to communicate with.

Complicating this further is the fact that due to the decentralized nature of the Internet, there is no singular authority that collects data on the entire Internet. Existing measures of Internet access exhibit notable flaws: the largest datasets describing Internet access and cost are survey-based, and suffer from the issue of comparability across locations or time periods. Non-survey-based datasets are not common, and do not generally cover wide areas, and while proxies for Internet access and cost exist, many are outdated due to technical changes, while others require heroic efforts to collect.

The problem facing economists is thus two-part: firstly, since the Internet is now the dominant communications technology, we need to develop a way of measuring a bilateral Internet communication cost. Secondly, we ideally need for this new measure to be readily generated from the fragmentary data that is available.

In this dissertation, I develop novel sources of data which I use to construct measures of Internet access, focusing on the extensive, cost and quality margins. These data are publicly available and are based on fundamental technical aspects of how the Internet is run, making them unlikely to become obsolete any time soon. The processes I use to construct my measures are designed to be computationally accessible (requiring only modest computational resources to compute and estimate), and I show that they can be used to answer research questions including:

- Does expanded Internet access result in a reduction of barriers to trade?
- Do specific economic sectors' trading behaviors respond differently to changes in Internet access?
- Does cross-border portfolio investment respond to improved Internet access in the same way as trade?

In the first chapter, I first demonstrate the explanatory power in these data by constructing simple, unilaterally-varying measures of communication cost which I apply in replications of earlier work from the literature (Freund and Weinhold (2004) and Allen (2014)). In the second chapter, I expand on these measures, adapting a structural model of physical transportation in such a way as to estimate a bilateral Internet communication cost analogous to iceberg trade costs. In this and the third chapter, I demonstrate the explanatory power of the estimated communication costs in gravity models of trade and portfolio investment. Results of the replication and gravity models show that these measures contain explanatory power comparable with previous measures used in the literature. The results of my gravity models also confirm results seen in Keller and Yeaple (2013).

CHAPTER II

ROUTING DATA AS A MEASURE OF INTERNET ACCESS: AN INTRODUCTION TO THE DATA AND ITS EXPLANATORY POWER

2.1 Introduction

In this chapter, I introduce the first of my novel data sources, the Internet routing data compiled by the Oregon Route Views Project. Based on this data, I propose a pair of new measures of Internet access. These measures are straightforward to compute, and can be computed at varying levels of geographic detail with little added effort. I then demonstrate empirically that these measures perform comparably to measures previously used in the literature by adapting the work of Freund and Weinhold (2004) and Allen (2014), and that the two measures capture largely separate aspects of Internet access.

2.2 Literature Review

Perhaps the earliest available measure of Internet access is the UN Statistics Division's measure of "Internet Users per 100 Inhabitants," which is available at the country-year level starting in 1990. However, this measure is compiled from surveys administered by the statistical agencies of many different nations, and therefore suffers from comparability issues. The metadata for this data series states upfront that there may be discrepancies when the age scope of national surveys differs[1]), when the survey administrators use different definitions of "Internet user[2]," or when the number of Internet users is estimated from a number of Internet subscriptions. Other survey-based measures suffer from similar limitations, or are limited in scope to single countries.

[1] Did the survey include minors, who use the Internet with greater intensity than, e.g., senior citizens?

[2] Is someone who gets email but does not browse websites an Internet user?

Similar surveys used in the literature include the World Bank Investment Climate Surveys, which measure the percentage of manufacturers with Internet access, and the surveys used by the International Telecommunications Union (Clarke and Wallsten (2006)). These surveys suffer from the same issues of comparability across countries. In studies of a smaller scope, such as those limited to a single country, the issue of comparability is less problematic, as it becomes possible to use single surveys which are presumably administered in a uniform manner (Fan and Salas Garcia (2018)).

In order to avoid issues of comparability on a global scale, Freund and Weinhold (2004) uses a proxy for Internet access, consisting of a count of web hosts[3] attributed to each country. This approach has flaws, as the authors are aware: hosts which end in generic domains such as ".com," ".edu," etc. cannot be attributed to any particular country, and additionally, even hosts with country-specific domains could be physically located anywhere: the country-specific domain only indicates the audience that the website is aimed at.

Also, since the publication of Freund and Weinhold, the Internet Corporation for Assigned Names and Numbers (ICANN), an NGO non-profit which regulates some aspects of the Internet, has greatly expanded the set of top-level domains to include such generic suffixes as ".community" and ".horse." These generic suffixes likewise cannot be attributed to a particular country, and it is likely that a growing proportion of webhosts will use these domains in the future. If so, this limits the usefulness of the webhost-counting measure of Internet access moving forward.

[3] Such as www.bbc.co.uk, registered in the United Kingdom, or www.amazon.nl, registered in the Netherlands.

Allen (2014), while not directly focusing on Internet access, deals with the related topic of information frictions, using a measure derived from data on cell phone tower construction in the Philippines. This data appears to no longer be available: the Asia Pacific Policy Center (APPC), the NGO which compiled this data, seems to have closed its doors, and I have been unable to determine the current custodian of its data.

Even if this data were available, Allen states that the APPC expended "substantial effort" in digitizing the registration records of the universe of Philippine cell towers, and this dataset is, naturally, limited to the Philippines. This approach to measuring information frictions does not appear to be scalable to analyses of wider scope.

Finally, all of these measures address only how widely Internet (or cell phones) are available in a country: there are few measures which address the quality of Internet access, as described by latency, reliability, or cost. One proxy for quality of access used in the literature is whether a firm subscribes to broadband Internet (Grimes, Ren, and Stevens (2012)), but this is merely a coarse proxy[4], and again relies on a micro-survey approach.

In looking for an objective measure of Internet access, I have been inspired by the approach taken in Chen and Nordhaus (2011), in which the authors use luminosity data as a proxy for economic activity. The advantage of this approach is that, despite being only an indirect measure of economic activity, the luminosity data is easily obtained, easily processed, and objective: it is measured in the same way in every country, and thus serves as an excellent proxy when more accurate and detailed data are not available. Given that this is a similar situation, in which

[4] "Broadband Internet" can refer to a range of technologies with different qualities, not to mention costs.

detailed and/or accurate data on Internet access are not always available, a similar approach is justifiable here.

2.3 Data

This chapter centers on two novel measures of Internet access which can be computed from a previously unused, publicly available source of data on Internet routing—the process by which information is transmitted via the Internet. Before defining these measures, I will first briefly summarize how Internet routing works in order to provide context for the two measures which I propose.

2.3.1 Routing Data. The Oregon Route Views Project (ORVP), hosted at the University of Oregon, hosts an archive of routing data from sources around the world. This data is output from the Border Gateway Protocol (BGP), the algorithm which Internet-connected devices use to share routing information with each other, and describes the routes that select collectors would use to send communication to destinations around the world.

The Internet is not a uniform network; rather, it is comprised of many smaller networks linked together into a whole. These networks include Internet Service Providers, telecommunication companies, Internet-based companies like Google and Amazon, and in fact any collection of network infrastructure run as a cohesive unit by a single organization. These networks can of course route communication between their own users, but in order to communicate outside of their own boundaries, they must send communication through other networks; to do this effectively, they must communicate information about good routes to use. Internet Exchange Points (IXPs) are points at which many networks are connected together and exchanging routing information, resulting in the IXP being highly-informed about the best routes available. The routing data hosted by the ORVP is collected by select IXPs, and because of their highly-connected nature, these collectors provide an unusually good viewpoint on Internet routing.

The unit of observation in this routing data is a route, defined here as a sequence of networks that can be used to pass communication from the collector to a target device elsewhere in the Internet. Routes generally serve a block of devices identified by their Internet Protocol (IP) addresses. There are commonly multiple routes that serve a single block, and there are cases in the data where one route serves a block that is a subset of another block with its own set of routes.

There are around 30 collectors contributing to the ORVP, some of which have been contributing observations taken every two hours since 2003. Rather than process the entirety of this vast body of data, I take samples from the first of each month at three collectors: PAIX, the Palo Alto Internet eXchange, EQIX, the Equinix-Ashburn exchange, and LINX, the London InterNet eXchange, chosen for their size and long-running contribution to the ORVP.

2.3.2 Internet Routing in Action. Every Internet-connected device possesses an IP address, a number which uniquely identifies the device to other Internet-connected devices. IP addresses are allocated to Internet Service Providers (ISPs) and other entities, which in turn assign addresses to consumer devices.

When an Internet-connected device needs to communicate with another, it consults its internal routing table, which contains instructions for communication with other devices. These instructions take the form of a table of identifying numbers (Internet Protocol or IP addresses) for other devices, along with a sequence of other devices that can act as intermediaries to forward communication to those devices. This sequence of devices can be described as a route or path.

In the case of most consumer devices, the routing table is not particularly detailed, and does not contain specific instructions for a large number of other devices: when attempting to communicate with devices that are not directly connected to them, they instead instruct the device to forward its communication to an Internet router with a more complete routing table and let it take over. The router may forward it again, to a better-connected router, as may the next, and so on, until the communication reaches a router that actually does have instructions in its routing table that enable it to send the communication to its actual destination.

A typical consumer device's routing table will only contain a handful of entries, since it only needs to communicate with a handful of other devices: the next hop in virtually all of its routes will be a router operated by the user's ISP. However, highly-connected devices, such as the routers in Internet Exchange Points (IXPs) where multiple ISPs connect their networks together, have much more detailed routing tables. At very large IXPs, devices' routing tables may have detailed information on how to communicate with ~95% of the roughly 4 billion IP addresses in existence.

To use an analogy, the routing table is similar to a set of instructions for sending physical mail: when a typical consumer wishes to send a letter or package, they do not need to know the exact route that it will take to reach the recipient. All they need to know is how to get the package to the post office, after which it is up to the postal service to get the package to its recipient. The postal service, on the other hand, needs to have detailed instructions about how to get the package from any given point A to point B[5].

[5] These detailed instructions need not even be available at individual post offices: each post office only needs to know where to send the package next, based on its destination, and this is similar to how Internet routing works in practice. However, somewhere, perhaps at a

Extending this analogy, if one were to obtain the post office's master list of instructions for how to ship packages between any origin and destination, one could draw conclusions about where people who used the post office were located, based on where there were concentrations of post offices, or approximate the speed of mail service between towns A and B based on how many steps were in the instructions. The measures of Internet access I develop in this chapter are based on a similar line of thinking, in that the routing data possessed by large IXPs conveys information about where Internet users are located and how good their Internet access is.

2.3.3 Counting IP Addresses. The first measure which this data allows me to construct is a simple count of how many IP addresses are in use in any given country or province. This can be done in other ways–for instance, by consulting the official registry of IP addresses allocated to each country, or a geolocation database that maps IP addresses to provinces. However, the official registry only captures the number of IP addresses allocated, and it is relatively common for addresses to be allocated but not assigned (attached to an end-user who is using it to actively communicate). Geolocation databases face a similar issue, in that they are commonly based on the blocks of IP addresses allocated to ISPs, without consideration for whether those addresses are assigned. Counts of IP addresses based on these data sources can thus only be interpreted as an upper bound on the number of IP addresses actually assigned and in use.

I attempt to refine these approaches by coupling a commercial geolocation database (Maxmind) with routing data. Because the collectors which provide my routing data are located in highly-connected IXPs, any routes listed in their routing data are valid, and can actually be used to reach the block of IP addresses

headquarters, the authority responsible for distributing instructions to post offices must have a master list of instructions, and this is analogous to the data I actually use.

associated with it.[6] The IP addresses observed in in this routing data can therefore be viewed as the set of all IP address blocks which have recently been observed in use by the collector. While not a perfect measure–it is still only an upper bound on the number of assigned IP addresses, since blocks may include some unassigned IP addresses–when coupled with geolocation data it provides a tighter and thus superior bound than a count based purely on the geolocation data.

In constructing this measure, I begin by constructing an exhaustive list of each unique IP address block for which the routing data contains at least one route[7]. Then, using the Maxmind geolocation database[8] I identify each block's location at the country and province level. This enables me to construct a variable tracking the number of IP addresses in use, again at the country or province level, over the time period from 2003 to 2018. Having done this for all three of the collectors I draw data from, I then take the mean of the three values.

This measure does not directly capture the extent of Internet access, defined as the number or percentage of residents in an area that have Internet access. It is complicated by multiple factors: since users may be associated with multiple IP addresses (one at their home and one at their place of work, for example), it does not map one-to-one into a count of Internet users, and the ratio of Internet users to IP addresses is likely to vary across countries. This is further complicated by the unobserved presence of unassigned IP addresses; both factors would be likely

[6] Without going into excessive technical detail, the Border Gateway Protocol ensures that when an IP address block can no longer be reached via a route, that knowledge propagates rapidly and the route is struck from the routing table.

[7] As the address blocks may overlap, I process them to form a set of disjoint blocks while preserving the associated route data.

[8] This is the same kind of database used by websites and advertisers to determine the location of webpage visitors: if you've ever visited a webpage that appeared to know where you were, it was probably using a similar database.

to cause measurement error and thus attenuation bias when I apply these measures in my empirical section. The measure also does not directly capture the intensity of Internet access, defined as the volume of communication that an area produces. However, it is plausibly correlated with both of these quantities: ceteris paribus, one would expect a region with more IP addresses to have more residents with Internet access, and to generate more Internet communication.

2.3.3.1 Descriptive Statistics. Table 1 presents descriptive statistics for the IP address count at two different levels of aggregation: the country-level, and the province-level within the Philippines.[9]

Table 1. Descriptive Statistics: IP Address Count

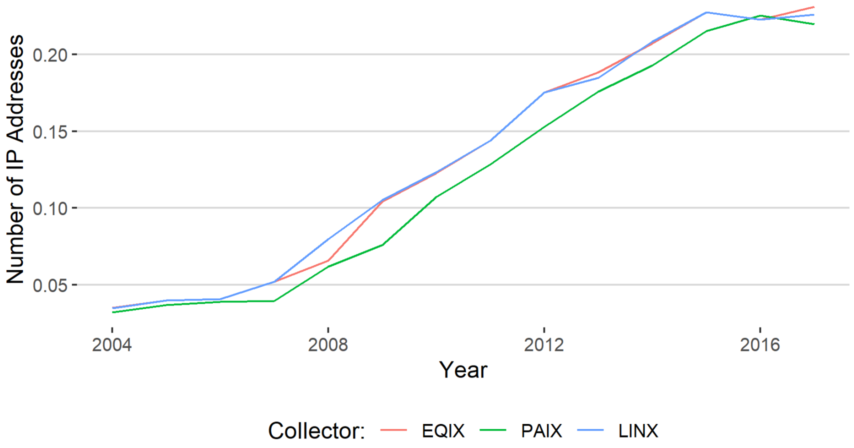
Statistic	N	Mean	St. Dev.	Min	Max
World, by Country	10,337	11,726,450	78,863,537	256	2,172,239,716
Philippines, by Province	3,053	54,366	132,317	256	1,000,448

There is substantial variation in the IP address count across time and location. Part of this is due to variation in national population size, but much of the variation remains in the per-capita IP address count: Figure 1 shows the per-capita IP addresses in the median country over time, as observed from the three collectors, while Figure 2 shows per-capita IP address counts around the world in 2004, 2010, and 2016.

The three collectors which I use data from report highly-correlated IP address counts. The correlation coefficient is 0.952 between the values reported by EQIX and PAIX, 0.862 between EQIX and LINX, and 0.904 between PAIX and LINX, indicating a high degree of consistency between collectors.

[9] I have chosen the Philippines in particular because of relevance to Allen (2014).

Figure 1. Median National Per-Capita IP Addresses



2.3.4 Measuring Route Length.

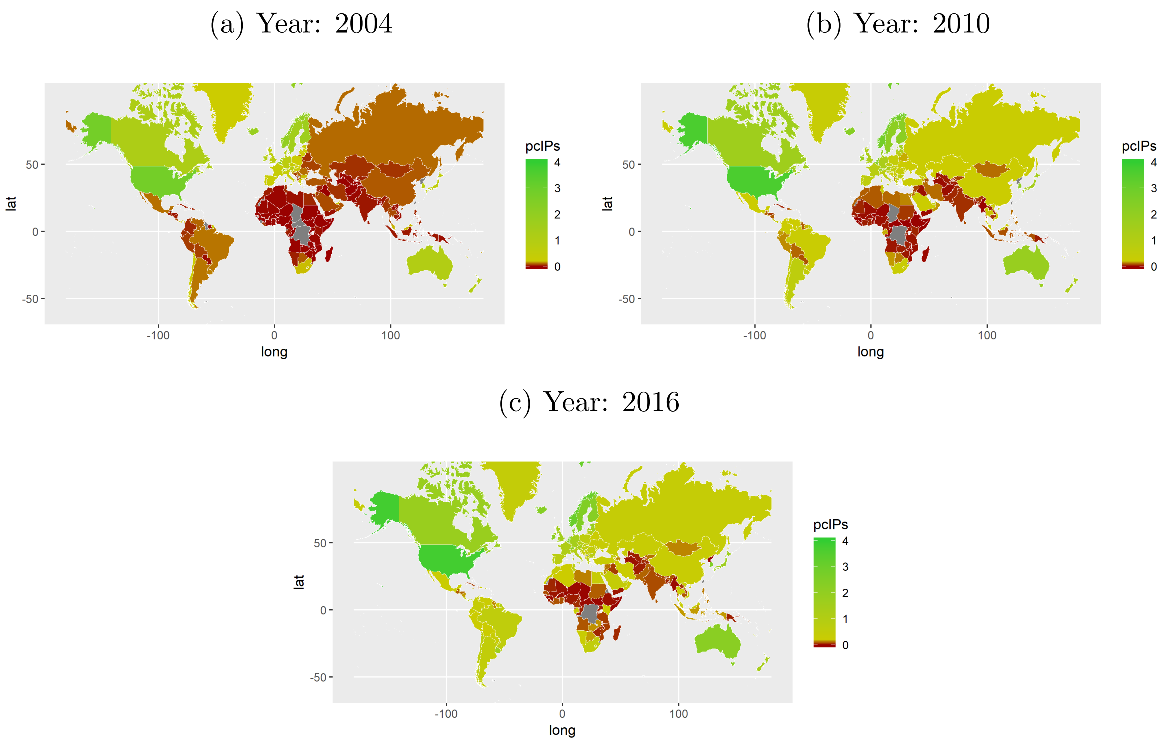
The second of my proposed measures is Aggregate Route Length (ARL). This is a measure of how complicated it is to send data from a major IXP to a target location.

One of the variables in the routing data is the Autonomous System Path (ASP)[10], an ordered list of all the networks that a route would pass through. It can be thought of as a list of all the organizations whose cooperation is necessary to use the route. I use this as a measure of route length, and for the purposes of this paper, any references to "route length" refer specifically to the length of the ASP. Route length is plausibly correlated with several factors relating to the quality of Internet access experienced by the IP addresses in the block:

- The cost of using a route is plausibly positively correlated with the route's length. Each network incurs costs (monetary costs such as electricity and wear and tear on equipment, and opportunity costs that may be incurred from congestion if a network is heavily trafficked) to carry Internet

[10] "Autonomous System" being the formal name for what I am otherwise referring to as a "network."

Figure 2. Per-Capita IP Addresses Around the World



(a) Year: 2004

(b) Year: 2010

(c) Year: 2016

communication to its destination: economically, we would only expect the network to do so if it can extract some benefit from doing so. This benefit may be monetary (an access fee), reciprocity (an agreement with another network to carry each others' traffic), or some other form, but in each case, every network in a route adds some cost to the route. To phrase this differently, if it were possible to remove a network from a route without altering the rest of the route[11], the modified route would be expected to be lower-cost. Assuming that networks only participate in a route if they can recoup their costs, these costs would thus be passed on to the end-user who

[11] Specifically, if it were hypothetically possible to remove network B from a route "A-B-C", without altering the points at which the route left A and entered C.

pays for the service. Under this assumption, regions that are served by longer routes would be expected to have costlier Internet access.

- The latency (time required to transmit communication) of Internet access through a route is plausibly positively correlated with the route's length. Each network in the route represents additional computational steps that slow the transmission of communication.

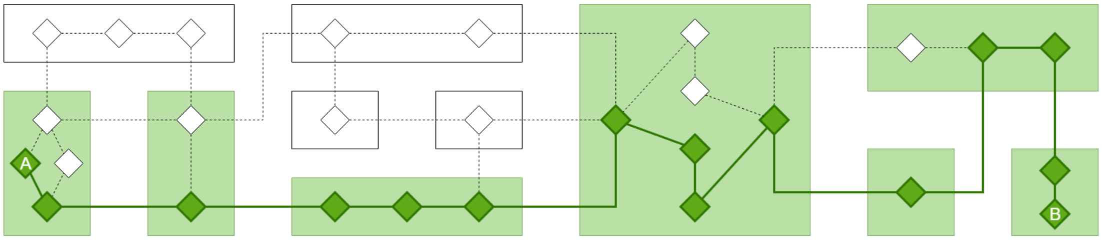
- The reliability (percent uptime, rate of successful communication, etc.) of Internet access through a route is plausibly negatively correlated with the route's length. In addition to adding computational steps, each network in a route adds more potential points of failure to the route. These include opportunities for hardware failure (severed cables or crashed servers) that shut down routes and transmission errors (dropped or corrupt packets) that prevent communication from being successfully delivered.

Route length can thus be used as a measure of the quality margin of Internet access, although it is not readily possible to disentangle it further into cost, latency, and reliability. As IP address blocks can be served by multiple routes, the cost, latency, and reliability of Internet access to a block are thus plausibly correlated with some aggregate (a simple mean, for example) of the lengths of the routes serving the block.

Figure 3 shows a simple section of Internet, with a particular route marked out in green. Its associated ASP is indicated by the light green boxes, with each such box representing one network.

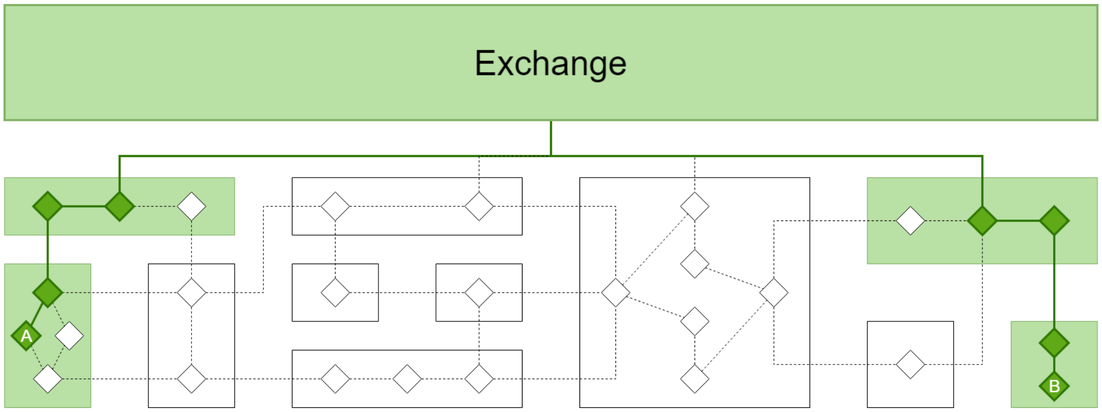
However, I am unable to observe these routes directly, as the routing data I use contains only routes that terminate in the three collectors I work with. Instead

Figure 3. A Sample Route



of seeing a route directly from A to B, I am only able to observe routes from a third, outside location to A and B, such as those illustrated in Figure 4. The routes from the Exchange to devices A and B shown in this diagram each have length 2, and I infer the existence (if not necessarily the optimality) of a route with length 5 that connects A and B via the Exchange. Such an inferred route could be used as the basis a bilateral measure of route length (and indeed I do use such a measure in some of my empirical models), but the route length taken from such a route would provide only an upper bound on the shortest-length route, as there might easily exist shorter routes that do not pass through the exchange and are thus unobserved.

Figure 4. Routes Through an Exchange



Based on these three assumptions, the length of the route contains information about the quality of Internet access for the end-users of the IP addresses at the end of a route. It should be noted, however, that the exact relation between route length and these metrics does not appear to have ever been quantified in the computer science literature: there appears to be a general consensus that path length is positively correlated with latency, but the computer science literature does not concern itself with quantifying this relationship in a way which economists would find satisfactory. (Da Lozzo, Di Battista, and Squarcella (2014); Doan, Bajpai, Ott, and Pajevic (2019))

I construct the ARL measure by the following steps:

- Firstly, because IXPs maintain records of multiple routes to many target IP address blocks, I select the route of minimal length (i.e. the route with the fewest networks) to each destination block. While this is a simplistic selection criterion, it is heavily used: in practice, route selection is carried out by automated processes that generally place high importance on route length: to cite some examples, the default selection algorithm on Cisco routers ranks route length as the fourth of eleven criteria, while Juniper ranks it fifth of fifteen and Huawei ranks it sixth of fourteen. (Cisco (2016), Juniper Networks (2020), Huawei (2019)) The higher-ranking criteria are generally values set by network administrators (such as a manually-set "local preference" attribute) and do not appear in my routing data. Thus, I treat the length of this minimal-length route as the route length for the associated IP address block.

- I then aggregate to the province or country level, using the Maxmind geolocation database to determine the location of the IP addresses in the block. Because IP addresses are reassigned periodically—they "move around" over time—I use a set of historical Maxmind databases from different time periods, so that the geolocation data is as close as possible to contemporaneous with the routing data. Because IP address blocks may be of wildly different sizes, I construct a mean of the route length for devices within a region, weighting by the size of the IP address block so that each individual IP address receives equal weight. This is the ARL for a given location.

- Finally, having performed this aggregation using routing data from all three of the EQIX, PAIX, and LINX IXPs, I take the mean of the Aggregated Route Length across the three collectors, in order to obtain a measure which is more representative of access to the global Internet, as opposed to access to a particular IXP.

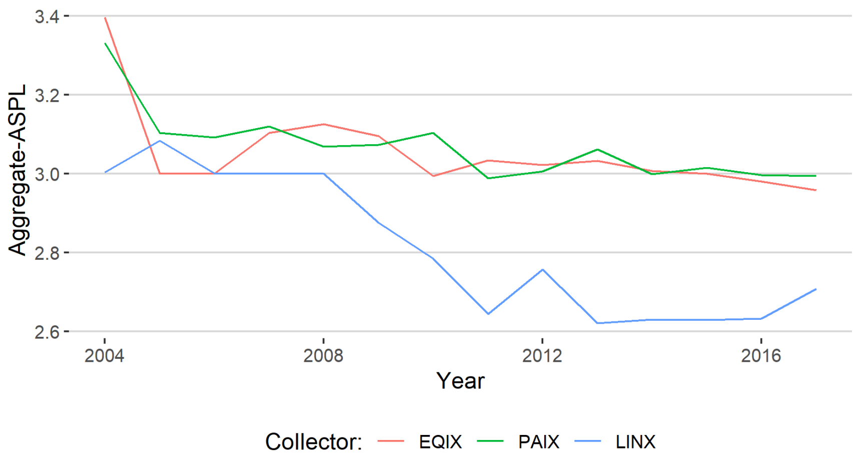
In essence, ARL captures how difficult or complex it is for Internet users to receive data from a non-local Internet Exchange Point (and by extension, how difficult it is to send data back). This is the major bottleneck in sending data to geographically distant destinations: much of the difficulty in sending data internationally is in getting the data from the sender to an IXP, and then from an IXP to the recipient. In the middle (between IXPs), the data can often be sent via an Internet Backbone, a high-speed, high-bandwidth, international connection.

This approach is limited by two factors: firstly, my method of identifying the actual route used to reach a block is limited by the available data: there are criteria in route selection that would outweigh route length, and without data on the higher-ranking criteria, this may not be accurately measuring the route length of the actual routes used. Secondly, my approach to geolocating IP address blocks is only as accurate as the Maxmind database–which certainly has its flaws, especially when trying to geolocate at the province-level or below outside of its ”favored” countries, as described in Poese, Uhlig, Kaafar, Donnet, and Gueye (2011).

Table 2. Descriptive Statistics: Aggregate Route Length

Statistic	N	Mean	St. Dev.	Min	Max
World, by Country	10,337	3	1	1	14
Philippines, by Province	3,053	3	1	2	9

Figure 5. Median National Aggregate Route Length



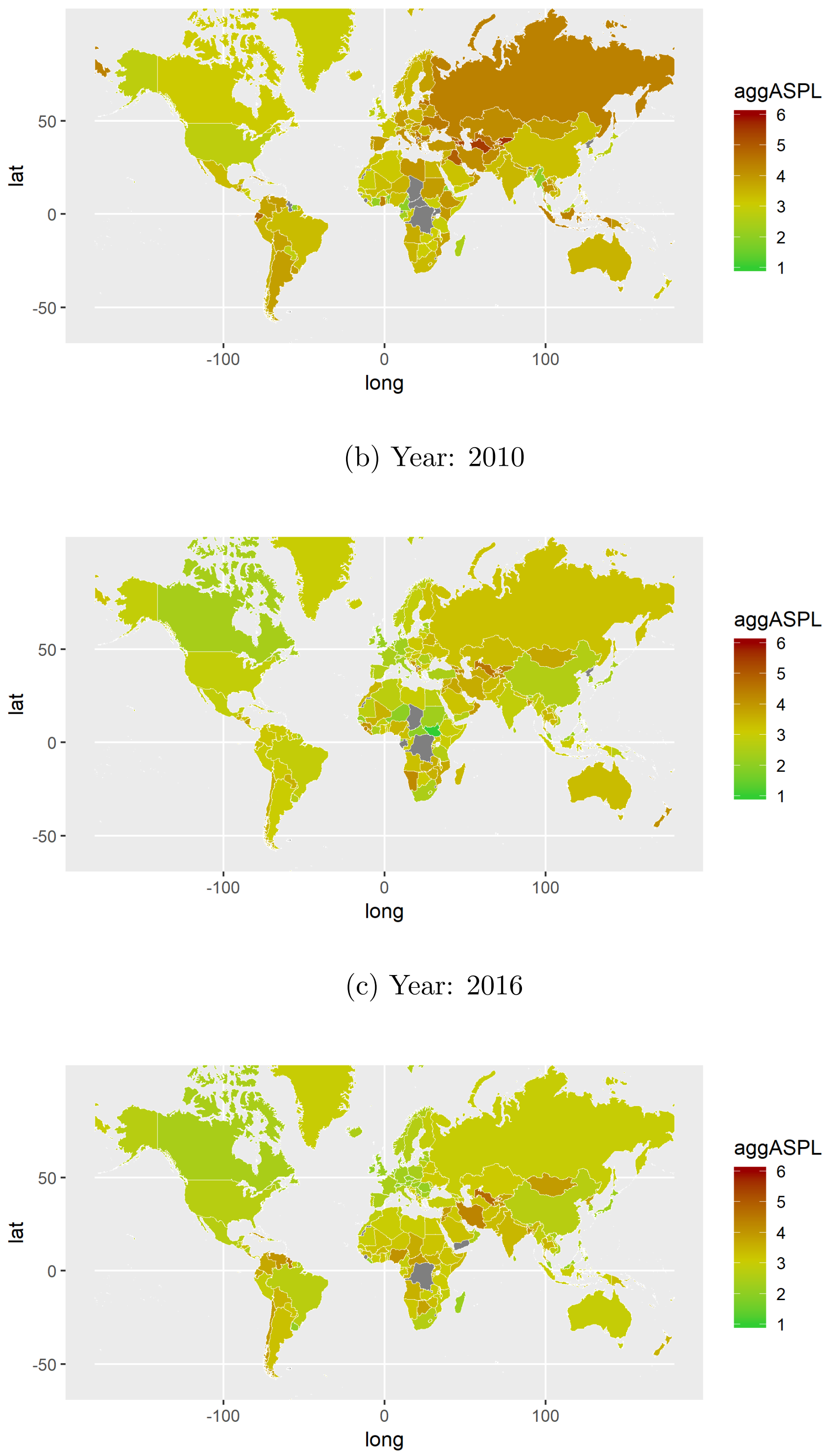
2.3.4.1 Descriptive Statistics. Table 2 presents descriptive statistics for the ARL measure at the same two levels of aggregation.

There is again significant variation across time and location, as seen in Tables 5 and 6. However, this ARL measure contains noticeably more noise, both year-to-year and between collectors. Notably, while PAIX and EQIX report similar median ARL values, LINX reports a significantly lower median.

Much as with the IP address count, there is a positive correlation between the ARL values reported by different collectors, with coefficients of 0.667 between EQIX and PAIX values, 0.865 between EQIX and LINX, and 0.663 between PAIX and LINX. These correlations are somewhat weaker for Aggregate Route Length than for the IP address count—as would be expected, since the route to a target location would be heavily influenced by the location of the collector. Again, LINX

Figure 6. Aggregate Route Length Around the World

(a) Year: 2004



(b) Year: 2010

(c) Year: 2016

reports shorter routes than PAIX, which may be due to London's advantageous position at the end of multiple undersea cables, allowing for shorter routes to many distant locations.

2.3.5 Limitations to Geolocation Data.

While the Maxmind geolocation data used in the computation of both of these measures appears reliable on the country level, it is less reliable at the province level—particularly in developing countries. In the Philippines, the reported accuracy radii are sufficiently large (in many cases, greater than 50 km), and the provinces are sufficiently small, that the true location of an IP address block may lie in neighboring provinces. Additionally, there are indications that the location attributed to IP address blocks is based on the location of the ISP or other organization which owns them, rather than the location of end-users.

The practical effect of this limitation is that there exist some Philippine provinces which, according to this geolocation data, contain no IP addresses (and thus, no Internet users) at all. This seems improbable, but in the absence of better alternatives, I have proceeded to use this data in my empirical work. This flaw does not appear when working at the country level, but to the extent that it does exist, it would cause attenuation bias due to measurement error.

There exists a commercial version of the free datasets which I use in this process, which purports to offer greater accuracy at the province level and below. It may therefore be possible to refine the province-level geolocation process in future work.

2.3.6 Economic and Other Data.

My empirical work demonstrating the value of these measures revolves around replications of two papers: Freund and Weinhold (2004) and Allen (2014). Where possible, I have obtained economic and demographic data from the same sources as the original authors: in the case of Freund and Weinhold, I use trade-flow data from the IMF Direction of Trade Statistics, as well as other economic and demographic data series from the IMF. In the case of Allen, I use the data provided by the author in his replication files.

2.4 Empirical Results

2.4.1 Freund and Weinhold Replication.

I begin by closely replicating the trade-growth models from Table 3 of Freund and Weinhold (2004), substituting my proposed measures of Internet access. In this replication, I take trade data from UN Comtrade (United Nations (2003)) as well as control variables from the Penn World Table (Zeileis (2021)). First, I estimate a baseline model

$$\begin{aligned} gExports_{ijt} = {} & \beta_0 ln(Export_{ij})_{04} + \beta_1 (gGDP_j)_t + \beta_2 log(Distance_{ij}) + \\ & FE_t + \epsilon_{ijt} \end{aligned} \tag{2.1}$$

Here, i indexes origin (exporting) countries, j indexes destination (importing) countries, and t indexes year. $gExports_{ijt}$ is the growth in exports from i to j between years $t-1$ and t, $(gGDP_j)_t$ is the growth in j's GDP and $log(Distance_{ij})$ is the log of the distance between the centroids of i and j. Results of this estimation are reported as Model (1) in Table 3.

I next introduce my IP address count, using the specification

$$\begin{aligned} gExports_{ijt} = {} & \beta_0 (gNumIPs_i)_{t-1} + \beta_1 (gNumIPs_j)_{t-1} + \beta_2 ln(NumIPs_i)_{04} + \\ & \beta_3 ln(NumIPs_j)_{04} + \beta_5 ln(Export_{ij})_{04} + \beta_6 (gGDP_j)_t + \\ & \beta_7 log(Distance_{ij}) + FE_t + \epsilon_{ijt} \end{aligned} \tag{2.2}$$

where $(gNumIPs_i)_{t-1}$ and $(gNumIPs_j)_{t-1}$ are the growth in the count of IP addresses contained in countries i and j, respectively. $ln(NumIPs_i)_{04}$ and

Table 3. Freund and Weinhold Replication Using IP Address Count

	Dependent variable:				
	Growth of exports from country 1 to country 2				
	(1)	(2)	(3)	(4)	(5)
Lag Orig. IP Growth		0.074***	0.031	0.028	0.029
		(0.029)	(0.019)	(0.020)	(0.020)
Lag Dest. IP Growth		0.016	0.032*	0.031*	0.020
		(0.023)	(0.018)	(0.018)	(0.016)
Log 2004 Orig. IPs		0.009***	0.005**	0.007	0.008
		(0.003)	(0.002)	(0.005)	(0.005)
Log 2004 Dest. IPs		0.007***	0.005**	0.007*	0.011***
		(0.003)	(0.002)	(0.004)	(0.004)
Log 2004 Exports	−0.011***	−0.013***	−0.009***	−0.012***	−0.025***
	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)
Dest. GDP Growth	0.300***	0.313***	0.322***	0.250***	0.238***
	(0.064)	(0.067)	(0.053)	(0.056)	(0.054)
Log Distance	0.002	−0.005	−0.0004	−0.008	−0.021***
	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
Orig. Real Exch. Rate Growth				−0.123**	−0.154***
				(0.060)	(0.059)
Dest. Real Exch. Rate Growth				−0.057	−0.141***
				(0.051)	(0.050)
Log 2004 Orig. GDP				−0.006	0.002
				(0.008)	(0.007)
Log 2004 Dest. GDP				−0.005	−0.007
				(0.007)	(0.007)
Log 2004 Orig. Pop.				0.014***	0.020***
				(0.004)	(0.004)
Log 2004 Dest. Pop				0.012***	0.021***
				(0.004)	(0.004)
Lag Export Growth					−0.333***
					(0.009)
Fixed Effects	t	t	t	t	t
Observations	48,125	42,657	42,091	41,605	41,652
R^2	0.013	0.015	0.022	0.024	0.178
Adjusted R^2	0.013	0.015	0.022	0.024	0.177

*p<0.1; **p<0.05; ***p<0.01

$ln(NumIPs_j)_{04}$ are the logged count of IP addresses in those countries in 2004, the first year of the sample, used as a control for initial conditions. This model is reported as Model (2) in Table 3, and Model (3) of that table is the same specification but estimated after deleting all observations which had a residual more than four standard deviations from zero in model (2).[12]

In Models (4-5) of Table 3, I introduce additional controls, making the specification

$$\begin{aligned} gExports_{ijt} = {} & \beta_0(gNumIPs_i)_{t-1} + \beta_1(gNumIPs_j)_{t-1} + \beta_2 ln(NumIPs_i)_{04} + \\ & \beta_3 ln(NumIPs_j)_{04} + \beta_5 ln(Export_{ij})_{04} + \beta_6(gGDP_j)_t + \\ & \beta_7 log(Distance_{ij}) + \beta_8 X_{ijt} + FE_t + \epsilon_{ijt} \end{aligned} \tag{2.3}$$

For Model (4), I introduce controls for economic factors, including fluctuations in each country's USD exchange rate, initial GDP, and initial population. For Model (5) I additionally introduce a lag of the dependent variable to control for autocorrelation.

Table 4 reports results from re-estimating the models from Table 3, but substituting ARL for the IP address count. ARL is non-significant at the 5% level in all models, although it should be noted that the signs on the growth of ARL in the origin country are exactly opposite the signs on their counterpart measure from Table 3. This is consistent with the hypothesis that larger numbers of IP addresses are representative of easier access to the Internet (and a corresponding easing of information frictions), while longer routes represent more difficult access.

In both sets of regressions, which closely follow the models used by Freund and Weinhold, much of the models' explanatory power appears to come from

[12] A step taken by Freund and Weinhold, which in my data deletes about 1.3% of the sample.

Table 4. Freund and Weinhold Replication Using Aggregate Route Length

	Dependent variable:				
	Growth of exports from country 1 to country 2				
	(1)	(2)	(3)	(4)	(5)
Orig. ARL Growth		−0.090	−0.064	−0.049	−0.078*
		(0.059)	(0.043)	(0.043)	(0.041)
Dest. ARL Growth		0.032	0.038	0.051	0.051
		(0.053)	(0.039)	(0.039)	(0.037)
Log 2004 Orig. ARL		−0.018	−0.015	−0.022	−0.006
		(0.031)	(0.024)	(0.025)	(0.024)
Log 2004 Dest. ARL		0.011	−0.001	−0.010	−0.004
		(0.023)	(0.019)	(0.019)	(0.018)
Log 2004 Exports	−0.011***	−0.006***	−0.004***	−0.012***	−0.024***
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Dest. GDP Growth	0.300***	0.287***	0.316***	0.261***	0.261***
	(0.064)	(0.065)	(0.052)	(0.056)	(0.054)
Log Distance	0.002	0.009	0.007	−0.005	−0.017***
	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)
Lag Orig. Real Exch. Rate Growth				−0.106*	−0.153***
				(0.059)	(0.057)
Lag Dest. Real Exch. Rate Growth				−0.053	−0.143***
				(0.051)	(0.050)
Log 2004 Orig. GDP				0.004	0.011***
				(0.004)	(0.004)
Log 2004 Dest. GDP				0.004	0.008**
				(0.004)	(0.003)
Log 2004 Orig. Pop.				0.012***	0.017***
				(0.003)	(0.003)
Log 2004 Dest. Pop				0.010***	0.016***
				(0.003)	(0.003)
Lag Export Growth					−0.332***
					(0.009)
Fixed Effects	t	t	t	t	t
Observations	48,125	42,657	42,092	41,606	41,651
R^2	0.013	0.015	0.022	0.024	0.177
Adjusted R^2	0.013	0.014	0.022	0.024	0.176

*p<0.1; **p<0.05; ***p<0.01

variables which control for initial conditions (e.g. Log(EXPORT$_{12}$)$_{95}$) and the lag of the dependent variable introduced in model (5) to account for the time-series nature of the data. When the measures of Internet penetration are significant, it is mainly the variables which account for initial conditions (e.g. Log(NumIPs$_2$)$_{95}$ in Table 3).

2.4.1.1 Comparison to Freund and Weinhold.

It is easiest to compare Table 3 to the results of Freund and Weinhold, as the count of IP addresses is similar to their measure of Internet usage, which was a count of registered webhosts. I find that the estimated coefficients in my models (3)-(5) are comparable in size to their counterparts in Freund and Weinhold, but far less significant. In fact, in model (5), only one variable (Log(NumIPs$_j$)$_{04}$) derived from the IP address count is at all significant, and it is a variable which controls for initial conditions—not year-to-year growth.

Comparing Table 4 to Freund and Weinhold is considerably harder, as they do not use any variables analogous to the ARL. I can draw no direct comparisons between the coefficients on ARL variables, other than to point out that I find them to be far less significant in these models than Freund and Weinhold's measures of Internet access.

2.4.2 Freund and Weinhold Adaptation.

In Table 5, I now modify the original Freund model to use origin-destination fixed effects as a substitute for the variables controlling for initial conditions. Adapted models (1)-(3) in this table are of the form

$$gExports_{ijt} = \beta_0(gNumIPs_i)_{t-1} + \beta_1(gNumIPs_j)_{t-1} + \beta X_{ijt} + FE_{ij} + FE_t + \epsilon_{ijt}. \tag{2.4}$$

Table 5. Freund and Weinhold Adaptation Using IP Address Count

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
Lag Orig. IP Growth	0.087***	0.086***	0.051		
	(0.029)	(0.029)	(0.032)		
Lag Dest. IP Growth	0.013	0.009	−0.004		
	(0.030)	(0.029)	(0.030)		
Lag Joint IP Growth				0.047	0.046
				(0.082)	(0.084)
Dest. GDP Growth		0.155	0.184		
		(0.122)	(0.142)		
Lag Orig. Real Exch. Rate Growth		−0.083	−0.136		
		(0.154)	(0.158)		
Lag Dest. Real Exch. Rate Growth		−0.063	−0.035		
		(0.120)	(0.135)		
Lag Export Growth			−0.400***		−0.407***
			(0.011)		(0.011)
Fixed Effects	ij, t	ij, t	ij, t	it, jt, ij	it, jt, ij
Observations	44,238	44,238	44,238	44,238	44,238
R^2	0.091	0.091	0.238	0.143	0.286
Adjusted R^2	−0.039	−0.039	0.128	−0.014	0.155

*p<0.1; **p<0.05; ***p<0.01

Here, FE_{ij} is an origin-destination fixed effect, while all other variables are defined as they were previously.

Model (1) is a baseline model, including no controls. Model (2) introduces controls for destination GDP and the real USD exchange rates in the origin and destination countries. Model (3) introduces a lag of the dependent variable.

As can be seen here, it is only the growth in IP addresses within the origin country which are significant[13]—and that significance is lost with the introduction of the lagged dependent variable, suggesting that the count of IP addresses largely captures some underlying economic trend.

Models (4) and (5), instead of using separate variables for the growth of the IP address count in origin and destination countries, use the growth of the total

[13] This is similar to the result found by Freund and Weinhold.

number of IP addresses in both countries combined:

$$gExports_{ijt} = \beta_0(gNumIPsTotal_{ij})_{t-1} + \beta X_{ijt} + FE_{ij} + FE_{it} + FE_{jt} + \epsilon_{ijt}.$$

Here, $(gNumIPsTotal_{ij})_{t-1}$ is the lagged growth of total IP addresses in i and j combined. FE_{it}, FE_{jt}, and FE_{ij} are origin-year, destination-year, and origin-destination fixed effects, respectively. It is only possible to use the origin-year and destination-year fixed effects in this specification because there is bilateral variation in $(gNumIPsTotal_{ij})_{t-1}$; however, this metric makes no distinction between growth in the origin vs. destination countries. Model (4) includes no controls; model (5) introduces a lag of the dependent variable similar to model (3).

Even with the use of additional control variables, there is no gain in significance for the joint measure of IP address growth. This is likely because, as seen in models (1) and (2), it is only the count of IP addresses in the origin province which matter.

Table 6 repeats the models from Table 5, substituting the ARL measure for the count of IP addresses. Here, the measure of Internet access only becomes significant after introducing the lagged dependent variable—suggesting that part of the noise in the measure is based upon underlying trends—but again, it is only the route length in the origin province which is at all significant.

Finally, Table 7 includes both measures of Internet access simultaneously. These models demonstrate that the results from using each measure independently do not suffer from including both together, and indeed, there is a small gain of significance for the origin-country count of IP addresses. This suggests that the two measures capture largely different aspects of Internet access, although in this context, it appears that the ARL remains the more useful measure.

Table 6. Freund and Weinhold Adaptation Using Aggregate Route Length

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
Orig. ARL Growth	−0.113	−0.113	−0.126**		
	(0.073)	(0.073)	(0.063)		
Dest. ARL Growth	0.042	0.039	0.058		
	(0.068)	(0.067)	(0.059)		
Lag Joint ARL Growth				1.044	0.598
				(0.980)	(0.712)
Dest. GDP Growth		0.152	0.175		
		(0.122)	(0.141)		
Lag Orig. Real Exch. Rate Growth		−0.104	−0.148		
		(0.152)	(0.156)		
Lag Dest. Real Exch. Rate Growth		−0.069	−0.041		
		(0.119)	(0.134)		
Lag Export Growth			−0.400***		−0.407***
			(0.011)		(0.011)
Fixed Effects	ij, t	ij, t	ij, t	it, jt, ij	it, jt, ij
Observations	44,238	44,238	44,238	44,238	44,238
R^2	0.091	0.091	0.238	0.143	0.286
Adjusted R^2	−0.040	−0.039	0.128	−0.014	0.155

*p<0.1; **p<0.05; ***p<0.01

Table 7. Freund and Weinhold Adaptation Using Both Measures

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
Lag Orig. IP Growth	0.089***	0.088***	0.053*		
	(0.027)	(0.028)	(0.030)		
Lag Dest. IP Growth	0.014	0.010	−0.003		
	(0.029)	(0.028)	(0.029)		
Orig. ARL Growth	−0.120*	−0.120*	−0.130**		
	(0.067)	(0.067)	(0.062)		
Dest. ARL Growth	0.044	0.040	0.057		
	(0.069)	(0.068)	(0.059)		
Lag Joint IP Growth				0.047	0.046
				(0.082)	(0.084)
Lag Joint ARL Growth				1.042	0.596
				(0.977)	(0.709)
Dest. GDP Growth		0.150	0.176		
		(0.120)	(0.140)		
Lag Orig. Real Exch. Rate Growth		−0.083	−0.136		
		(0.156)	(0.161)		
Lag Dest. Real Exch. Rate Growth		−0.067	−0.040		
		(0.119)	(0.134)		
Lag Export Growth			−0.400***		−0.407***
			(0.011)		(0.011)
Fixed Effects	ij, t	ij, t	ij, t	it, jt, ij	it, jt, ij
Observations	44,238	44,238	44,238	44,238	44,238
R^2	0.091	0.091	0.238	0.143	0.286
Adjusted R^2	−0.039	−0.039	0.128	−0.014	0.155

*p<0.1; **p<0.05; ***p<0.01

In all of these models, the coefficients on ARL growth, where significant, hold the opposite sign compared to the coefficients on the count of IP addresses, which is again consistent with the hypothesis that longer routes indicate more difficult or costly Internet access.

Also of note is that, while the lag of the dependent variable still accounts for a large fraction of the explained variation when it is introduced in each table's model 3, these models can explain much more of the variation without relying on the persistent trends.

From Tables 5 and 6, I conclude that, while Internet access does have an impact on trade, it does so largely through a channel associated with the origin country. (This is also what Freund and Weinhold found.) A possible explanation for the differential impact on origin and destination countries is that exporters (origin countries) use the Internet to publicize information about products available for export, while importers (destination countries) use the Internet to view this information. Reliable and cheap Internet access is therefore more beneficial to exporters, who must constantly maintain a website or other Internet presence, while importers only require occasional Internet access when searching for product information—and are thus less impacted by unreliable or expensive Internet access.

2.4.2.1 Comparison to Freund and Weinhold. Again, the coefficients on growth in IP address count can be directly compared to their counterparts in the original Freund and Weinhold paper. I find that growth in origin-country IP addresses has a noticeably larger effect on growth in exports than Freund and Weinhold's measure—in Model (3) of Table 7 (where the coefficient is marginally significant after the introduction of controls), the coefficient is roughly twice as large as its Freund and Weinhold counterpart.

However, what I find more interesting is that changes in origin ARL have an effect of similar magnitude to changes in destination GDP (which is used to control for the size of the importing market): a 1% decrease in ARL is estimated to cause roughly 2/3 the increase in exports that a 1% increase in importer GDP would. This is a considerably larger effect than any which Freund and Weinhold found, which may be due to the fact that typical values of ARL lie within a relatively small band: a small percentage change in ARL can therefore have a large impact.

2.4.3 Allen Replication.

Allen (2014) analyzes several unusual patterns in trade of agricultural products among provinces of the Philippines. I adapt his methodology (and much of his original data, provided as part of his replication files) using my measures of Internet access.

2.4.3.1 Simultaneous Import and Export.

The first of these patterns is that many Philippine provinces simultaneously import and export the same product. Allen demonstrated that this market failure can be partly explained by information frictions; specifically, he found that provinces which contained cell phone towers were less likely to simultaneously import and export.

In Table 8, I perform the same analysis, using Internet access as the proxy for information frictions instead of cell phone access. In this table, all models are of the form

$$ImpExp_{itc} = \beta NetworkAccess_{it} + FE_i + FE_c + \epsilon_{itc}. \qquad (2.5)$$

Here, i represents province or port, t represents year, and c represents agricultural commodity. $ImpExp_{itc}$ is an indicator variable which takes the value 1 if location i both imported and exported commodity c in year t. $NetworkAccess_{it}$ is an indicator variable which takes the value 1 if province i had at least one IP address in year t. FE_i and FE_c are location and commodity fixed effects.

Table 8. Allen Replication: Simultaneous Import/Export Using IP Address Count

	Dependent variable:			
	Simultaneously imported and exported			
	Prov.-prov., annual		Port-port, 4th quarter	
	(1)	(2)	(3)	(4)
Has IP Addresses	-0.036** (0.018)	-0.064* (0.032)	-0.020*** (0.007)	-0.061** (0.025)
Sample Provinces/Ports	All	Trading	All	Trading
Fixed Effects	i, c	i, c	i, c	i, c
Mean of dep. variable	0.263	0.406	0.059	0.201
R-squared	0.497	0.445	0.411	0.440
Observations	5181	3361	14407	4224

*p<0.1; **p<0.05; ***p<0.01

Models (1) and (2) of Table 8 are estimated at the province level. Model (1) includes all provinces, while Model (2) excludes provinces which neither imported nor exported commodity c in year t. Models (3) and (4) repeat this exercise at the port level.

The data used to estimate these models, as well as those of Tables 9 to 13, represent the period from 2004 to 2009, which is the period in which my routing data overlaps with the data provided in Allen's replication files.

As can be seen from Table 8, Internet access makes it substantially less likely that a province will experience this type of market failure.

In Table 9, I incorporate the ARL measure into this analysis. Here, I restrict the sample to only those location-years for which $NetworkAccess_{it} = 1$, and estimate the additional impact which route length has upon this market failure. In this table, all models are of the form

$$ImpExp_{itc} = \beta ARL_{it} + FE_i + FE_c + \epsilon_{itc}. \tag{2.6}$$

Table 9. Results: Simultaneous Import/Export in Internet-Connected Provinces

	Dependent variable:			
	Simultaneously imported and exported			
	Prov.-prov., annual		Port-port, 4th quarter	
	(1)	(2)	(3)	(4)
ARL	0.041**	0.047**	0.019***	0.043**
	(0.017)	(0.024)	(0.006)	(0.019)
Sample Provinces/Ports	All	Trading	All	Trading
Fixed Effects	i, c	i, c	i, c	i, c
Mean of dep. variable	0.336	0.514	0.064	0.199
R-squared	0.516	0.432	0.409	0.449
Observations	2622	1715	8905	2865

*p<0.1; **p<0.05; ***p<0.01

Here, $ARLit$ is the ARL from the PAIX exchange in San Francisco[14] to province i (or the province containing port i). All other variables are defined as in Table 8, and the models follow the same order as before.

I find that longer routes make a location substantially more likely to simultaneously import and export a commodity. In fact, in some models this effect is large enough to completely offset the benefit of gaining Internet access in the first place. It is counterintuitive to think that poor Internet access (as defined by having longer routes) is worse than no Internet access at all, and so I suspect that part of this finding is driven by limitations to my geolocation data, in particular the fact that it identifies many provinces as lacking any IP addresses at all.

2.4.3.2 Price Pass-Through. Allen next investigates the effect which information frictions have upon price pass-through. Again using cell tower access as a proxy for information frictions, Allen finds that price pass-through is

[14] Chosen because it is the closest collector to the Philippines.

substantially more complete in origin-destination province pairs which have a cell phone connection (i.e. which both contain a cell tower).

Table 10. Results: Internet Access and Price Pass-through

	Dependent variable:			
	Change in log destination price ratio			
	(1)	(2)	(3)	(4)
	OLS	2SLS	OLS	2SLS
Change in Log Orig. PR	0.828***	0.752***	0.831***	0.762***
	(0.053)	(0.117)	(0.053)	(0.113)
Change in Log Orig. PR × Has IP Addresses			-0.125	-0.110
			(0.188)	(0.188)
Fixed Effects	t	t	t	t
R-squared	0.645	0.641	0.645	0.643
Observations	229	229	229	229

*p<0.1; **p<0.05; ***p<0.01

As before, I first replicate Allen's models, substituting my measure of Internet access for the cell tower data. Results are shown in Table 10: models (1) and (2) are of the form

$$dLogDestPR_{ijt} = \beta dLogOrigPR_{ijt} + FE_t + \epsilon_{ijt} \tag{2.7}$$

and models (3) and (4) are of the form

$$\begin{aligned} dLogDestPR_{ijt} = {} & \beta_0 dLogOrigPR_{ijt} + \beta_1 dLogOrigPR_{ijt} \times Connection_{ijt} + \\ & FE_t + \epsilon_{ijt} \end{aligned} \tag{2.8}$$

In both forms of the model, i represents origin province, j represents destination province, and t represents year. $dLogDestPR_{ijt}$ is the change in the log price ratio of corn to rice in the destination province; $dLogOrigPR_{ijt}$ is the same quantity measured in the origin province. $Connection_{ijt}$ is an indicator variable which takes the value 1 if both provinces i and j each have at least one IP address in year t.

Models (1) and (3) are estimated using OLS. Models (2) and (4) are estimated using 2SLS: as in the original Allen paper, the change in origin price ratio is instrumented with a vector of changes in origin-province rainfall. These weather variables are likely to affect prices in the origin province itself—via their impact on crop yields—but are plausibly uncorrelated with the price ratio in other (destination) provinces.

Table 11. Results: Internet Access and Price Pass-through Using Aggregate Route Length

	Dependent variable:	
	Change in log destination price ratio	
	(1)	(2)
	2SLS	2SLS
Change in Log Orig. PR	0.764***	0.682***
	(0.113)	(0.113)
Change in Log Orig. PR × Has IP Addresses	-0.726	
	(1.996)	
Change in Log Orig. PR × ARL	0.105	
	(0.338)	
Change in Log Orig. PR × Has IP Addresses		1.483**
		(0.631)
Change in Log Orig. PR × ARL		-0.271
		(0.190)
Change in Log Orig. PR × Has IP Addresses		1.031
		(1.562)
Change in Log Orig. PR × ARL		-0.626
		(0.569)
Fixed Effects	t	t
R-squared	0.643	0.665
Observations	229	229

*p<0.1; **p<0.05; ***p<0.01

In Table 11, I next incorporate ARL into this analysis. Here, model (1) is of the form

$$\begin{aligned} dLogDestPR_{ijt} = {} & \beta_0 dLogOrigPR_{ijt} + \beta_1 dLogOrigPR_{ijt} \times Connection_{ijt} + \\ & \beta_2 dLogOrigPR_{ijt} \times ARL_{ijt} + FE_t + \epsilon_{ijt} \end{aligned} \tag{2.9}$$

and model (2), which draws upon results from my earlier replication of Freund and Weinhold, is of the form

$$\begin{aligned} dLogDestPR_{ijt} = {} & \beta_0 dLogOrigPR_{ijt} + \beta_1 dLogOrigPR_{ijt} \times Connection_{it} + \\ & \beta_2 dLogOrigPR_{ijt} \times ARL_{it} + \beta_3 dLogOrigPR_{ijt} \times Connection_{jt} + \\ & \beta_4 dLogOrigPR_{ijt} \times ARL_{jt} + FE_t + \epsilon_{ijt} \end{aligned} \tag{2.10}$$

As in my results from adapting Freund and Weinhold, I find that my measures of Internet access are most significant when split into separate measures of the origin and destination provinces, and that when this is done, only Internet access in the origin province is significant. ARL remains non-significant in both provinces, although the signs of both coefficients are as predicted. ARL does become more significant in model (2)—and again, the measure in the origin province is more significant than that in the destination province.

Because even the non-significant coefficients in these models have the expected signs, I suspect that my measures of Internet access are noisy. Also, since this was not an issue with my adaptation of Freund and Weinhold, which used country-level data, I would conclude that this noise is more prevalent on the province level. I again suspect that this may be due to inaccuracies in my province-level geolocation data.

Complete price pass-through, in which shocks to the price of a commodity in the origin province are fully felt in destination provinces, would result in the total coefficient on *LogQuantity* and its appropriate interactions being equal to 1. I use a one-sided test here, because in some cases the total coefficient is so much greater than 1 that a two-sided test rejects the null hypothesis due to passthrough being "more than complete."

Table 12. Results: Tests of Complete Passthrough

	p-values		
H_0 : Complete pass-through between provinces...	2004	2008	Overall
...with no IP addresses	0.002***	0.002***	0.002***
...with 95th percentile ARL	0.028**	0.039**	0.043**
...with 75th percentile ARL	0.025**	0.025**	0.028**
...with 50th percentile ARL	0.027**	0.026**	0.026**
...with 25th percentile ARL	0.027**	0.026**	0.050**
...with 5th percentile ARL	0.300	0.179	0.428

*p<0.1; **p<0.05; ***p<0.01
All tests performed using model (2) of Table 11.
All tests are one-sided.

As can be seen from the table, it is possible to reject the hypothesis of complete (or more than complete) pass-through at the 5% level for provinces which contain no IP addresses, as well as those which have ARL in the 95th, 75th, 50th, and 25th percentiles.[15] In the case of provinces with ARL in the 5th percentile,[16] it is not possible to reject this hypothesis.

2.4.3.3 Farmer Trade Search. The final part of Allen's analysis that I replicate here is the analysis of farmer trading behavior. Allen found that larger farmers were more likely to incur freight costs (i.e. "trade"), but that access

[15] It is important to remember that longer routes suggest worse Internet access; provinces with ARL above the 95th percentile are therefore the 5% of provinces with the worst Internet connection by this measure.

[16] i.e. the 5% of provinces which have the best Internet access by this measure.

Table 13. Results: Internet Access and Farmer Search Patterns

	Dependent variable:			
	Farmer seached for trade			
	(1)	(2)	(3)	(4)
Log Quantity	0.017***	0.024***	0.028***	0.114***
	(0.001)	(0.003)	(0.002)	(0.044)
Has IP Addresses		0.112***		
		(0.025)		
Log Quantity × Has IP Addresses		−0.026***	−0.042***	
		(0.004)	(0.003)	
Log Quantity × 5th-24th Percentile ARL				−0.081*
				(0.046)
Log Quantity × 25th-49th Percentile ARL				−0.128***
				(0.044)
Log Quantity × 50th-74th Percentile ARL				−0.109**
				(0.044)
Log Quantity × 75th-95th Percentile ARL				−0.146***
				(0.044)
Log Quantity × 95th+ Percentile ARL				−0.129***
				(0.045)
Fixed Effects	pymc	pyc, mc	pymc	pyc, mc
Sample Provinces	All	All	All	Connected
Dep. Var. Mean	0.139	0.139	0.139	0.065
Observations	365,297	365,297	365,297	84,809
R^2	0.672	0.635	0.674	0.555
Adjusted R^2	0.655	0.628	0.656	0.529

*p<0.1; **p<0.05; ***p<0.01

to mobile phones closed the gap between small and large farmers. I adapt Allen's methodology and display the results in Table 13.

Model (1) is a baseline model, not incorporating any measurements of Internet access, of the form

$$FarmerTraded_{iymc} = \beta_0 logQuantity_{iymc} + FE_{pymc} + \epsilon_{iymc}. \quad (2.11)$$

Here, i describes farmers, y and m describe year and month, c describes agricultural commodities, and p describes the province in which farmer i operates. $FarmerTraded_{iymc}$ is an indicator variable which takes the value 1 if farmer i incurred freight costs for commodity c in year y and month m. $logQuantity_{iymc}$ is the log of the quantity of commodity c that farmer i produced in year y and month m. FE_{pymc} is a province-commodity-time fixed effect.

Model (2) is of the form

$$FarmerTraded_{iymc} = \beta_0 logQuantity_{iymc} + \beta_1 InternetAccess_{pym} + \\ \beta_2 logQuantity_{iymc} \times InternetAccess_{pym} + FE_{pyc} + FE_{mc} + \epsilon_{iymc} \tag{2.12}$$

in which $InternetAccess_{pym}$ is an indicator variable which takes the value 1 if province p contains at least one IP address. FE_{pyc} and FE_{mc} are province-commodity-year and commodity-month fixed effects, respectively.

Model (3) is of the form

$$FarmerTraded_{iymc} = \beta_0 logQuantity_{iymc} + \\ \beta_2 logQuantity_{iymc} \times InternetAccess_{pym} + FE_{pymc} + \epsilon_{iymc}. \tag{2.13}$$

In Model (4), I restrict the sample to farmers in provinces which contain at least one IP address, and examine the effect of ARL. Rather than attempt to interact the two continuous variables for log-quantity and ARL, I instead generate indicator variables which take the value 1 if ARL is within a specified percentile range, and interact these with the log-quantity:

$$\begin{aligned} FarmerTraded_{iymc} = {} & \beta_0 logQuantity_{iymc} + \\ & \beta_1 logQuantity_{iymc} \times Pct05\_24_{pym} + \\ & \beta_2 logQuantity_{iymc} \times Pct25\_49_{pym} + \\ & \beta_3 logQuantity_{iymc} \times Pct50\_74_{pym} + \\ & \beta_4 logQuantity_{iymc} \times Pct75\_94_{pym} + \\ & \beta_5 logQuantity_{iymc} \times Pct95Plus_{pym} + \\ & FE_{pymc} + \epsilon_{iymc} \end{aligned} \tag{2.14}$$

In the first three models, I find that Internet access has a stronger effect than Allen found for cell phone access. Where Allen's results suggest that smaller farmers are less likely to trade, even with access to cell phones, I find that Internet access completely removes this difference (as in Model (2)), or even reverses it, so that it is in fact smaller farmers who are more likely to trade (as in model (3)).

When I incorporate ARL into this analysis in model (4), it appears to be the provinces where ARL is above the 25th percentile which drive this result: below the 25th percentile, larger farmers are more likely to trade; above the 25th percentile, larger farmers appear no more likely to trade, or possibly even less likely (as in the 75th-94th percentiles).

It is difficult to explain why small farmers export more than large farmers when given poor Internet access, but not when given good Internet access, or no Internet access at all. A possible explanation might be that, when Internet access is of poor quality, it still suffices to ease the information frictions experienced by small farmers. This would allow them to compete in the export market—but as the quality of Internet access improves (offering lower latency, for example), it offers some competitive advantage which only large farmers are able to exploit: this might result from some economy of scale, or it might be that it requires a greater degree of literacy or human capital associated with larger, more prosperous farmers.

2.4.3.4 Comparison to Allen. My proposed measures of Internet access perform comparably to the cell tower data used as measures of information friction in Allen (2014). The IP address count functions well as a direct replacement for the cell tower count, and the use of ARL offers an additional dimension by which to measure Internet access, which allows me to explain additional variation among Internet-connected locations.

However, based on a lack of significance in some models—primarily the models of price pass-through—I remain concerned about the precision of my method of geolocating IP addresses at the province level. There exists commercial data which purports to offer greater accuracy, and it is possible that with this additional data, I may be able to remove some of the noise from my measures.

2.5 Conclusions

Based on the empirical results from adapting previous papers, I conclude that my proposed measures possess similar or greater explanatory power when compared to previously-used measures of Internet access. Additionally, these measures may be computed over large geographic areas, at a finer level of detail, using an automated script, making the measures far easier to compute and use in a variety of models.

This is not to say that these measures are a perfect measure of Internet access: they are intended to serve as proxies when more reliable data is not available (a state of affairs which is unfortunately common). In this role, the measures already appear to serve well.

There remains some room for improvement, naturally: it is quite likely that the computed measures contain noise due to lack of precision in the geolocation data used for aggregation, which may be fixable with the use of commercial data.

CHAPTER III

THE COSTS OF INTERNATIONAL INTERNET COMMUNICATION: MEASUREMENT AND IMPLICATIONS FOR TRADE

3.1 Introduction

Information frictions are a significant component of barriers to international trade. These barriers include the costs of locating buyers or suppliers, arranging the transportation and delivery of goods, and monitoring foreign market conditions, among others. An important driver of information frictions are communication costs, which are reduced by advances in communications technology allowing traders to exchange information more easily. Advances in communication technology began in the 19th century with the invention of the telegraph, and especially with the installation of trans-Atlantic telegraph cables. Per Steinwender (2018), this advance reduced the trans-Atlantic information lag from 10 to 1.3 days, by making it possible to communicate information without sending a physical message aboard ship. Further advances, such as the telephone and fax machine, further reduced this and other information lags, while the decreasing cost of these technologies allowed smaller and smaller traders to acquire them. The Internet is now the primary medium of communication in most of the world, and it is increasingly important that economists be able to measure the costs of Internet communication effectively.

In this chapter, I develop a methodology to estimate bilateral Internet communication costs on a country-to-country network, by adapting the model of network transportation costs of Allen and Arkolakis (2019). The structure of the Internet is a network sufficiently similar to the environment of this model that it is easily adaptable, and by coupling this model with novel data sources describing the volume of Internet communication sent through a large communication hub, and the routing of that communication as described in II, I am able to extract measures of communication costs analogous to the iceberg trade costs already in wide use in the trade literature. I further demonstrate that these extracted communication costs possess explanatory power when applied in gravity models of trade, and produce results similar to those previously seen in Keller and Yeaple (2013). The data used in this methodology are publicly accessible without requiring the significant data-gathering efforts necessary for previous approaches to modeling information frictions, such as Allen (2014), and the methodology requires only modest computational resources to process the data and estimate the model.

Despite its obvious importance for understanding economic transactions, particularly on the international level, the cost of communication is difficult to measure. In particular, it is difficult to measure bilateral costs of communication, especially Internet communication. We are currently seeing the Internet expand explosively into new markets: cellular Internet, which provides "last-mile" connections via cellular phone towers and their accompanying infrastructure, allows for Internet access to households and businesses without a landline phone connection, or even access to an electrical grid: a cellular phone can access the Internet from anywhere with a cell tower nearby, and can be charged from a gasoline generator or solar panel. In the near future, satellite Internet constellations such as the Starlink project promise to remove even the need for cellular towers. However, unlike telegraph or telephone communication, for which bilateral rates were readily available (such as used in Fink, Mattoo, and Neagu (2005)), Internet communication is commonly billed on a per-month or per-gigabyte basis that obscures the bilateral costs that would be most useful in the trade literature.

Further complicating the matter, the Internet lacks a central authority that compiles cost, traffic, or infrastructure data, meaning that no single dataset covering the entire global Internet exists. My methodology overcomes this restriction by using data gathered at a single point, the Chicago Equinix Internet Exchange Point, to produce estimates of bilateral Internet communication costs on a global country-to-country network. The two datasets from Chicago Equinix that I use provide a viewpoint on how much communication passes through that facility, and how it is routed. From this single, incomplete viewpoint, I am able to impute the amount of traffic generated by this communication on the links of a country-to-country network. While Chicago Equinix is the only readily-available source of the paired datasets necessary to perform this imputation, this single source is sufficient to apply the resultant traffic distribution to the model adapted from Allen and Arkolakis (2019), and additional sources of data could only improve the accuracy of estimation.

This methodology also avoids several flaws present in earlier approaches to information frictions in the literature: the data sources I use are based on fundamental aspects of the Internet, and are unlikely to become obsolete in the near future, as has already occurred to the measures of Internet access in Freund and Weinhold (2004). The data sources, being publicly available with minimal fixed costs of access, do not require significant data-gathering efforts, as were reported in Allen (2014), and are unlikely to become inaccessible, as also appears to have happened to the data sources of Allen (2014). Finally, the model of Allen and Arkolakis (2019) is scalable, and can be applied to scopes ranging from the province-level to the global Internet; this allows it to be used in the context of international trade, but it can also be scaled down to contexts as detailed as those found in Leuven, Akerman, and Mogstad (2018), using micro-survey data.

3.2 Literature Review

My work in this chapter connects three threads of the economic literature: one examining the effects that communications technology has on trade, and a second which examines the costs of trade on a network of ports or countries. However, rather than apply this transport network literature directly to trade flows, I adapt this literature to apply it to the costs of communication on a network. This work is also informed by some elements of the computer science literature.

3.2.1 Communication Costs and Trade. Previous work has addressed the effects that expanding access to Internet and other communications technologies have had on trade, but largely take the approach of measuring communication costs via some easily-attainable proxy. Work in this vein includes Freund and Weinhold (2004), Allen (2014), and Leuven et al. (2018), which use counts of registered webhosts, cell tower access, and broadband Internet access, respectively, to proxy for communications costs. A further branch of this literature represented by Fink et al. (2005), Lew and Cater (2006), Ejrnæs and Persson (2010), and Steinwender (2018) focuses on historic contexts and the costs of communication by what are now obsolete technologies (the telegraph and, arguably, the landline telephone) for which costs were more readily measurable. I depart from both these branches of the literature by developing a model which can estimate otherwise unobservable or difficult-to-observe communications costs, as an "iceberg communication cost" analogous to the iceberg trade costs already widely known in the trade literature. Further, the data used in this model are generated by processes which are fundamental to the functioning of the Internet, and so are unlikely to become outdated (as has arguably happened to Freund and Weinhold (2004)'s count of webhosts), relatively straightforward to obtain (unlike the cell tower data painstakingly gathered in Allen (2014)), and more universally applicable (unlike the Norwegian micro-survey data used in Ejrnæs and Persson (2010)).

A separate section of this literature deals with substitution patterns related to communications costs, as exemplified Keller and Yeaple (2013), Cristea, Anca D. (2015) and Gokan, Kichko, and Thisse (2019). These papers concern the choices of multinational firms when undertaking production overseas from their headquarters; a pattern that emerges is that expensive communication leads multinationals to engage in "embodied knowledge transfer," in the terminology of Keller and Yeaple, or the use of local knowledge to produce complex goods which are then physically transported (as an alternative to communicating that knowledge directly).

Diverging slightly from the topic of trade, Blonigen, Cristea, and Lee (2020) finds that information frictions, specifically monitoring costs, resulting from physical and cultural distance have significant negative impacts on cross-border merger and acquisition (M&A) activity. The effect is less pronounced in the manufacturing sector, owing to the lesser importance of monitoring activity, which is an important factor in the disproportionate emphasis on manufacturing in such M&A. Costs of communication are a major factor in these monitoring costs, as modern communications technology can potentially reduce the importance of physical distance when available. Such costs, however, are difficult to measure directly.

3.2.2 Trade Costs on Networks. Another thread in the literature addresses the estimation of trade costs on networks. This is relevant to my work, not because I will be estimating a network trade cost directly, but because the

Internet is ultimately a similar kind of network. Specifically, the Internet is a communication network structured with strong similarities to the global trade networks seen in these papers, and the costs of communication on this network can be estimated using similar methodologies to those developed to estimate the costs of trade on a global trade network.

Kikuchi (2002) provides a theoretical model predicting that countries with communications networks that are interconnected (or, by extension, interconnected to a greater degree) will have a comparative advantage in the trade of business services.

Anderson and van Wincoop (2004) lay out a basic framework for the estimation of trade costs using a gravity model, or from purchasing power parity data. They also present a summary of available data on trade costs, derived from records, surveys of national non-tariff barriers, and other sources. However, this primarily addresses "tangible" costs and barriers to trade, leaving out intangibles such as communication costs and information frictions.

An entire sub-thread of this literature deals with transportation over a defined network, with an emphasis on enabling detailed counterfactuals: notable studies in this vein include Donaldson and Hornbeck (2016), S. Redding (2016), Nagy (2016), Sotelo (2015), and Ganapati, Wong, and Ziv (2020). Most relevant is Allen and Arkolakis (2014), which establishes a very general framework for modeling economic activity on surfaces with highly-adaptable topology; applying this framework, Allen and Arkolakis (2019) provides a more specific framework for estimating the costs of each link in a transportation network, which is applied to the context of inter-city trade along the US highway network. The structure of the highway network is similar to that of the Internet, and the structure of this model is convenient to adapt to cases where complete traffic data (i.e. data describing the entire universe of traffic throughout a network) is hard to come by.

3.2.3 Endogenous Trade Costs.

I also take some inspiration from the literature on endogenous trade costs, in which the costs of trading are part of an equilibrium and are determined partly by the distribution of trade flows. As with the literature of trade costs on networks, this chapter will apply models of endogenous trade costs to the related problem of endogenous communication costs, and demonstrate that with minimal adaptations these models can be used to produce useful results in this novel context. Works central to this literature include Anderson and van Wincoop (2004), Head and Mayer (2014), Hummels (2007), and Limão and Venables (2001).

Additional works model determinants of endogenous trade costs that have analogues in the context of communication costs. Specific examples include search frictions between exporters and bulk carrier ships as modeled by Brancaccio, Kalouptsidi, and Papageorgiou (2020), analogous to similar frictions between local ISPs and the telecom companies that operate the global Internet backbone, and port efficiency as modeled by Clark, Dollar, and Micco (2004) and Blonigen and Wilson (2008), analogous to efficiencies at large Internet Exchange Points. Other papers with contributions in this vein include Kleinert and Spies (2011), Behrens and Picard (2011), Jonkeren, Demirel, van Ommeren, and Rietveld (2011), Brancaccio, Kalouptsidi, Papageorgiou, and Rosaia (2020), Gruber and Marattin (2010), S. J. Redding and Turner (2014), and Hummels, Lugovskyy, and Skiba (2009). This subset of the literature provides additional justification for the use of models from the endogenous trade cost literature, on the basis that communication costs via the Internet are influenced by a range of analogous determinants.

Also of note is Duranton and Storper (2008), which uses a model of industry location with endogenous transaction costs to explain a juxtaposition between rising total trade costs and falling transport costs. This model suggests that due to increased use of complex, specialized machinery, transaction costs in the form of extensive communication between machinery manufacturer and client have offset the reduced cost of actual transport. In addition to contributing to the endogenous trade costs literature, this also motivates interest in communication costs—which are likely much lower now than in 2008, thanks to further advancements in Internet infrastructure and communication technologies.

3.2.4 Relevant Computer Science Literature. One of the key components of my approach is geolocation of Internet end-users as well as networks. Precisely identifying a network's geographic footprint remains a thorny problem, but Rasti, Magharei, Rejaie, and Willinger (2010) provides a novel method of doing so. However, this method creates what is in effect a probabilistic mapping of networks to countries, which vastly complicates the process of approximating global Internet traffic as discussed in section 3.3.5 of this chapter. The increase in computational complexity has proven difficult to surmount without using advanced computing resources (i.e. without a supercomputing cluster or intensive use of cloud computing) and so I have opted to use a simpler and more accessible methodology. Appendix C.3 contains a brief discussion of an alternate methodology based on Rasti et al. (2010), which could be implemented given sufficiently-long time constraints or an abundance of computing power.

3.3 Data

In order to estimate the costs of communication, it is first necessary to somehow measure the amount of communication which takes place–ideally, the amount of communication activity generated by an international trade transaction. This is a complex problem: at the micro level, there is little data measuring how many emails or phone calls a trading firm sends in the process of arranging a trade, and at the macro level, it is not feasible to separate trade-related communication from other communications. Therefore, a novel approach to measuring communication will be necessary. My solution combines two novel datasets that describe Internet routing and Internet communication as observed from the same position within the Internet.

3.3.1 Analogy to Physical Transportation.

It is perhaps easiest to explain what this data captures by first establishing an analogy to the transportation of physical goods. A common question in the trade literature is how to determine the costs of trade among various locations in a transportation network: this may be country-to-country, port-to-port, or even city-to-city, depending on context. Perhaps the ideal dataset for such an application consists of two parts: measurements of trade (where goods start out, where they end up, and how valuable they are), and measurements of shipping (the value of the goods transported along various links in the network, irrespective of origin and destination).

Given these two pieces of information, it is possible to draw conclusions about the costs of the links in the transportation network. To give the simplest possible example, if there are only two routes which connect nodes A and B in the network, and the majority of goods shipped from A to B are sent along the first route, one can reasonably conclude that the first route is the less costly to use. If similar behavior can be observed for many pairs of A and B nodes, then one can begin to draw conclusions about the factors which make routes costly by comparing the characteristics of the less-used routes.

The data in this ideal dataset could be obtained from commerce and transportation authorities, such as UN COMTRADE, the US Department of Commerce (DOC), and/or the US Department of Transportation (DOT). Unfortunately, there is no analogue to these authorities when it comes to data on Internet communication: the many distinct networks comprising the Internet may collect relevant data within their own borders, but there is no central authority which aggregates this data or ensures that the entities collecting it use a standardized process. Therefore, in this analogy, suppose that there is no federal DOT or DOC.

Even with this restriction, it is possible to get a partial picture of where road traffic occurs by conducting a survey of drivers in a single location. Suppose that I survey drivers as they leave Eugene, Oregon, asking each one where they are driving to. I can then get directions to their destination from Google Maps, Apple Maps, or a variety of other sources, and record the sections of road which these directions say to drive on. Having done this for a large number of drivers, I can then count the number of times that a driver from Eugene will (probably) use each section of road in the US highway network. While this would only produce a measurement of traffic resulting from drivers passing through Eugene, a more complete picture could be obtained by repeating the procedure in a densely-populated, centrally-located, or heavily-traveled area, e.g. Portland, Chicago or New York.

3.3.2 Internet Routing and Routing Data. Briefly, “routing data” as used in this chapter refers to records of the routes that can be used to transmit communication over the Internet, analogous to driving directions on a road network. In its typical format, each observation of this data describes a route that can be used by a sending device (typically the device recording the data) to communicate with a contiguous block of receiving devices. Sending and receiving devices are uniquely identified by Internet Protocol (IP) addresses. The route is described as a sequence of unique identifying numbers for the distinct networks that it passes through. A sample of what this data may look like is provided in Appendix A.1.

The Internet is not monolithic: rather, as described in Chapter II, it is composed of many distinct computer networks that have developed protocols for cooperating and connecting with each other. Each of these networks, or rather their administrators, independently select a set of routes that they prefer to use to send their users' communication to its destination. In highly-connected locations such as Internet Exchange Points (IXPs), which are datacenters where many networks connect to each other, routing data is very detailed, and very complete: it contains listings of routes which can be used to communicate with the vast majority of Internet-connected devices in the world. It can generally be assumed that at an IXP, the network has effectively perfect information about the routes available to them, and has chosen the best possible route (i.e. there exist no routes which an IXP would strictly prefer to use but does not know about).

A network's administrators do not manually select routes, for reasons of scale: rather, administrators design a metric based on multiple criteria by which a computer can select the "best" route to a block of IP addresses. The most widely-used criterion is the directness of the route. In the vast majority of cases, the selected route will be the most direct one: not necessarily the physically shortest route, but rather the route which passes through the fewest intermediary networks. The exceptions occur as a result of idiosyncratic variations which are not observable in this data: to give a simple hypothetical, if the administrator of network A has an old college friend at network B, they may be able to get favorable terms making routes passing through network B less costly even if they are not the most direct.

It is common that the selected routes observed in routing data contain a multiplicity of routes that can reach the same block of IP addresses: this is partly a precaution against service disruptions, e.g. a route being cut off due to a backhoe hitting a buried cable.

I once again use routing data compiled by the Oregon Route Views Project (ORVP), previously discussed in Chapter II. In my empirical exercise, I this time focus on routing data from one particular collector, that being the Equinix Chicago IXP, due to a conveniently-available set of complementary communication data, discussed in the next section. Equinix Chicago is the only IXP contributing to the ORVP for which such matching data is readily available.

3.3.3 Internet Communication and Trace Data. For information on communication volumes, I turn to the Center for Applied Internet Data Analysis (CAIDA)'s Anonymized[1] Internet Traces Dataset. A brief description of the raw data is presented in Appendix A.2.

To make the distinction between routing and communication data clear, while the routing data describes the paths communication can take to reach a destination, it does not describe how often each of those paths is used. Conversely,

[1] The anonymization referred to here obscures the *exact* identities of senders and recipients of data, but in a way that still allows it to be geolocated. For further information, see Appendix A.2.1.

this communication data describes how much communication is sent to and from a myriad of devices, but does not describe the path used to get it there.

Since 2008, CAIDA has taken periodic “snapshots” of the traffic flowing through select devices on the high-speed Internet backbone. This data consists of observations of individual “packets” of information transmitted over the Internet, including the origin IP address of the packet, the destination IP address, and the size of the packet. Each snapshot captures roughly an hour of packets, and the snapshots are taken irregularly, but several times per year in the period that I’m focusing on.[2]

What makes this particular dataset useful is that one of the sources of this trace data is the Equinix Chicago IXP, which is also a contributor to the ORVP. By matching this facility’s trace data with its contemporary routing data, I can determine which links of the network each packet would likely use, and how much traffic they would create on those links. I can then aggregate to the link level in order to construct measurements of the amount of traffic originating from the IXP on each link in the network—which is a core component of my model. However, this is also somewhat of a limitation, because similar matching datasets are uncommon: the model I use will rely on the conjunction of routing and trace data from the same source, or hypothetically from related sources for which it can be argued that the routing data represents the true routing of the packets in the trace data. It is for this reason that I assume that Equinix Chicago is representative of the rest of the US, rather than simply using additional data sources to get a more complete picture.

[2] To contextualize the size of this dataset, one snapshot takes up roughly 100 GB in its compressed form, and contains observations of roughly 20 billion distinct packets. The content of the packets themselves are not included in the data, merely their metadata.

3.3.3.1 First-Party vs. Third-Party Communication Flows.

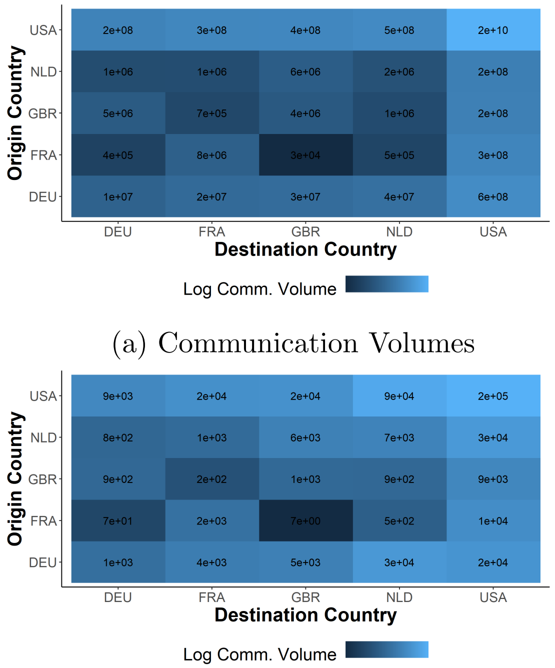
Although the Equinix Chicago IXP is located on the global Internet backbone (the skeleton of long-distance, high-bandwidth communication lines that facilitate most international communication), it is still a US-based facility, and communication within the US is heavily overrepresented in the trace data. This can be seen graphically in Figure 7a, which compares the volume of communications passing through Equinix Chicago among a subset of five countries[3]. As can be seen here, this dataset captures a much higher volume of communication between the US and partner countries than it does among those partner countries: for example, Chicago Equinix's observed volume of communication from the US to Germany or the Netherlands is more than an order of magnitude greater than that between Germany and the Netherlands.

This result is robust to controlling for the populations of origin and destination countries, as can be seen in Figure 7b, in which I plot communication volume weighted by origin and destination populations.

I therefore take the data to be representative only of international communication to and from the US: I discard observations of intra-national communication, which is unlikely to leave the US to begin with, but I also discard third-party communication flows (those which neither originate nor end in the US), on the basis that they are vastly underrepresented in the dataset and cannot thus be representative of the true sizes of those flows. This leaves me with only communication flows arriving at Chicago Equinix from outside the US, and leaving Chicago Equinix for destinations outside the US. This means that this data will be

[3] The US, the Netherlands, the UK, France, and Germany—all countries that are significant trading partners and that the US sends significant volumes of communication to.

Figure 7. Comparison of Observed First vs. Third-Party Communication



(a) Communication Volumes

(b) Communication Volumes adjusted for Partner Populations

best used in a model focused on US-origin and US-destination communication, and I focus my later empirical analysis on this context.

3.3.3.2 Use of Multiple Collectors. In reality, the US is large enough, and contains enough IXPs comparable to Equinix Chicago, that it is likely not completely representative of the US. There is nothing special about this particular source, other than the fact that it makes trace and routing data from overlapping time periods readily accessible. Thus, to obtain a more representative dataset, similar matching datasets could be obtained in the future by partnering with similar organizations in other countries, allowing this model to be applied in a broader context. In the sections to come, I specify my model to explicitly allow for the use of data from multiple collectors, each taken to be representative of a country, indexed by the subscript c.

3.3.4 Counting Internet Users.

One straightforward way to employ this data is to simply count the number of Internet-connected devices that a collector has a working route to. This can be simply done by geolocating each block of IP addresses observed in the routing data, then summing the size of each block in a country.

This measure is not a perfect measurement of the amount of communication originating from a country, as it can be affected by variations in the number of devices per user (considerably higher in developed countries) and the intensity with which a device is used (difficult to measure, but also likely higher in developed countries). However, it has one advantage over comparable measures (such as a count of the number of IP addresses officially registered to a country), as only devices which have been connected to the Internet relatively recently, and therefore have an IP address, will be observed in the data: IP addresses which have been allocated to a country but which are not in service will not be counted. Furthermore, a count of IP addresses per country will be necessary in constructing my measure of Internet traffic.

Table 14 provides summary statistics for the number of Internet-connected devices observed in the routing data, in both total and per-capita terms. Although there are idiosyncratic variations, the general trend is for both of these measures to increase over time.

3.3.5 Procedure for Constructing Link Traffic Measurements.

While the count of IP addresses is useful, it does not actually measure communication, or provide insight into which links in the global Internet are heavily used. To construct a measure of traffic, I couple the IXP's trace and

Table 14. Summary Statistics: Reachable IP addresses

Variable	**Year**	**n**	**Min**	**Median**	**Mean**	**Max**	**Std.Dev.**
IP Addresses (thousands)	2016	55	0.3	89.2	2845.5	73850.9	10824.5
	2017	55	0.5	102.4	3023.1	74454.1	11027.1
	2018	55	1.0	165.2	3180.7	75994.1	11270.0
	2019	55	4.1	224.6	3340.4	78533.9	11651.4
	all	220	0.3	141.6	3097.4	78533.9	11122.1
IP Addresses (per 1K population)	2016	51	0.3	14.0	118.9	1280.3	277.3
	2017	51	0.3	21.5	128.5	1283.3	284.5
	2018	51	0.4	26.4	139.2	1445.6	300.2
	2019	51	0.9	30.2	209.5	3879.2	617.3
	all	204	0.3	24.0	149.0	3879.2	395.2

routing datasets together, approximating the traffic generated by that facility across this network.

I begin by geolocating the origin and destination IP addresses of all observed packets, using a commercial geolocation dataset by Maxmind. I then discard all packets that are not US-origin or US-destination. I denote as $IPComm_{cij}$ the total size of all the packets observed by collector c being sent from country i to country j.[4] I also let $NumIP_j$ be the number of unique IP addresses[5] active in country j and $NumIP_\eta$ the number of addresses in block η, located in country j and observed in the routing data.

Absent any observable characteristics distinguishing IP addresses, I assume that each IP address in a country receives an equal share of communication bound for that country. I therefore assign each unique block of IP addresses in the routing data an amount of communication from c as follows:

$$BlockComm_{c\eta} = CountryComm_{cij} \times \frac{NumIP_\eta}{NumIP_j} \tag{3.1}$$

[4] Because I discard all third-party communication, in all observations of $IPComm_{cij}$, either i or j will be the country represented by c.

[5] Computed as described in Section 3.3.4.

I now couple this dataset with the matching routing data: For each block of destination IP addresses η, I identify the set of routes $R_{c\eta}$ which c would use to reach it. Since there are frequently a multiplicity of usable routes in $R_{c\eta}$, I cut down this set using the most-direct-route criterion mentioned previously, and denote the set of most-direct routes (those with fewest intermediary networks) as $R_{c\eta}^{min}$. Further, as I only observe routes from the collector to other devices, I make the assumption that if communication from c to j uses a particular route, communication from j to c uses the same route in reverse.

Having identified the routes which are the most direct way of reaching η, it is now necessary to divide the volume of communication sent to η among them. In cases where there is only one most-direct route, this is trivial, but in many cases there is a multiplicity of most-direct routes. As the routing data does not identify which routes are chosen, and does not contain values which can be used to condition on, I simply assign each route an equal share of communication from c to each most-direct route as follows:

$$RouteComm_{rc\eta} = \begin{cases} \frac{BlockComm_{c\eta}}{|R_{c\eta}^{min}|} & \text{if } r \in R_{c\eta}^{min} \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

where $|R_{c\eta}^{min}|$ is the size of $R_{c\eta}^{min}$, or the multiplicity of most-direct routes serving η.

Since this volume of communication will be sent over each link in the route, I next denote as $Traffic_{kl}(c, \eta, r)$ the amount of traffic across link kl generated by communication from the IXP to η over route r:

$$Traffic_{kl}(c, \eta, r) = \begin{cases} RouteComm_{rc\eta} & \text{if } kl \in r \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

Finally, the amount of traffic originating from c, and present on link kl, is given by summing over blocks η and routes r:

$$TotalTraffic_{ckl} = \sum_{\eta} \sum_{r} Traffic_{kl}(c, \eta, r) \tag{3.4}$$

Illustrated toy examples of this procedure's application can be found in Appendix B. A more precise method of performing this construction, which has proven vastly more computationally intensive and therefore infeasible, is discussed in Appendix C.2.

3.4 Model

I adapt the framework developed in Allen and Arkolakis (2019) (from here on, referred to as the "AA framework") to estimate two sets of communication costs (the costs of using individual country-to-country links, t_{kl}, and the expected costs of end-to-end communication between countries, τ_{ij}) using this data. This framework is well-suited to this application due to the generic nature of the trade network which it models: while previous applications include road and ocean transportation networks, the structure of the Internet is sufficiently similar that it requires minimal modifications.

This model relies on two key components: a measurement of end-to-end communication between countries, and a measurement of traffic (either total traffic, or only the traffic which ultimately originates from a particular node) flowing across each link in the network. Communication can be measured by summing the total size of all packets exchanged by a pair of countries, while the traffic measurement can be constructed as described earlier.

3.4.1 The Nature of Costs.

The τ_{ij} and t_{kl} estimated from this model are analogous to the iceberg trade costs in common usage in the trade literature. If the cost of transmitting a single unit of communication within a single network (i.e. from one device to another on the same ISP's network) is normalized to 1, then t_{kl} represents the cost of transmitting that same unit directly (without intermediary networks) from a network in country k to a network in country l.[6] The expected end-to-end communication cost τ_{ij} is likewise the expected cost of transmitting that unit all the way from a network in country i to one in country j, by whatever routes are optimal.

These costs do not represent costs directly paid by Internet users, but rather costs paid by ISPs, which are aggregated and passed on indirectly to users. In order to provide Internet access to their subscribers, ISPs must be able to connect subscribers to any other device on the Internet. If a subscriber wishes to communicate with another device on the same ISP's network, this is straightforward—but since ISPs are small, relative to the size of the entire Internet, it is far more common that an ISP must connect a subscriber with a device outside of their network. An ISP must therefore form some sort of agreements with other networks, to be allowed to send data outside of their own network. Such an agreement requires that the ISP pay a cost: this may be a monetary cost (an access fee to use a high-speed, long-distance cable, for example), or it may be an implicit cost: reciprocity agreements (akin to a barter transaction, in which a pair of ISPs simply agree to carry each other's communication) are common, but these come with added demands on an ISP's hardware and infrastructure, and thus indirectly impose a cost on each partner in the agreement.

Additionally, such costs need not be purely monetary: it may be more accurate to describe these costs as the cost of successfully transmitting information: if a link is unreliable, requiring repeated attempts to transmit a packet without

[6] In the special case where $k = l$, this is the cost of transmitting the unit from one network to another in the same country, which is observed to happen in the data.

errors, or high-latency, making it difficult to transmit time-sensitive information, this too will be captured in t_{kl} and τ_{ij}. However, these link cost measures do not map directly into a monetary cost such as "price per gigabyte" any more than iceberg trade costs map immediately into "price per 40-foot container."

The costs of constructing and maintaining an Internet link scale with distance. Longer cables naturally cost more to purchase and then install, and a longer cable also means more places that it can suffer damage from being hit by a backhoe (Poulsen (2006)), severed by a dropped anchor (Limer (2016)), or bitten into by a shark (Carter et al. (2009)).

τ_{ij} represents the cost to an ISP in i of transmitting information, on behalf of a user, to a recipient in j. It is uncommon for ISPs to price-discriminate on the basis of the destination of a user's communication; ISPs more commonly charge a lump-sum periodic subscription fee or a per-unit rate which does not vary depending on where information is sent. The costs ultimately faced by an Internet user in i could potentially be indexed by

$$C_i = \sum_j \left(\tau_{ij} \times \frac{CountryComm_{ij}}{TotalComm_i} \right) \tag{3.5}$$

where $TotalComm_i = \sum_j CountryComm_{ij}$. This is the expected cost of transmitting a unit of information from i, given the distribution of destinations for traffic originating in i. However, the different market conditions (competition, regulation, etc) in each country likely obfuscate these costs by inducing varying degrees of markup, which would make it difficult to draw a direct comparison between this index and data on Internet prices.

3.4.2 Model Environment. Let there exist a network of nodes (representing countries) connected by links (an aggregation of cables and other lines of communication). There exist a continuum of "traders", who seek to transmit information from an origin node i to a destination node j. To accomplish this, traders seek out the lowest-cost route from i to j.

A route p consists of a series of nodes $p_n,\ n = 0, 1, 2, \ldots, N$. A route from i to j begins at $p_0 = i$ and ends at $p_N = j$. The baseline cost of such a route is the product of the costs t_{kl} associated with each link along the route,

$$\tilde{\tau}_p = \prod_{n=1}^{N} t_{k_n l_n} \tag{3.6}$$

where $k_n = p_{n-1}$ and $l_n = p_n$.

However, each trader also has a personal cost of using each route, which is determined by the baseline cost and an idiosyncratic multiplicative cost factor $\epsilon_{p,\nu}$, so that the cost to trader ν of using route p is

$$\tau_{p,\nu} = \tilde{\tau}_p \epsilon_{p,\nu} \tag{3.7}$$

Allen and Arkolakis show that, when this idiosyncratic multiplier is Frechet distributed with shape parameter θ, the traders' routing choice problem yields an analytical solution for the traffic generated by a set of link costs t_{kl}. Let A be the matrix $[t_{kl}^{-\theta}]$, and let $B = (I - A)^{-1}$, the Leontief Inverse[7] of A. Finally, let X be a matrix of observed communication flows. Then, the volume of traffic induced by these costs and communication flows is given by

$$\Xi = A \odot B'(X \oslash B)B' \tag{3.8}$$

[7] In order to compute the Leontief Inverse, it is necessary for the spectral radius of A, i.e. the supremum of the absolute values of its eigenvalues, to be less than 1. In practice, this condition may be violated when the traffic matrix contains a large number of zero elements on its diagonal. The method of computing link traffic detailed earlier is not guaranteed to produce a traffic matrix with an adequate number of non-zero diagonal values, but in my experience, it has never failed to do so. This results from the fact that a sufficient number of most-direct routes in the data include "domestic" links connecting networks within a single country; enough traffic passes over these links to result in sufficient non-zero diagonal values that the spectral radius requirement is met.

where the $\odot$ and $\oslash$ operators represent Hadamard (element-wise) multiplication and division, respectively. Here, the element Ξ_{kl} is the volume of traffic flowing along link kl in the communication network.

As I am using measurements of traffic from only a single origin in the US, it is now necessary to extract from the Ξ matrix a similar measure of single-origin traffic. Allen and Arkolakis provide a convenient formula for the fraction of trade, or in this context communication, from i to j which is routed across a link kl, denoted by $\pi_{ij,kl}$:

$$\pi_{ij,kl} = (\rho \frac{\tau_{ij}}{\tau_{ik} t_{kl} \tau_{lj}})^{\theta}, \tag{3.9}$$

where $\rho \equiv \Gamma\left(\frac{\theta-1}{\theta}\right)$. Using this formula, I am able to compute the amount of Chicago-origin traffic across links kl, given communication costs τ_{ij} and t_{kl}, and volumes X_{cj}:

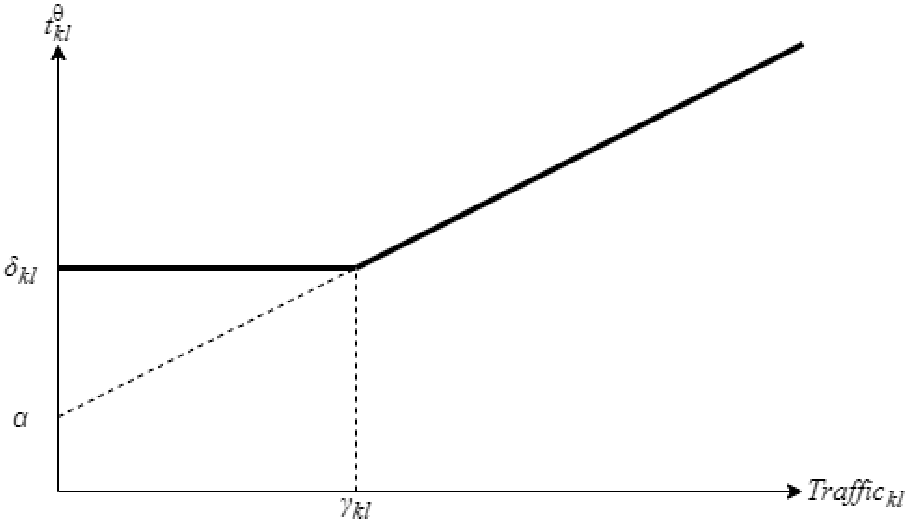
$$\Xi^{c}_{kl} = \sum_{j} \left[X_{cj} (\rho \frac{\tau_{cj}}{\tau_{ck} t_{kl} \tau_{lj}})^{\theta} \right]. \tag{3.10}$$

Link costs t_{kl} can be parameterized as a function of observable characteristics and potentially traffic levels, if congestion is expected to affect costs. Due to the significantly different factors affecting link costs in a communications network, I impose the functional form

$$t^{\theta}_{kl} = \min \left[\tilde{\delta}_{kl}, \alpha + \frac{\tilde{\delta}_{kl} - \alpha}{\tilde{\gamma}_{kl}} Traffic_{kl} \right] \tag{3.11}$$

Here, $\tilde{\delta}_{kl} \equiv \delta Z^{cost}_{kl}$ is a "baseline" cost of using link kl. This cost applies as long as the volume of traffic is less than $\tilde{\gamma}_{kl} \equiv \gamma Z^{cap}_{kl}$, the rated capacity of the link. Beyond this capacity, the cost of the link increases above the baseline cost, scaling linearly as illustrated in Figure 8. Z_{kl}^{cost} and Z_{kl}^{cap} are observables related to the cost and capacity of a link, respectively.[8] The parameters δ and γ are to be estimated.

Figure 8. Costs vs. Traffic



Given a functional form and a set of cost parameters ρ, there exists a single traffic matrix $\Xi_{pred}(\rho)$ which is rational given the costs which it induces. This traffic matrix can be found using a fixed-point algorithm which is iterated until the full traffic matrix $\Xi_{pred}(\rho)$ converges, at which point the single-origin traffic matrix $\Xi_{pred}^{c}(\rho)$ can be extracted.

The cost parameters can then be calibrated by an outer loop which searches the parameter space to minimize the distance between observed and predicted single-origin traffic,

$$|\Xi_{pred}^{c}(\rho) - \Xi_{obs}^{c}(\rho)| \tag{3.12}$$

[8] In an ideal world, they would be actual measurements of link cost and capacity, but no sufficiently complete dataset is readily available.

Furthermore, this operation can be repeated for multiple time periods t, in order to make use of panel data, so that the objective function to be minimized is

$$\sum_t |\Xi^{c,t}_{pred}(\rho) - \Xi^{c,t}_{obs}(\rho)| \tag{3.13}$$

3.5 Estimation

I initially estimate this model using routing and trace data from Equinix Chicago in 2015-2016. Due to the well-connectedness of large IXPs like this one, I make the assumption that the routes seen from this IXP are representative of the United States as a whole. However, owing to concerns that this IXP may not accurately capture the volume of "third-party" communication flowing between pairs of countries that are not the US, I restrict the dataset to only US-origin and US-destination communication.[9] The available data is sufficient to work with 171 partner countries, and covers the time periods February 2015 (the earliest available period for which routing and trace data are both available), January 2016, and April 2016 (the latest available).

3.5.1 Link Cost Parameterization. For the observables Z^{cost}_{klt} used in the parameterization of link cost, I use data on border adjacency and the presence of undersea cables, further interacted with geographical distance (as the distance crossed by a link will naturally increase its construction and maintenance costs). The undersea cable data I obtain from a GitHub repo made available by TeleGeography (TeleGeography (2020)). Z^{cost}_{kl} is thus parameterized as

$$\tilde{\delta}_{klt} = \delta_{dist}(dist_{kl}) + \delta_{adjdist}(adj_{kl} \times dist_{kl}) + \delta_{cabledist}(cable_{klt} \times dist_{kl}) + \delta_{dom}(dom_{kl}) \tag{3.14}$$

[9] A method for bypassing this restriction, given appropriate covariates, is presented in Appendix C.1

where $dist_{kl}$ is the centroid distance between countries k and l, adj_{kl} is an indicator taking the value 1 if k and l share a land border, $cable_{kl}$ is an indicator taking the value 1 if k and l are connected to the same undersea cable, and dom_{kl} is an indicator taking the value 1 when $k = l$ (used to set a cost for domestic links). Intuitively, the cost of an international link will depend in large part on the distance that the link must cover: the presence of a shared border (allowing a terrestrial cable to run directly from k to l without passing through a third country) or an undersea cable (which have generally lower maintenance costs owing to the lack of backhoes at the bottom of the ocean) merely alters the effect of distance on link cost.

Due to the scarcity of similarly detailed data on international bandwidth availability[10], I initially parameterize the bandwidth constant $\tilde{\gamma}_{kl}$ simply as a constant, i.e. $\tilde{\gamma}_{kl} = \gamma$.

I also initially allow these parameters to remain constant over time. The only cost variable which is time-varying is $cable_{klt}$, owing to a small number of new undersea cables which came online during this time period.

3.5.2 Initial Values and Estimates.

The Nelder-Mead variant used to solve the minimization problem requires a set of initial values, and unfortunately, the fixed-point algorithm in the inner loop results in an objective function with a multitude of local minima. As a result, the minimization is sensitive to the choice of initial values. In an early version of the estimation procedure, I first selected initial values by initially iterating through a discretized parameter space and selecting what were the sole set of initial values from this space which produced parameter and cost estimates satisfying two minimally-restrictive criteria:

[10] The TeleGeography dataset does include some information on cable bandwidth, which is unfortunately too incomplete to rely on.

- $\gamma_{dist} > 0$, $\gamma_{dist}+\gamma_{adjdist} > 0$, $\gamma_{dist}+\gamma_{cabledist} > 0$, and $\gamma_{dist}+\gamma_{adjdist}+\gamma_{cabledist} > 0$ so that the cost of a link between any pair of countries is increasing in distance.

- The link costs t_{kl} are all less than 10. This condition was chosen on the basis that initial runs of the model using randomly-chosen initial parameters tended to produce either costs less than 10, or extremely high values (in excess of 1000) that strained credulity in the context of iceberg costs, with little middle ground.

I have used the same initial values in successive versions of the procedure with results of similar quality in all cases.

The coefficients estimated by the model (using only US-origin and US-destination communication) are reported in Table 16, in the Baseline column. The γ and α parameters scale with the units that $Traffic_{kl}$ is measured in (e.g. converting from bytes (B) to megabytes (MB $= 1 \times 10^6 B$) would allow the γ parameters to be scaled up and the α down by 10^6.

Table 16. Coefficient Estimates

Parameter	Baseline	Varying Gamma	Discounted Parameters
δ_{dist}	$1.189e+12$	$1.186e+12$	$1.186e+12$
$\delta_{adjdist}$	$-1.181e+12$	$-1.178e+12$	$-1.178e+12$
$\delta_{cabledist}$	$-7.279e+09$	$-7.718e+09$	$-7.718e+09$
δ_{dom}	$-1.556e+06$	$-1.556e+06$	$-1.556e+06$
$\tilde{\gamma}$	$4.274e+06$		$4.358e+06$
$\tilde{\gamma}_{Feb2016}$		$4.358e+06$	
$\tilde{\gamma}_{Jan2016}$		$4.358e+06$	
$\tilde{\gamma}_{Apr2016}$		$4.358e+06$	
α	38.84	38.84	38.84
θ	27.335	28.085	28.085
λ			0.129
κ			1.05

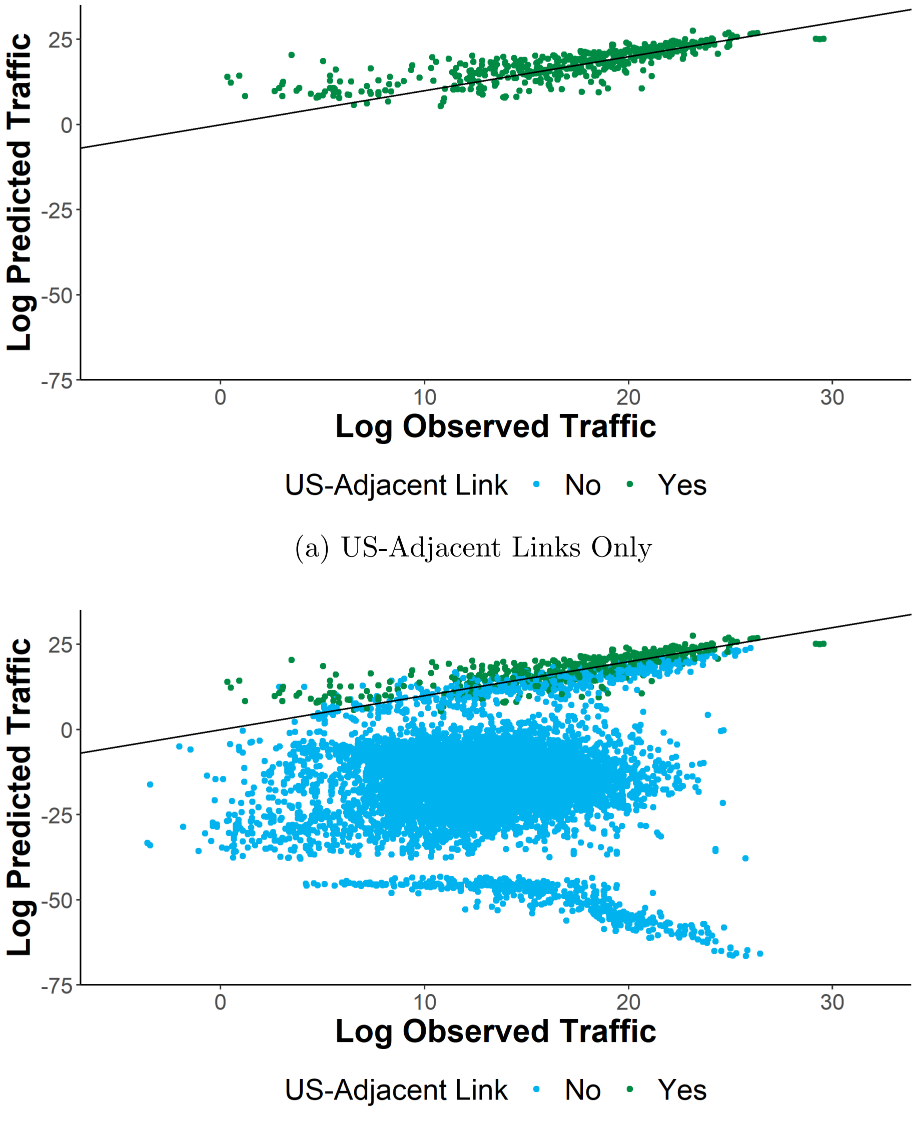
3.5.2.1 Model Fit. This model achieves a 0.796 correlation coefficient between observed and model-predicted log-traffic levels along US-adjacent links, which is where the model is expected to be the most accurate, owing to the US-centric nature of the data. Overall, the model achieves a 0.159 correlation between observed and predicted link traffic levels.

As seen in the scatter plot in Figure 9, the model accurately predicts volumes of traffic along a visually-distinguishable subset of links (recognizable in the plot as those which are close to the diagonal “45-degree” line), but vastly underestimates the traffic across other links. The links on which traffic is accurately predicted include most US-adjacent links as well as a subset of non-US-adjacent links. There is no discernable pattern which explains which links are accurately predicted: the geographical distance covered by these links varies widely, and there are links in this set that have shared land borders, undersea cables, both, or neither.

The interpretation which emerges from these results is that this model over-costs some links in the network, resulting in the drastic underprediction of traffic on those links. Given the sparseness of the cost parameterization, I now begin to examine alternate parameterizations:

3.5.2.2 Estimated Costs. The distribution of link and expected communications costs estimated using this data are shown in Figure 10a. As can be seen, expected communications costs are only slightly greater than link costs, indicating that it is rare for a route to be significantly more expensive (taking into account the Frechet-distributed idiosyncratic route cost multiplier) than the direct connection with no intermediate nodes.

Figure 9. Observed vs. Predicted Traffic (Log Scale)



(a) US-Adjacent Links Only

(b) All Links

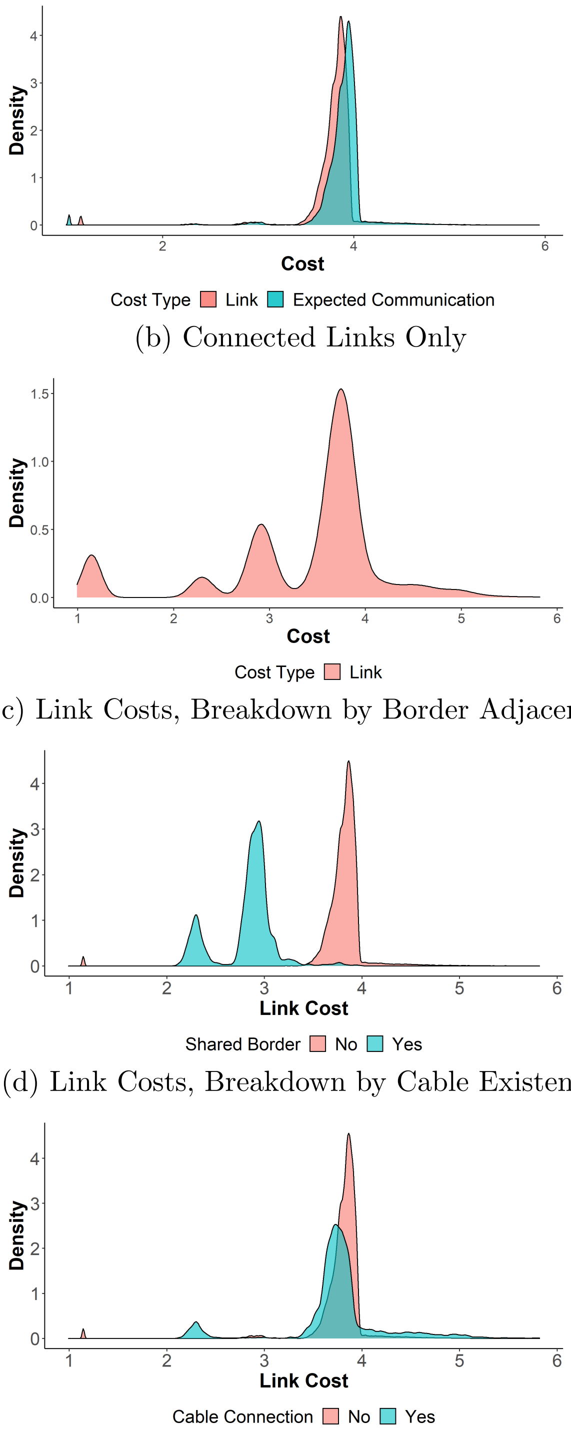
Additionally, it can be seen that expected communication costs for domestic links (seen at the far left of the diagrams, with costs only slightly larger than 1) are in fact less than the corresponding link costs. The interpretation of this result is that there must be some agents in this model whose draw of the idiosyncratic cost multipliers makes the cost of sending purely domestic communication out of and then back into the country less expensive than routing it purely within the country. (Or to phrase it differently, if the least-costly route for domestic communication were always the purely domestic route, the expected communication costs would be distributed with their mean at the domestic link cost.) While counterintuitive, this is actually a recognized phenomenon in Internet routing, called tromboning. It occurs when networks are not sufficiently interconnected for a direct domestic route to be cheaper than the most direct international route.

Figure 10b shows the distribution of link costs for just the "connected" links between countries with shared land borders or cable connections. As seen there, these costs fall into three rough categories: the category with lowest costs consists largely of links with both a shared border and a cable, the middle category consists mostly of links with only a shared border, and the high-cost category, which includes the right tail of the distribution, consists of those links with neither shared border or cable connection. (Links with only a cable connection are scattered throughout the middle and upper groups, but are relatively rare.) Figures 10c and 10d illustrate these breakdowns further. It should be noted that the fat right tail of the distribution, in which the costs are greater than 4, is largely composed of transoceanic links to and from the US, which are expensive due to sheer distance.

3.5.3 Time-Varying Bandwidth. As a robustness check, I examine whether allowing the cost parameters to vary over time impacts the results

Figure 10. Distributions of Link and Expected Communication Costs

(a) Overall Distributions



(b) Connected Links Only

(c) Link Costs, Breakdown by Border Adjacency

(d) Link Costs, Breakdown by Cable Existence of the model. Admittedly, 2015 to 2016 is not a wide time interval, but given Moore's Law[11], it seems plausible that there could be significant reductions in communication costs year-to-year. I first allow the gamma parameter, representing the bandwidth cap of undersea cables, to vary over time, using the specification

$$t^{\theta}_{klt} = \min\left[\tilde{\delta}_{kl}, \alpha + \frac{\tilde{\delta}_{kl} - \alpha}{\tilde{\gamma}_{klt}} Traffic_{klt}\right] \tag{3.15}$$

Parameters are reported in Table 16 under the Varying Gamma column. The $\tilde{\gamma}$ parameters are extremely similar, but not completely identical; interestingly, allowing the gamma parameters to vary over time using the same initial values has resulted in a slightly higher estimate of θ. A comparison of estimated costs and predicted traffic is shown in Figure 11: as seen here, the change of specification has little impact on predicted traffic volumes, but slightly reduces link costs from their values in the baseline estimation. Correlation between observed and predicted traffic levels is similar to that in the baseline model.

3.5.4 Time as Proxy for Quality of Connection.

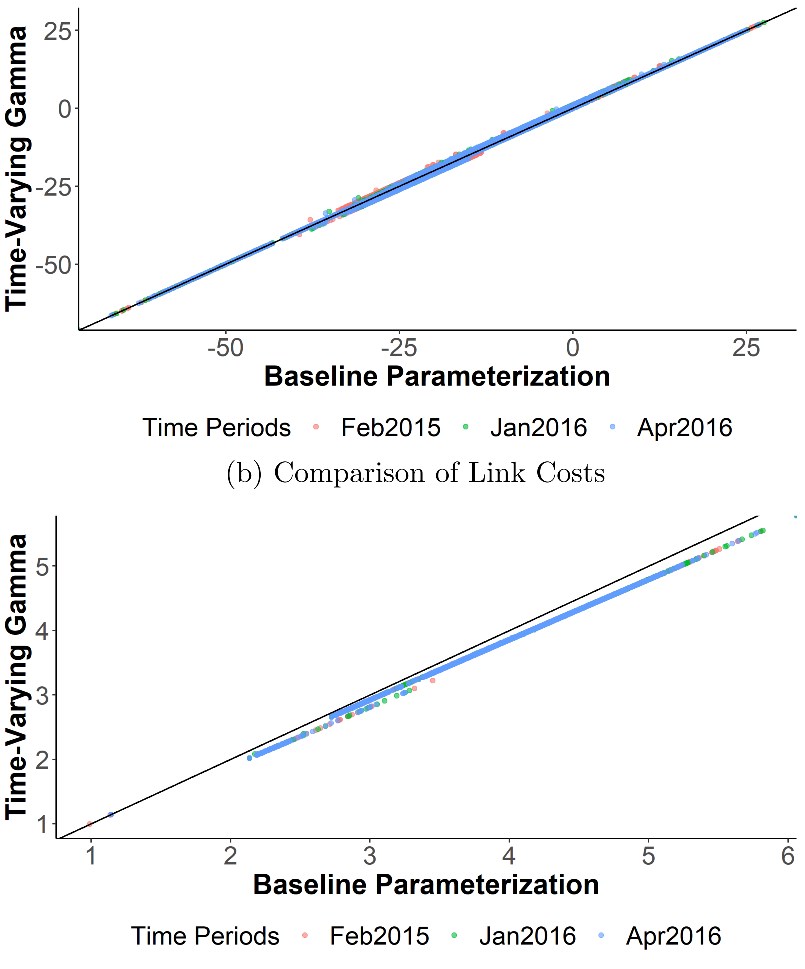
While data on the operation cost or rated capacity of undersea cables does exist, it is not complete enough to apply in this context. However, it can be assumed that the quality of Internet infrastructure improves over time, while the cost of such infrastructure decreases. Since the data on undersea cables from TeleGeography does include the date of activation for each cable, it is possible to use the time since the last cable on a link became active as a proxy for the quality of the link. This allows me to redefine the $\tilde{\delta}$ and $\tilde{\gamma}$ parameters, from the original parameterization, as follows:

$$\tilde{\delta}_{klt} = \delta_{dist}(dist_{kl}) + \delta_{adjdist}(adj_{kl} \times dist_{kl}) + \lambda^{t-t'}\delta_{cabledist}(cable_{klt} \times dist_{kl}) + \delta_{dom}(dom_{kl}) \tag{3.16}$$

[11] An informal but widely-accepted observation that computing power tends to halve in cost, or double in effectiveness holding cost constant, every one to two years

Figure 11. Baseline vs. Varying-Gamma Cost Specifications

(a) Comparison of Log Predicted Traffic



(b) Comparison of Link Costs

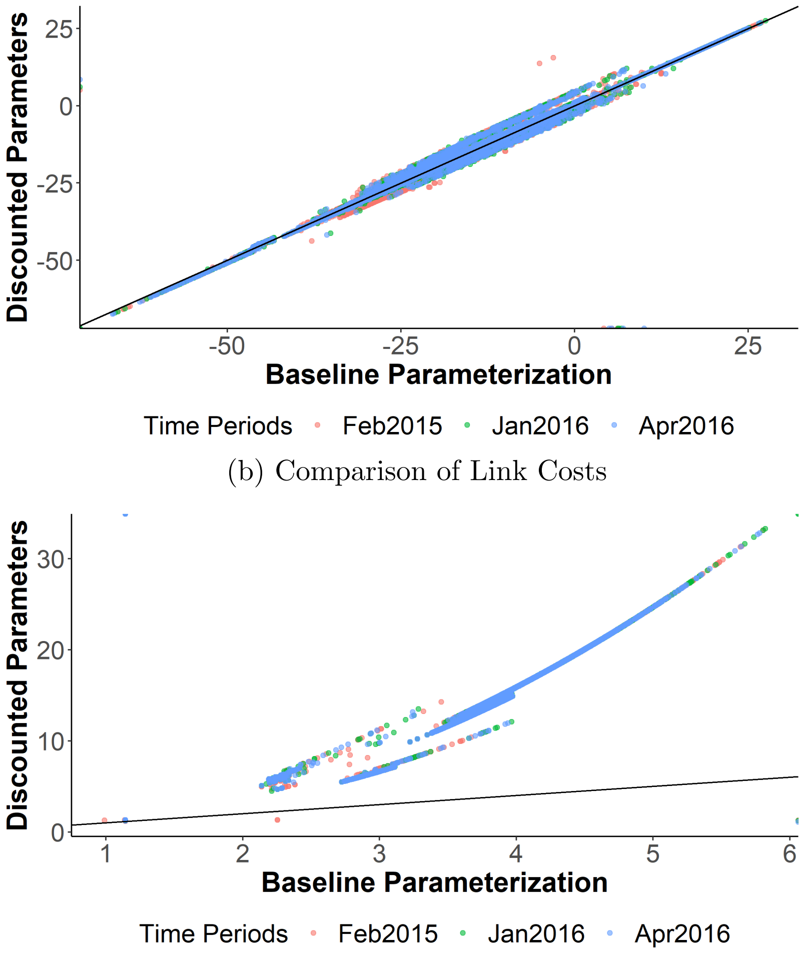
$$\tilde{\gamma}_{klt} = \begin{cases} \kappa^{t-t'}\gamma \text{ if } cable_{klt} = 1 \text{ and } adj_{kl} = 0 \\ \\ \gamma \text{ otherwise} \end{cases} \tag{3.17}$$

Here, κ and λ are constants between 0 and 1, and $t - t'$ is the elapsed time, in years, between the time period t and the time at which the last undersea cable serving the link was constructed, t'. The κ and λ factors apply geometric discounting to the constants governing cost of a link equipped with an undersea cable and the rated bandwidth of such a link, respectively. This allows for operating cost to increase and rated bandwidth to decrease for links where the cables are older. This rests upon the assumption that for connections other than undersea cables, there is constant small-scale investment keeping the connection's technology up-to-date, as opposed to undersea cables which require significant lump-sum investment to build, replace, or update.

Parameters are reported in Table 16 under the Discounted Parameters column. The $\tilde{\tilde{\delta}}$ parameters are similar to those already estimated, but are closest to those estimated for the time-varying Gamma model. Interestingly, the discount factor λ is quite small at 0.129, indicating that the value of an undersea cable connection drops off rapidly after coming into service—far more rapidly, indeed, than conventional wisdom such as Moore's Law[12] would suggest. The 0.129 estimate would suggest that the effectiveness of undersea cable technology to reduce communication costs doubles every 4 months, such that a 1-year old cable is roughly 1/8 as effective as a modern equivalent, which is difficult to believe, since it vastly exceeds the commonly-accepted rate of technological advancement suggested by Moore's Law.

[12] A observation by noted engineer and former Intel CEO Gordon Moore that the number of transistors on a silicon chip doubles every 1-2 years, but often generalized to mean that computing power or the general effectiveness of computing technology doubles in that period.

Figure 12. Baseline vs. Discounted-Parameters Cost Specifications



(a) Comparison of Log Predicted Traffic

(b) Comparison of Link Costs

A comparison of estimated costs and predicted traffic is shown in Figure 12: as seen here, the change of specification still has little impact on predicted traffic volumes, but due to the introduced discounting factors, vastly increases the estimated costs of the links in a nonlinear fashion. Despite this, correlation between observed and predicted traffic levels is similar to that in the baseline model.

3.5.5 Underprediction of Traffic. All of these cost specifications produce very similar predictions of traffic, and these predictions understate the amount of traffic on a large number of links. This, in turn, implies that the costs of these links are overestimated. Why is this the case in all three specifications? Consider the following possibilities:

1. **Omitted variables in the cost parameterization**: At its core, the functional form I use for link costs contains few variables, owing to the scarcity of complete data on the infrastructure associated with these links. There may be important cost-reducing factors which I do not have data for. In particular, the cost parameterization would be improved by a more complete dataset on undersea cable bandwidth, or even better, the bandwidth of terrestrial cables crossing land borders. Such data would allow for γ, the parameter governing link bandwidth, to be given a more nuanced functional form than the constant or discounted-constant value it takes in my specifications.

2. **Flawed assignment of communication to redundant routes:** Recall that when a multiplicity of routes exists, I assign an equal share of communication to each route, as visualized in Figure B.2b. I do this due to a lack of observable characteristics upon which to base a more nuanced division of traffic (and, also, because it would take a prohibitive amount of time to parameterize this split and search for the ideal parameter values in my estimation process). However, it may be the case that, by assigning communication in this simplistic way, I am creating an observed traffic dataset which overstates the amount of traffic among some links, by assigning too much communication to routes which are only in the routing data as a redundant backup. With more detailed information about how a route is selected, it would be possible to refine the method by which communication is assigned to routes.

3. **Internet entities undercost some links:** A third possibility, and one that I do not place any particular emphasis on, is that the routes chosen by Internet entities are not necessarily optimal, or are optimal given some constraint which I do not model. If the routes observed in the routing data are not themselves optimal for the simplified environment I model, then the traffic I compute from it would also not be an optimal distribution of traffic. This, again, might be fixable with improved information from the providers of the routing data, as it might be possible to model the factors which affect routing choice as part of the cost function.

Each of these possibilities represents a hypothesis which is non-trivial to test, owing to their reliance on data which is so far not readily available. I therefore consider the hypotheses to be fertile avenues for further research.

3.6 Explanatory Power Applied to Trade Volumes

With the expected communication costs estimated, I now turn to applying them in a straightforward application: a gravity model of international trade. Using trade data from COMTRADE (United Nations (2003)) for the years 2015-2016 and the communication costs estimated using non-adjusted data, I estimate the following simple gravity models:

$$log(Trade_{ijt}) = \beta_0 log(dist_{ij}) + FE_i + FE_j + FE_t + \epsilon_{ijt} \tag{3.18}$$

$$log(Trade_{ijt}) = \beta_1 \tau_{ijt} + FE_i + FE_j + FE_t + \epsilon_{ijt} \tag{3.19}$$

$$log(Trade_{ijt}) = \beta_0 log(dist_{ij}) + \beta_1 \tau_{ijt} + FE_i + FE_j + FE_t + \epsilon_{ijt} \tag{3.20}$$

where $Trade_{ijt}$ is COMTRADE's measure of trade volume, $dist_{ij}$ is the same centroid distance used earlier, and τ_{ijt} is the expected trade cost extracted using my model. This model uses the simplest possible fixed effects, comprising origin, destination, and year. Results are shown in Table 17.

As can be seen in the table, by themselves the extracted communication costs have an interpretation similar to that of distance, i.e. as a resistance term in the gravity equation, while possessing somewhat less explanatory power. When coupled together, distance absorbs much of the explanatory power of the extracted communication cost, which is to be expected considering that distance is explicitly a factor which contributes to link costs in my parameterization of the link cost function. The coefficient on communication costs is highly significant in all models where it is included, with p-values less than $2.2 \times e^{-16}$.

Thanks to the inclusion of multiple years of data (albeit condensing both 2016 observations into one year by taking the mean communication cost), I also estimate the second and third models replacing the origin, destination, and year fixed effects with origin-year and destination-year fixed effects. I omit origin-destination fixed effects, as they would absorb the distance term, which I wish to retain for comparison. As can be seen in columns (1-2) of Table 18, this has very little impact on the estimated coefficients or explanatory power of the models, indicating that the estimated communications costs do not simply proxy for other origin-year- or destination-year-varying factors.

Columns (3-4) of Table 18 repeats this exercise with trade in services replacing trade in goods. Results are qualitatively similar, although note that the elasticity of trade value, with respect to either distance or communication cost, is smaller for services than for goods.

Table 17. Regression Results: Simple Fixed Effects

	Dependent variable:		
	Log Trade in Goods		
	(1)	(2)	(3)
Log Distance	-2.094^{***}		-1.976^{***}
	(0.016)		(0.021)
Log Comm. Cost		-15.677^{***}	-1.980^{***}
		(0.197)	(0.230)
Fixed Effects	i, j, t	i, j, t	i, j, t
Observations	33,154	33,154	33,154
R^2	0.760	0.698	0.760
Adjusted R^2	0.757	0.695	0.758
Residual Std. Error	2.028 (df = 32849)	2.274 (df = 32849)	2.026 (df = 32848)

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 18. Regression Results: Interacted Fixed Effects

	Dependent variable:			
	Log Trade in Goods		Log Trade in Services	
	(1)	(2)	(3)	(4)
Log DIstance		-1.972^{***}		-1.416^{***}
		(0.022)		(0.029)
Log Comm. Cost	-15.868^{***}	-2.005^{***}	-7.159^{***}	-0.635^{***}
	(0.199)	(0.233)	(0.221)	(0.235)
Fixed Effects	it, jt	it, jt	it, jt	it, jt
Observations	33,154	33,154	7,798	7,798
R^2	0.701	0.762	0.814	0.860
Adjusted R^2	0.695	0.758	0.797	0.847
Residual Std. Error	2.272 (df = 32558)	2.026 (df = 32557)	1.334 (df = 7140)	1.160 (df = 7139)

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

3.6.1 Heterogeneity Analysis: Breakdown by Categories of Goods and Services.

I now turn my attention to the search for deeper patterns of trade, specifically, for sectors of the economy which are affected more severely by elevated communication costs.

3.6.1.1 Heterogeneity in Trade of Goods.

I begin by breaking down trade in goods in more detail: since the sheer number of goods classifications in my COMTRADE data makes analysis on that level difficult, I instead apply the classification of goods used in Rauch (1996), which divides goods into categories of commodities, reference-priced products, and differentiated products. Using Rauch's published concordance of SITC codes to goods categories and a dataset on US exports of goods broken down by SITC code, I estimate the models

$$log(Trade_{jtg}) = \beta_0 log(\tau_{jt}) \times r(g) + FE_j + FE_t + FE_{r(g)} + \epsilon_{jtg} \tag{3.21}$$

$$log(Trade_{jtg}) = \beta_0 log(\tau_{jt}) \times r(g) + \beta_1 X_{jt} + FE_j + FE_t + FE_{r(g)} + \epsilon_{jtg} \tag{3.22}$$

where the subscript g refers to goods classified by SITC code, and $r(g)$ is a vector of indicator variables corresponding to the three Rauch classifications, each of which takes the value 1 if good g is of that classification, and 0 otherwise. In the second model, X_{jt} is a vector of destination-year controls including real GDP and population. I again condense both sets of 2016 communication costs into an average cost corresponding to the 2016 trade data, limiting the analysis to two time periods, 2015 and 2016. This specification allows for the elasticity of trade in goods to vary depending on the degree of heterogeneity a category of goods exhibits, represented by β_0 being a vector of coefficients corresponding to Rauch classifications. I omit physical distance, as it is absorbed by the destination-time fixed effect. Results are somewhat counterintuitive: as seen in Table 19, exports of commodities (about which little communication is necessary to establish the properties of the good) are reduced the most by increased communication costs; exports of reference-priced goods are reduced to a lesser degree, and in the case of differentiated goods, the effect is non-significant.

Table 19. Regression Results: Heterogeneity by Rauch Classification

	Dependent variable:	
	Log Trade in Goods	
	(1)	(2)
Log Comm. Cost× Commodity	-1.408^{***}	-1.343^{***}
	(0.170)	(0.173)
Log Comm. Cost× Ref-Priced	-0.434^{***}	-0.400^{**}
	(0.162)	(0.165)
Log Comm. Cost× Differentiated	0.146	0.183
	(0.159)	(0.162)
Log GDP		1.280^{***}
		(0.199)
Log Population		1.127
		(1.335)
Log Capital Stock		-2.442^{***}
		(0.657)
Observed Flows	US Exports	US Exports
Fixed Effects	j, t, r(g)	j, t, r(g)
Observations	467,616	466,162
R^2	0.258	0.258
Adjusted R^2	0.258	0.258
Residual Std. Error	2.675 (df = 467457)	2.677 (df = 466009)

*p<0.1; **p<0.05; ***p<0.01

This is the reverse of what conventional wisdom suggests should occur, in which differentiated products, which may require significant amounts of description to convey their unique product characteristics, should suffer the most from increased communication costs. A potential explanation for this pattern can be found in Keller and Yeaple (2013), which finds that multinational firms respond to greater communication costs ("disembodied knowledge transfer costs," in the terminology of Keller and Yeaple) by importing goods from their foreign affiliates that require greater knowledge to produce (an increase in the "embodied knowledge" embedded in their affiliate imports). In other words, multinational firms may respond to increased communication costs by centralizing production of more sophisticated, i.e. differentiated, products and importing the completed good, rather than importing intermediate goods which are then assembled locally.

This result is robust to the inclusion of destination-year controls, as seen from the minimal differences between the common coefficients of Models 1 and 2 in Table 19. I therefore couple this data with a dataset measuring what is nominally the universe of imports and exports to and from the US among related parties. [13]

Using the related-party trade data allows me to estimate the regression models

$$log(RelatedTrade_{jtg}) = \beta_0 log(\tau_{jt}) \times r(g) + FE_j + FE_t + FE_{r(g)} + \epsilon_{jtg} \qquad (3.23)$$

$$log(RelatedTrade_{itg}) = \beta_0 log(\tau_{it}) \times r(g) + FE_i + FE_t + FE_{r(g)} + \epsilon_{itg} \qquad (3.24)$$

Here, $r(g)$ is a vector of indicator variables corresponding to the three Rauch classifications, each of which takes the value 1 if good g is of that classification, and 0 otherwise.

Results are reported in Table 20. As can be seen there, the τ_{ijt} cost measure has its largest effects on related-party trade in commodities, while reference-priced goods are not affected to a significant degree, and imports of differentiated goods in fact increase as communication costs rise. This result is suggestive evidence in

[13] While the dataset does represent the universe of flows observed by the US Bureau of Customs and Border Protection, documentation on the dataset does acknowledge that importers and exporters do not always report the indicator that identifies a shipment as a related-party transaction.

favor of multinational corporations shifting away from trade in commodities and towards trade in differentiated goods, which tend to be more complex and therefore embody a greater concentration of knowledge. Again, these results are robust to the addition of country-time control variables, as seen in Models 3 and 4 of Table 20.

Table 20. Regression Results: Heterogeneity in Related-Party Trade

	Dependent variable:			
	Log Trade in Goods			
	(1)	(2)	(3)	(4)
Log Comm. Cost× Commodity	-3.244^{**} (1.481)	-15.005^{***} (1.925)	-2.853^{*} (1.516)	-15.041^{***} (1.967)
Log Comm. Cost× Ref-Priced	0.151 (0.950)	−1.372 (1.296)	0.214 (0.973)	−1.481 (1.341)
Log Comm. Cost× Differentiated	−0.267 (0.855)	2.523^{**} (1.149)	−0.146 (0.876)	2.259^{*} (1.194)
Log Dest. GDP			0.343 (0.667)	
Log Dest. Population			−3.660 (5.944)	
Log Dest. Capital Stock			2.819 (2.467)	
Log Orig. GDP				−0.899 (0.955)
Log Orig. Population				1.017 (9.082)
Log Orig. Capital Stock				0.737 (3.714)
Observed Flows	US Exports	US Imports	US Exports	US Imports
Fixed Effects	j, t, r(g)	i, t, r(g)	j, t, r(g)	i, t, r(g)
Observations	5,685	4,715	5,408	4,541
R^2	0.536	0.519	0.532	0.512
Adjusted R^2	0.522	0.501	0.518	0.494
Residual Std. Error	2.312 (df = 5512)	2.910 (df = 4544)	2.323 (df = 5246)	2.917 (df = 4380)

*p<0.1; **p<0.05; ***p<0.01

However, it is still possible to look deeper, and couple this related-party trade data with the industry knowledge intensity measures used in Bahar, Hausmann, and Hidalgo (2014) and Bahar (2019). These measures combine worker-level information to quantity the degree of "tacit knowledge" used in industries classified by SITC and NAICS codes. Using this data, I estimate the regressions

$$\begin{aligned} log(Exports_{jtg}) =& \beta_0 log(\tau_{jt}) \times r(g) + \beta_1 log(\tau_{jt}) \times log(Knowledge_g) \times r(g) \quad (3.25) \\ &+ FE_j + FE_t + FE_{r(g)} + \epsilon_{jtg} \\ log(Exports_{itg}) =& \beta_0 log(\tau_{it}) \times r(g) + \beta_1 log(\tau_{it}) \times log(Knowledge_g) \times r(g) \quad (3.26) \\ &+ FE_i + FE_t + FE_{r(g)} + \epsilon_{itg} \end{aligned}$$

Here, $Knowledge_g$ is the industry-knowledge measure for industry g, taken from the Bahar data.

Table 21. Regression Results: Heterogeneity Controlling for Knowledge-Intensity

	Dependent variable:	
	Log Trade in Goods	
	(1)	(2)
Log Comm. Cost× Ref-Priced	7.655***	−0.067
	(0.852)	(1.306)
Log Comm. Cost× Differentiated	−8.949***	−4.299***
	(0.754)	(1.085)
Log Comm. Cost× Ref-Priced× Knowledge	0.053	0.516***
	(0.124)	(0.185)
Log Comm. Cost× Differentiated× Knowledge	1.728***	1.076***
	(0.047)	(0.067)
Observed Flows	US Exports	US Imports
Fixed Effects	j, t	i, t
Observations	6,075	4,943
R^2	0.651	0.607
Adjusted R^2	0.640	0.593
Residual Std. Error	2.021 (df = 5902)	2.686 (df = 4772)

*p<0.1; **p<0.05; ***p<0.01

Unfortunately, coupling the Bahar data with the related-trade dataset results in a highly imbalanced panel due to missing observations; there is only one commodity good observed in the related-party trade data that can be matched to the knowledge data, which creates issues of multicollinearity between fixed effects. I therefore drop the problematic commodity-goods classification. Once again I estimate the model separately for imports and exports, with results reported in Table 21. The estimated coefficients for differentiated goods show, firstly, that all else equal, communication costs do negatively impact trade in differentiated goods, but secondly, that this effect is reduced or reversed for differentiated goods from knowledge-intensive industries. At the mean level of knowledge-intensity (weighted by the size of the export flow), the total coefficient on $log(\tau_{ijt})$ (calculated as $\beta_0 + \beta_1 \times log(\overline{Knowledge_g})$) is 0.100 for exports and 1.308 for imports, which confirms the earlier result suggesting that differentiated goods are traded more in situations with greater communication costs.

This regression also provides a more nuanced analysis of effects on reference-priced goods, which experience a net increase in trade volume from communication costs. There is some difference between effects on exports of reference-priced goods, which are driven primarily by communication costs with no significant effect from knowledge-intensity, and on imports, where the reverse is true. However, in both cases, the total coefficient on log_{ijt} at mean levels of knowledge-intensity is much larger than the corresponding total coefficient for differentiated goods (at 7.887 for exports and 2.595 for imports). Thus, controlling for industry knowledge-intensity, it is now apparent that, at least with trade among related parties, communication costs drive a shift away from trade in commodity goods and towards more complex goods that allow for knowledge to be embodied instead of requiring difficult and expensive international communication.

3.6.1.2 Heterogeneity in Trade of Services. A similar analysis can be conducted with trade in services, aided by the fact that services are grouped into vastly fewer categories by EBOPS codes, and therefore bilateral service trade data is much less time-intensive to acquire through COMTRADE. Using a dataset of services trade flows reported by 133 countries, I estimate the model

$$log(Trade_{ijtg}) = \beta_0 log(dist_{ij}) + \beta_{1g} log(\tau_{ijt}) + FE_{it} + FE_{jt} + FE_g + \epsilon_{ijtg} \tag{3.27}$$

which contains a wider array of fixed effects and allows me to have significantly more observations. A summary of results is shown in Table 22.

This model again produces results which go against the conventional wisdom that increased communication costs should have a purely negative effect on the value of trade in services. Instead, I observe communication costs having a mixture of positive and negative effects on trade volumes.

When I consider the types of services which experience positive or negative effects on trade volume from communication costs, a pattern emerges.

1. One broad category of goods, which I refer to as communication-delivered goods, includes those which can be exported using the Internet or other forms of communication, or which are made significantly easier to export. This type of service includes such items as health services (e.g. via telemedicine), construction abroad (which benefits from rapid exchanges of architectural plans, etc.), auditing, bookkeeping and tax consultation (all of which involve by their very nature extensive exchanges of financial data). These goods generally experience a decrease in trade value when communication costs rise.

2. A second type of service, which I refer to as communication-produced goods, includes those which can be more easily produced with easy access to communication: this type includes computing services (such as web hosting, online payment processing, etc.), research and development, advertising

Table 22. Most Heavily-Affected Service Sectors by EBOPS Code

Top 10 categories most positively affected by communication cost	
Transportation/Air/Passenger	7.356^{***}
Other business services/Miscellaneous/Research and development	6.443^{***}
Other business services/Miscellaneous/Advertising, market research, and public opinion polling	5.005^{***}
Insurance services/Auxiliary services	4.93^{***}
Other business services/Merchanting and other trade-related services/Other trade-related services	4.128^{***}
Communications services/Telecommunications services	3.935^{***}
Other business services/Miscellaneous/Legal, accounting, management consulting, and public relations/Legal services	3.851^{***}
Other business services/Miscellaneous/Other business services	3.279^{***}
Royalties and license fees/Other royalties and license fees	3.044^{***}
Insurance services/Reinsurance	2.753^{***}
Top 10 categories most negatively affected by communication cost	
Personal, cultural, and recreational services/Other personal, cultural, and recreational services/Health services	-10.31^{***}
Transportation/Other/Passenger	-8.316^{***}
Transportation/Other/Freight	-8.173^{***}
Transportation/Sea transport/Passenger	-7.789^{***}
Construction services/Construction abroad	-7.087^{***}
Other business services/Miscellaneous/Agricultural, mining, and on-site processing services/Waste treatment and depollution	-6.996^{***}
Government services, n.i.e./Embassies and consulates	-6.698^{***}
Personal, cultural, and recreational services/Other personal, cultural, and recreational services/Other	-6.16^{***}
Transportation/Other/Other	-6.024^{***}
Insurance services/Life insurance and pension funding	-3.689^{***}

and market research. These services generally experience an increase in trade value when communication costs rise, as countries with generally high communication costs find it difficult to produce these goods and services for themselves and substitute towards importing. (To give one example: a country with high communication costs would find internet hosting services expensive to produce domestically, leading consumers in these countries to host their websites abroad, in countries with lower communication costs.)

3. The third category of services are essentially substitutes for communication, largely restricted to transportation and telecommunication services. Physical transportation can be used to transport personnel in lieu of telecommunication, or to export goods as a substitute for exporting services, while telecommunication services naturally become more expensive as communication costs rise; as such, it is expected for the value of telecommunication service exports to rise with communication costs unless there is a price effect causing volume to decrease by a large amount. The effect of increased communication costs is erratic here, with some categories (such as air passenger transportation) seeing large increases in volume with increased costs, and others (such as sea passenger transport) seeing similar decreases in volume.

Communication-delivered goods experience negative effects on trade volumes as a result of increased communication costs—exactly what the conventional wisdom suggests would occur, since these costs make it more expensive to export such goods. The other two categories experience positive effects on trade volume instead, which on close consideration seems entirely plausible:

In the case of communication-produced goods, a country which finds itself with expensive communications will also find it expensive to produce these goods. To give one straightforward example, in a country with expensive communications, web-hosting companies will face greater costs to provide a fixed level of service (as measured by latency, reliability, etc.). Alternately, these companies may choose to produce an inferior level of service. Neither of these options lends itself to producing web-hosting services domestically, and in fact this country may become a net importer of web-hosting (i.e. individuals and firms in this country may pay to host their websites and data in countries with cheaper communications).

In the case of communication-substitute goods, the rationale behind increasing trade volumes as a result of increasing communication costs is even more straightforward: faced with communication costs making long-distance communication impossible, firms may opt to send personnel between countries (to gain first-hand experience, confer with colleagues in person, or perform complicated procedures). This is akin to the outcome described in Duranton and Storper (2008), in which it becomes cost-prohibitive to export complex machinery when communication costs are high, as the amount of physical travel necessary to convey the client's specifications for a machine becomes significant. Alternately, countries which have expensive communication may choose to specialize in producing goods, not services, which also increases the quantity of transportation services necessary.

3.7 Conclusions

The approach I have described allows for measures of communication cost to be extracted from reasonably-accessible data on Internet routing and communication. These measures have explanatory power when used to model trade volumes, and allow for the effect of physical distance on trade volumes to be separated from the effect of communication cost (which is affected by physical distance, but incorporates other components as well).

Analysis using this data reveals trends in how communication costs affect trade, that run counter to the conventional wisdom. Coupled with data on related-party trade and industry knowledge-intensity, these trends can be explained as a result of a substitution pattern: multinational firms with greater costs of communication, rather than coordinating complex global supply chains, instead perform a greater degree of transformational work in individual countries so that institutional knowledge can be "embedded" into complex goods. Given that large portions of global trade are performed by multinationals (estimates suggest values ranging from a third to a half), this substitution pattern is a noteworthy line of future inquiry.

There remain several avenues for further work on the estimation of communication costs: supplemental data regarding Internet infrastructure remains scarce, and this scarcity restricts what variables can be used to parameterize costs. This likely contributes to the major flaw of the model, which is its tendency to underpredict traffic along links far from the source of the Internet data. However, neither the scarcity of supplemental data nor the underprediction problem represent insurmountable barriers to the use of this method as a way of measuring communication costs.

CHAPTER IV
INTERACTIONS BETWEEN COMMUNICATION COSTS AND LANGUAGE BARRIERS: IMPLICATIONS FOR CROSS-BORDER INVESTMENT FLOWS

4.1 Introduction

Much as communication costs and other information frictions may be barriers to trade in goods and services, they may also be barriers to investment. A lack of cheap or effective communication translates into greater difficulty in judging the value of a foreign asset, greater monitoring costs, and simply greater transaction costs to acquire an asset. All of these factors are greatly alleviated by the rapid, low cost communication enabled by access to the Internet, meaning that Internet access may have effects on international financial flows.

In this chapter, I employ the Internet communication costs described and estimated in the previous chapter in an analysis of cross-border portfolio investment flows, using the Finflows (Nardo, Ndacyayisenga, Pagano, and Zeugner (2017)) and Treasury International Capital (US Treasury (2022)) datasets. I am unable to reject the hypothesis that the effect of communications costs on portfolio investment is zero when using a full complement of bilateral fixed effects, but when replacing some fixed effects with controls find that the effect is significant and negative. This combination of results suggests that while my measure of communication costs does not itself explain variation in cross-border portfolio investment, it does act as a proxy for other country-pair varying factors that do. Thus, this measure of communication costs may be useful as an explanatory variable in contexts that do not permit the use of said origin-destination fixed effects (such as a cross-sectional analysis without time variation).

4.2 Literature Review

4.2.1 Information Frictions and Portfolio Investment.

This chapter builds on the literature surrounding information frictions and trade, including the large body of work on communication costs and trade (Freund and Weinhold (2004), Allen (2014), Leuven et al. (2018), Fink et al. (2005), Lew and Cater (2006), Ejrnæs and Persson (2010), Steinwender (2018)). This literature employs a wide variety of proxies for communication costs, ranging from the time cost of communication to the costs of telegraph and telephone communication. My previous chapter provides a method of estimating communication costs from data on Internet communication and routing, which is a more direct measurement of the costs of Internet communication.

The finance literature deals with similar frictions and their effects on cross-border portfolio investment (Beneish and Yohn (2008), Berkel (2007)), specifically as a potential explanation for home bias, the tendency for investment portfolios to over-invest in their home countries. I therefore bring my previously-developed measure of Internet communication cost to bear on this question; in the literature there are widespread uses of common language, legal system, colonial backgrounds, etc. as proxies for information frictions, but these can all be described as proxies for the effectiveness of communication, rather than its cost.

As discussed in Chapter III, to date there have been few options for data that can be used to measure a cost of Internet communication, as many of the proxies in the literature are either associated with older communications technology, too small scale to be applied in a multi-national analysis, or too difficult to effectively compile. This is, therefore, a novel approach to examining the impact of Internet communication cost on cross-border investment activity.

4.2.2 Determinants of Portfolio Investment.

The literature on the determinants of cross-border portfolio investment is well-developed: Roque and Cortez (2014) provide an extensive summary of variables used in the literature to analyze equity portfolios. Although I initially employ models that make extensive use of fixed effects, I use a set of controls in alternate specifications when searching for heterogeneity in contexts that do not allow for the full set of fixed effects.

Roque and Cortez (2014) provide broad categories of determinants of equity investment, each containing a selection of appropriate variables used in the finance literature. I have included a selection of controls such as returns, GDP, population and capital stocks (Mishra (2007), Coeurdacier and Martin (2007), Faruqee and Yan (2004), barriers to cross-border investment (Lane and Milesi-Ferretti (2008), Ferreira and Miguel (2007), Mishra (2007)), and corruption (Daude and Fratzscher (2008), Coeurdacier and Martin (2007), De Santis and Gérard (2006)).

Of particular interest are controls that, like my communication cost measure, capture some aspect of information frictions. Information frictions are a well-defined determinant of portfolio investment: prior analyses have used geographical distance (Aggarwal, Kearney, and Lucey (2012), Daude and Fratzscher (2008), Lane and Milesi-Ferretti (2008), etc.), cultural distance (Aggarwal et al. (2012)), and common languages (Aggarwal et al. (2012), Daude and Fratzscher (2008), Lane and Milesi-Ferretti (2008), Ferreira and Miguel (2007), Mishra (2007), Coeurdacier and Martin (2007), Faruqee and Yan (2004)). My measure of bilateral communication cost represents a new dimension of information frictions not precisely like these or the other variables in use in the literature: neither these, nor any of the other papers compiled by Roque and Cortez (2014) attempt to use direct measurements or proxies for communication cost by Internet or other medium.

Table 23. Summary Statistics: Information Friction Variables

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
tauCostEst	7,320	3.986	0.367	2.233	3.865	4.031	5.926
languageDiff	9,660	0.857	0.227	0.007	0.822	0.998	1.000
cultureDist	4,140	0.299	0.092	0.035	0.238	0.368	0.534

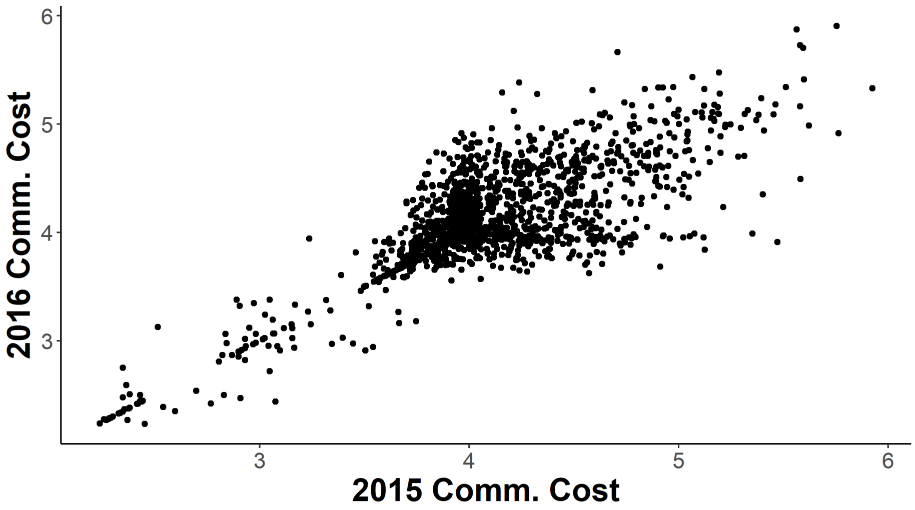
4.2.3 Gravity. The models I estimate employ a gravity framework, which has been applied extensively in the finance literature (Anderson and van Wincoop (2003), Okawa and van Wincoop (2012)). Several applications in this literature (Aggarwal et al. (2012), Karolyi (2016)) focus on the "gravity of culture," using cultural differences as resistance terms in a gravity model of portfolio investment (although in fact Aggarwal et al. (2012) finds that certain cultural differences act as attractors, not resistors, of investment).

4.3 Data

At the center of this analysis are the estimated communication costs taken from my previous chapter. These are bilateral values representing an "iceberg communication cost" analogous to the iceberg trade costs already in wide use in the literature, estimated using data on Internet communication and routing recorded in the United States in 2015-2016. While in theory these costs accurately measure the costs of communication between pairs of countries around the world, in practice they are most reliably estimated for pairs including the US. Summary statistics for this measure are reported in Table 23. A plot of communication costs in 2015 vs. 2016 is shown in Figure 13.

Data on portfolio investment volumes is taken from the European Commission's Finflows database (Nardo et al. (2017)). This data comprises bilateral portfolio investment flows among the EU countries and their trading

Figure 13. Year-to-Year Comparison of Estimated Communication Costs



partners, comprising 83 total countries and covering the period from 2001 onwards, although I restrict the analysis to the period 2015-2016, when my communication costs are most reliably estimated.

A second data source is the US Treasury's International Capital Data, consisting of monthly observations of portfolio transactions between US and foreign citizens (US Treasury (2022)). This dataset is disaggregated into different categories of asset, including US treasury bonds, federal agency bonds, US corporate stock and bonds, and foreign stock and bonds. However, because all observed flows are either to or from the US, this data does not contain any true bilateral variation, and additionally, it is a much smaller dataset. The advantage it offers is the ability to check for heterogeneity in how different asset categories are affected by communication costs, along with the fact that my communication costs (estimated using US-based data) are more reliably estimated for country-pairs including the US.

For control variables, I take data from a variety of sources. To control for a country's regulatory environment, I use Transparency International's Corruption Perceptions Index, an index of how corrupt a country's public sector is perceived to be. I also employ the capital controls dataset developed by Fernández, Klein, Rebucci, Schindler, and Uribe (2016), specifically the measures of inbound and outbound capital restrictions. Both data sources are varying at the country-year level.

To control for variation in general economic conditions, I take several variables from the Penn World Table (Zeileis (2021)), specifically real GDP, population, human and physical capital stocks, and prevailing rates of return, all varying at the country-year level.

Finally, to control for cultural differences, I use a combination of the US International Trade Commission's Domestic and International Common Language Database, and Hofstede's cultural dimensions. This latter dataset describes national culture along six axes, termed power distance (degree to which high- and low-power individuals are separated), individualism, masculinity, uncertainty avoidance, long term orientation, and indulgence (degree to which individuals are expected to seek their own interest rather than their organization's). The six indices are time-invariant and only vary by country, and the common-language data also varies only at the country-pair level. I also generate a country-pair varying "cultural distance" variable by computing the norm of the difference between a country-pair's cultural dimensions, $|Hofstede_i - Hofstede_j|$, for use alongside country-time fixed effects.

Summary statistics for the information-friction variables are reported in Table 23 alongside estimated communication cost. As the two origin-destination varying measures (language difference and cultural distance norm) capture elements of information friction, there is naturally some degree of correlation among them and the estimated communication cost at the center of this paper. Correlation coefficients are 0.421, 0.499, and 0.496, for log-cost vs. language difference, log-cost vs. culture distance, and language difference vs. culture distance, respectively.

4.4 Models and Results

In this section, as I discuss my models, I will in general proceed from a very naive model, containing only the explanatory variable(s), through the addition of control variables to the model, and finally to the introduction of fixed effects of increasing detail, which will replace some or all of the controls.

4.4.1 Finflows Data.

I begin my analysis with the Finflows data, which due to its broader geographic scope would seem the more broadly applicable source of data on cross-border financial transactions.

4.4.1.1 Naive Specifications.

I naturally begin with the estimation of an extremely naive model,

$$log(Inv_{ijt}) = \beta log(\tau_{ijt}) + \epsilon_{ijt} \tag{4.1}$$

, which I follow up with a trio of alternate specifications,

$$log(Inv_{ijt}) = \beta LanguageDiff_{ij} + \epsilon_{ijt} \tag{4.2}$$

$$log(Inv_{ijt}) = \beta CultureDiff_{ij} + \epsilon_{ijt} \tag{4.3}$$

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \epsilon_{ijt} \tag{4.4}$$

I include the specifications using only language and cultural differences as explanatory variables to examine the differences between these and my measure of

Table 24. Finflows Portfolio Investment: Effects of Information Frictions

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	3.764***			4.677***
	(0.966)			(1.042)
Language Difference		−1.606***		−1.538**
		(0.453)		(0.601)
Culture Difference			−3.921***	−2.576*
			(1.192)	(1.349)
Observations	236	236	180	180
R^2	0.061	0.051	0.057	0.159
Adjusted R^2	0.057	0.047	0.052	0.144
Residual Std. Error	1.853 (df = 234)	1.863 (df = 234)	1.706 (df = 178)	1.621 (df = 176)

*p<0.1; **p<0.05; ***p<0.01

communication costs, but as the communication costs vary at a more detailed level[1] it will, in the end, be the only one of these measures that will not be absorbed by the introduction of an origin-destination fixed effect. Results of these models are reported in Table 24.

Notably, communication costs have a positive, and highly significant, coefficient in the two models that include it. This is contrary to the conventional wisdom, that higher communication costs or information frictions in general should impede investment flows, and is the first sign that these communication costs may in part act as a proxy for some other quantity.

4.4.1.2 Introduction of Controls. I next introduce, step by step, a vector of controls to my model, now specified as

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \beta_3 X_{ijt} + \epsilon_{ijt} \tag{4.5}$$

X_{ijt} here represents a vector of controls, and I estimate this model multiple times using controls for the quality of governance and institutions, economic conditions,

[1] Origin-destination year, as opposed to only origin-destination in the case of language and culture differences.

cultural traits, and finally all three sets combined. Results of these models are reported in Table 25.

In the first three models of this form, the coefficient on log communication costs remains negative (although in one case non-significantly), as in the results from estimating Equation 4.4. It is not until all three sets of controls are used simultaneously that the more conventional result, of all three information-friction variables having significant and negative coefficients, is achieved.

4.4.1.3 Introduction of Fixed Effects.

I now introduce, in stages, a set of fixed effects to the model in order to eliminate the possibility of omitted variable bias. First, I incorporate simple origin, destination, and time fixed effects in the following specification:

$$\begin{aligned} log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \beta_3 X_{ijt} + \\ FE_i + FE_j + FE_t + \epsilon_{ijt} \end{aligned} \tag{4.6}$$

Because my cultural controls are non-time varying, and vary only at the country level, they are absorbed by FE_i and FE_j, and I thus omit them from this specification. The governance and economic controls are country-time-varying, and I retain them.

I follow this model by next using interacted fixed effects, in the specification

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + FE_{it} + FE_{jt} + \epsilon_{ijt} \tag{4.7}$$

All remaining controls being country-time-varying, they are now absorbed by the origin-time and destination-time fixed effects. These fixed effects are the most complete set I can include without also absorbing my measures of language and cultural difference, and so before adding the final, origin-destination fixed effect, I examine possible interactions between my measures of information friction, using

Table 25. Finflows Portfolio Investment: with Controls

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	4.292^{***}	0.554	3.614^{***}	-1.831^{***}
	(0.532)	(0.580)	(0.534)	(0.539)
Language Difference	-2.930^{***}	-3.047^{***}	-3.129^{***}	-2.021^{***}
	(0.335)	(0.333)	(0.345)	(0.317)
Culture Difference	-1.776^{**}	-1.955^{***}	-3.134^{***}	-2.403^{***}
	(0.701)	(0.660)	(0.702)	(0.646)
Origin Equity Restriction	-1.552^{***}			-0.835^{***}
	(0.179)			(0.190)
Destination Equity Restriction	0.411^{*}			0.974^{***}
	(0.238)			(0.234)
Origin Corruption	0.060^{***}			0.066^{***}
	(0.004)			(0.005)
Destination Corruption	0.038^{***}			0.043^{***}
	(0.004)			(0.006)
Origin GDP		-0.847^{***}		-0.791^{***}
		(0.089)		(0.095)
Origin Population		-0.003^{***}		-0.004^{***}
		(0.0004)		(0.0004)
Origin Human Capital		2.828^{***}		0.279
		(0.179)		(0.235)
Origin Physical Capital		0.303^{***}		0.308^{***}
		(0.024)		(0.025)
Origin Rates of Return		7.278^{***}		3.599^{**}
		(1.504)		(1.486)
Destination GDP		-0.450^{***}		-0.574^{***}
		(0.089)		(0.098)
Destination Population		-0.002^{***}		-0.003^{***}
		(0.0004)		(0.0004)
Destination Human Capital		1.032^{***}		-0.238
		(0.159)		(0.217)
Destination Physical Capital		0.195^{***}		0.239^{***}
		(0.025)		(0.027)
Destination Rates of Return		7.924^{***}		6.554^{***}
		(1.481)		(1.402)
Origin Power Distance			-0.024^{***}	-0.002
			(0.004)	(0.004)
Origin Individuality			0.021^{***}	-0.005
			(0.004)	(0.003)
Origin Masculinity			0.004	0.003
			(0.003)	(0.003)
Origin Uncertainty Aversion			-0.018^{***}	-0.012^{***}
			(0.003)	(0.003)
Origin Long Term Orientation			0.045^{***}	0.008^{**}
			(0.003)	(0.004)
Origin Indulgence			0.034^{***}	0.004
			(0.004)	(0.004)
Destination Power Distance			-0.0004	0.006
			(0.004)	(0.004)
Destination Individuality			0.024^{***}	0.008^{**}
			(0.004)	(0.003)
Destination Masculinity			0.008^{***}	0.009^{***}
			(0.003)	(0.003)
Destination Uncertainty Aversion			-0.011^{***}	-0.012^{***}
			(0.003)	(0.003)
Destination Long Term Orientation			0.026^{***}	0.012^{***}
			(0.003)	(0.004)
Destination Indulgence			0.029^{***}	0.022^{***}
			(0.004)	(0.003)
Constant	-4.750^{***}	-8.982^{***}	-4.246^{***}	-2.315^{*}
	(0.813)	(1.115)	(1.017)	(1.371)
Controls	Governance	Economic	Cultural	All
Observations	2,079	2,559	2,559	2,079
R^2	0.339	0.313	0.288	0.531
Adjusted R^2	0.336	0.309	0.284	0.524
Residual Std. Error	2.741 (df = 2071)	2.838 (df = 2545)	2.890 (df = 2543)	2.321 (df = 2049)

*p<0.1; **p<0.05; ***p<0.01

the specifications

$$\begin{aligned} log(Inv_{ijt}) =& \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 log(\tau_{ijt}) \times LanguageDiff_{ij} + \\ & \beta_3 CultureDiff_{ij} + FE_{it} + FE_{jt} + \epsilon_{ijt} \end{aligned} \tag{4.8}$$

$$\begin{aligned} log(Inv_{ijt}) =& \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \\ & \beta_3 log(\tau_{ijt}) \times CultureDiff_{ij} + FE_{it} + FE_{jt} + \epsilon_{ijt} \end{aligned} \tag{4.9}$$

Estimation results for Equations 4.6 through 4.9 are reported as the first four models of Table 26.

Table 26. Finflows Portfolio Investment: with Fixed Effects Replacing Controls

	Dependent variable:				
	Log Investment				
	(1)	(2)	(3)	(4)	(5)
Log Comm. Cost	-2.960^{***}	-3.308^{***}	-2.720^{**}	-6.612^{***}	-0.350
	(0.548)	(0.534)	(1.196)	(1.010)	(2.350)
Language Difference	-2.136^{***}	-2.529^{***}	-1.335	-2.542^{***}	
	(0.307)	(0.289)	(2.191)	(0.288)	
Culture Difference	-2.386^{***}	-1.706^{***}	-1.678^{***}	-21.900^{***}	
	(0.669)	(0.624)	(0.626)	(5.281)	
Log Comm. Cost × Language Difference			-0.882		
			(1.605)		
Log Comm. Cost × Culture Difference				14.619^{***}	
				(3.797)	
Fixed Effects	i, j, t	it, jt	it, jt	it, jt	it, jt, ij
Controls	Country-Time	None	None	None	None
Observations	2,079	2,559	2,559	2,559	4,556
R^2	0.659	0.673	0.673	0.675	0.916
Adjusted R^2	0.640	0.645	0.645	0.647	0.635
Residual Std. Error	2.018 (df = 1972)	2.035 (df = 2354)	2.035 (df = 2353)	2.029 (df = 2353)	2.241 (df = 1053)

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

In these models, the coefficients on the three information-friction variables all remain negative and for the most part[2] significant. However, the introduction of interactions produces an interesting result, in that the interaction between log communication costs and cultural differences has a positive and significant coefficient. This would suggest that the negative effect of communication costs is decreased for pairs of countries with significant cultural differences, a result that does not have an immediately obvious explanation. The simplest explanation, though it is one that cannot readily be tested with this data, is a selection effect. It may be that cultural differences drive foreign investors to select only the most transparent assets in a country, thus making extensive communication less necessary and by extension, making communication costs a less salient factor in the decision to invest. (An alternate phrasing of this possibility is that communication costs drive investors to select only the investments for which cultural differences are least salient.)

Finally, I estimate a model with a complete set of fixed effects that absorb all explanatory variables save for log communication costs,

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + FE_{it} + FE_{jt} + FE_{ij} + \epsilon_{ijt} \tag{4.10}$$

Results are reported as the last model of Table 26, but are disappointing: in conjunction with the origin-destination fixed effect, log communication cost has little significance, although it does have the negative coefficient that would be expected of an information friction. This is a stronger indication that my communication cost measure is in truth acting as a proxy for other origin-destination varying factors.

[2] The exception being in the model interacting communication costs with language differences, in which the coefficients are still negative but not all significant.

4.4.2 Likelihood of Investment Activity.

I also estimate a set of logit models to determine if the effect of communication costs on the likelihood of cross-border investment flows may be separate from its effect on the size of flows:

$$log\left(\frac{p(Act_{ijt})}{1-p(Act_{ijt})}\right) = \beta_0 log(\tau_{ijt}) + FE_{it} + FE_{jt} + \epsilon_{ijt} \tag{4.11}$$

$$\begin{aligned} log\left(\frac{p(Act_{ijt})}{1-p(Act_{ijt})}\right) &= \beta_0 log(\tau_{ijt}) + \beta_1 languageDiff_{ij} + \\ &\quad \beta_2 log(\tau_{ijt}) \times languageDiff_{ij} + \beta_3 cultureDiff_{ij} + \epsilon_{ijt} \end{aligned} \tag{4.12}$$

$$\begin{aligned} log\left(\frac{p(Act_{ijt})}{1-p(Act_{ijt})}\right) &= \beta_0 log(\tau_{ijt}) + \beta_1 cultureDiff_{ij} + \\ &\quad \beta_2 log(\tau_{ijt}) \times cultureDiff_{ij} + \beta_3 languageDiffij + \epsilon_{ijt} \end{aligned} \tag{4.13}$$

I omit the origin-destination fixed effects in these specifications for computational reasons.[3] Results are reported in Table 27.

Comparing the results of this estimation to those of previous models using the same origin-time and destination-time fixed effects, it becomes apparent that the effect of communication costs (or whatever origin-destination varying variable it proxies for) on the likelihood of investment is opposite its effect on investment value. On the surface, with increased communication costs leading to more-likely but less-valuable investment, it is possible to suggest some kind of crowding-out, in which low communication costs attract larger investors, making less frequent investments, but driving smaller investors out of the market. However, in the models with interactions between communication costs and other information frictions, the average marginal effect of communication cost is in fact negative

[3] In previous models with origin-time, destination-time, and origin-destination fixed effects, I was able to use the *felm* estimator in R, which is highly efficient at dealing with such numerous effects. *felm* does not, however, permit estimation of logit models, and the addition of a third family of fixed effects raises the complexity of the model to a point that R cannot estimate it within a reasonable timeframe.

Table 27. Likelihood of Finflows Portfolio Investment

	Dependent variable:			
	activity			
	(1)	(2)	(3)	(4)
log(tauCostEst)	4.979***	11.785***	22.984***	18.401***
	(0.245)	(0.651)	(1.565)	(1.247)
languageDiff		−6.278***	29.564***	−6.230***
		(0.646)	(4.056)	(0.736)
cultureDist		−3.977***	−4.836***	124.604***
		(1.194)	(1.383)	(11.259)
log(tauCostEst):languageDiff			−28.290***	
			(3.038)	
log(tauCostEst):cultureDist				−95.550***
				(8.264)
Fixed Effects	it, jt	it, jt	it, jt	it, jt
Observations	11,552	5,408	5,408	5,408
Log Likelihood	−2,625.123	−753.331	−646.632	−614.401
Akaike Inf. Crit.	5,856.246	1,924.663	1,713.265	1,648.802

*p<0.1; **p<0.05; ***p<0.01

Table 28. TIC Portfolio Investment: Effects of Information Frictions

	Dependent variable:			
	log(abs(flow))			
	(1)	(2)	(3)	(4)
log(tauCostEst)	3.764^{***}			4.677^{***}
	(0.966)			(1.042)
languageDiff		-1.606^{***}		-1.538^{**}
		(0.453)		(0.601)
cultureDist			-3.921^{***}	-2.576^{*}
			(1.192)	(1.349)
Observations	236	236	180	180
R^2	0.061	0.051	0.057	0.159
Adjusted R^2	0.057	0.047	0.052	0.144
Residual Std. Error	1.853 (df = 234)	1.863 (df = 234)	1.706 (df = 178)	1.621 (df = 176)

*p<0.1; **p<0.05; ***p<0.01

(-0.027 in Model 3 and -0.085 in Model 4), indicative of the more plausible case where higher communication costs lead to smaller and less frequent investments.

4.4.3 Robustness Check: TIC Data.

As a robustness check, I next re-estimate Equations 4.1 through 4.5, using the Treasury International Capital data as a substitute for the Finflows data. Results are reported, as before, in Tables 28 and 29. The results are broadly the same, though with reduced significance due to the much smaller sample size available in the TIC data.

Table 29. TIC Portfolio Investment: with Controls

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	4.002***	0.821	−0.186	−0.081
	(1.085)	(0.975)	(1.077)	(1.002)
Language Difference	−2.233***	−1.980***	−2.501***	−0.976
	(0.624)	(0.535)	(0.608)	(0.689)
Culture Difference	0.505	−1.163	−6.830**	−10.013***
	(1.588)	(1.218)	(3.159)	(3.158)
Origin Equity Restriction	−0.716			−0.817*
	(0.500)			(0.493)
Destination Equity Restriction	−0.535			−0.478
	(0.651)			(0.593)
Origin Corruption	0.009			0.015
	(0.012)			(0.013)
Destination Corruption	0.016			0.020
	(0.013)			(0.014)
Origin GDP		−0.335		−0.696*
		(0.319)		(0.372)
Origin Population		−0.0005		−0.0001
		(0.001)		(0.001)
Origin Human Capital		1.859***		1.660***
		(0.433)		(0.615)
Origin Physical Capital		0.158**		0.241***
		(0.071)		(0.085)
Origin Rates of Return		1.503		7.013*
		(3.023)		(3.673)
Destination GDP		−0.339		−0.624*
		(0.317)		(0.367)
Destination Population		−0.001		−0.001
		(0.001)		(0.001)
Destination Human Capital		1.766***		1.218**
		(0.430)		(0.614)
Destination Physical Capital		0.160**		0.224***
		(0.070)		(0.083)
Destination Rates of Return		1.870		6.972*
		(3.024)		(3.703)
Origin Power Distance			0.020**	0.031***
			(0.010)	(0.011)
Origin Individuality			−0.0005	−0.033**
			(0.012)	(0.013)
Origin Masculinity			−0.0005	0.001
			(0.008)	(0.008)
Origin Uncertainty Aversion			−0.011*	−0.006
			(0.006)	(0.007)
Origin Long Term Orientation			0.058***	0.019
			(0.009)	(0.012)
Origin Indulgence			0.008	0.019*
			(0.010)	(0.010)
Destination Power Distance			0.024**	0.037***
			(0.010)	(0.011)
Destination Individuality			0.002	−0.023*
			(0.012)	(0.013)
Destination Masculinity			−0.004	−0.004
			(0.008)	(0.008)
Destination Uncertainty Aversion			−0.009	−0.005
			(0.006)	(0.007)
Destination Long Term Orientation			0.059***	0.025**
			(0.009)	(0.012)
Destination Indulgence			0.007	0.017*
			(0.010)	(0.009)
Constant	2.732	−7.709***	7.037**	−5.977
	(2.031)	(2.674)	(3.001)	(4.248)
Controls	Governance	Economic	Cultural	All
Observations	168	180	180	168
R^2	0.243	0.466	0.479	0.646
Adjusted R^2	0.210	0.424	0.431	0.572
Residual Std. Error	1.557 (df = 160)	1.330 (df = 166)	1.322 (df = 164)	1.146 (df = 138)

*p<0.1; **p<0.05; ***p<0.01

Table 30. TIC Portfolio Investment: with Fixed Effects Replacing Controls

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	-0.111	-5.597^{**}	-0.582	-0.403
	(1.008)	(2.438)	(2.094)	(0.802)
Language Difference	-0.968	-17.596^{**}	-0.987	
	(0.692)	(6.785)	(0.698)	
Culture Difference	-10.025^{***}	-8.248^{**}	-12.638	
	(3.168)	(3.194)	(10.654)	
Log Comm. Cost × Language Difference		10.362^{**}		
		(4.207)		
Log Comm. Cost × Culture Difference			1.869	
			(7.272)	
Fixed Effects	t	t	t	ij,
Controls	All	All	All	Country-Time
Observations	168	168	168	200
R^2	0.647	0.662	0.647	0.986
Adjusted R^2	0.569	0.585	0.566	0.967
Residual Std. Error	1.150 (df = 137)	1.129 (df = 136)	1.154 (df = 136)	0.338 (df = 85)

*p<0.1; **p<0.05; ***p<0.01

I also estimate analogues to the models using fixed effects in Equations 4.6 through 4.10, using the closest degrees of fixed effects possible:

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \beta_3 X_{ijt} + FE_t + \epsilon_{ijt} \tag{4.14}$$

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 log(\tau_{ijt}) \times LanguageDiff_{ij} + \beta_3 CultureDiff_{ij} + FE_t + \epsilon_{ijt} \tag{4.15}$$

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + \beta_1 LanguageDiff_{ij} + \beta_2 CultureDiff_{ij} + \beta_3 log(\tau_{ijt}) \times CultureDiff_{ij} + FE_t + \epsilon_{ijt} \tag{4.16}$$

$$log(Inv_{ijt}) = \beta_0 log(\tau_{ijt}) + FE_{ij} + FE_t + \epsilon_{ijt} \tag{4.17}$$

Because the TIC data contains only US outflows and inflows, it is difficult to use origin or destination fixed effects, due to multicollinearity concerns. I instead begin with a simple time fixed effect, then proceed directly to introduce interactions among the information friction variables, and finally introduce the origin-destination fixed effect that absorbs all but log communication costs.

Results are broadly the same as those using Finflows data: the pattern of negative coefficients on information frictions and positive coefficients on interactions remains, although intriguingly, I note greater significance in the model interacting communications cost with language differences than the one interacting with culture differences. However, this seems most likely to be a matter of selection bias (as the TIC data contains flows to different trading partners than the Finflows data). Again, disappointingly, the introduction of an origin-destination fixed effect absorbs the significance of log communication costs.

4.4.4 Heterogeneity by Inflows vs. Outflows.

I next examine heterogeneity in the effects of communications cost on outflows vs. inflows of investment from and to the US. I begin by restricting the Finflows dataset to outflows of investment from the US and estimating the models

$$log(abs(Inv_{jt})) = \beta_0 log(\tau_{jt}) + \beta_1 X_{jt} + FE_t + \epsilon_{jt} \tag{4.18}$$

$$\begin{aligned} log(abs(Inv_{jt})) = {} & \beta_0 log(\tau_{jt}) + \beta_1 cultureDist_j + \\ & \beta_2 log(\tau_{jt}) \times cultureDist_j + +\beta_3 X_{jt} + FE_t + \epsilon_{jt} \end{aligned} \tag{4.19}$$

$$\begin{aligned} log(abs(Inv_{jt})) = {} & \beta_0 log(\tau_{jt}) + \beta_1 languageDiff_j + \\ & \beta_2 log(\tau_{jt}) \times languageDiff_j + +\beta_3 X_{jt} + FE_t + \epsilon_{jt} \end{aligned} \tag{4.20}$$

$$log(abs(Inv_{jt})) = \beta_0 log(\tau_{jt}) + FE_j + FE_t + \epsilon_{jt} \tag{4.21}$$

Results are reported in Table 31, showing very little significance, but coefficients broadly consistent in sign and magnitude with the previous sets of results. Lack of significance may easily be attributed to the sample size, which is roughly two orders of magnitude smaller.

I next estimate similar models (replacing j subscripts with i) on a dataset of investment inflows to the US, reported in Table 32. There is again little significance to communications cost in the model using origin and time fixed effects, but the models using controls produce significant coefficients consistent with prior results.

Table 31. Regression Results: Finflows Data, US Outflows Only

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	-2.436	-7.496	-6.605	-1.846
	(3.307)	(8.603)	(6.486)	(7.711)
Language Difference	-3.527^{*}	-16.658	-3.441	
	(2.009)	(20.675)	(2.027)	
Culture Difference	5.614	7.265	-18.083	
	(8.874)	(9.329)	(32.871)	
Log Comm. Cost × Language Difference		8.380		
		(13.132)		
Log Comm. Cost × Culture Difference			16.089	
			(21.476)	
Control Variables	None	All	All	None
Fixed Effects	t	t	t	j, t
Observations	48	48	48	80
R^2	0.774	0.777	0.778	0.882
Adjusted R^2	0.646	0.638	0.640	0.508
Residual Std. Error	1.607 (df = 30)	1.623 (df = 29)	1.618 (df = 29)	1.922 (df = 19)

*p<0.1; **p<0.05; ***p<0.01

Table 32. Regression Results: Finflows Data, US Inflows Only

	Dependent variable:			
	Log Investment			
	(1)	(2)	(3)	(4)
Log Comm. Cost	8.180^{***}	-12.073^{**}	-5.293	-2.188
	(2.065)	(4.598)	(4.098)	(3.269)
Language Difference	-1.993	-37.389^{***}	-0.398	
	(1.307)	(11.481)	(1.573)	
Culture Difference	-4.914^{*}	1.563	-36.680^{*}	
	(2.851)	(6.229)	(20.157)	
Log Comm. Cost × Language Difference		23.946^{***}		
		(7.362)		
Log Comm. Cost × Culture Difference			26.202^{*}	
			(13.656)	
Control Variables	None	All	All	None
Fixed Effects	t	t	t	i, t
Observations	75	67	67	94
R^2	0.255	0.718	0.680	0.923
Adjusted R^2	0.213	0.612	0.560	0.775
Residual Std. Error	2.153 (df = 70)	1.443 (df = 48)	1.536 (df = 48)	1.118 (df = 32)

*p<0.1; **p<0.05; ***p<0.01

Finally, I estimate a trio of models combining both one-way datasets and employing interactions with a direction indicator variable:

$$\begin{aligned} log(abs(Inv_{pdt})) =& \beta_0 log(\tau_{pdt}) + \beta_1 Direction_d + \beta_2 log(\tau_{pdt}) \times Direction_d + \\ & \beta_3 LanguageDiff_p + \beta_4 CultureDiff_p + \beta_5 X_{jt} + \epsilon_{pt} \end{aligned} \tag{4.22}$$

$$\begin{aligned} log(abs(Inv_{pdt})) =& \beta_0 log(\tau_{pdt}) + \beta_1 Direction_d + \beta_2 log(\tau_{pdt}) \times Direction_d + \\ & \beta_3 LanguageDiff_p + \beta_4 CultureDiff_p + \beta_5 X_{jt} + FE_t + \epsilon_{pt} \end{aligned} \tag{4.23}$$

$$\begin{aligned} log(abs(Inv_{pdt})) =& \beta_0 log(\tau_{pdt}) + \beta_1 Direction_d + \beta_2 log(\tau_{pdt}) \times Direction_d + \\ & FE_p + FE_t + \epsilon_{pt} \end{aligned} \tag{4.24}$$

Here, p indexes partner countries, and d indexes direction (into or out of the US). Results are reported in Table 33, and show no significance to either the direction indicator or its interaction with communication costs, implying a lack of differential impact on incoming vs. outgoing investment.

4.4.5 Heterogeneity by Asset Type. I next turn this analysis to examine the possibility of heterogeneous effects by asset category, using the TIC data. The TIC data classifies assets into six groups: marketable US Treasury and Federal Financing Bank bonds and notes, government corporation and federal agency bonds, US corporate and other bonds, US corporate stock, foreign bonds[4], and foreign stock. I first re-estimate the models specified in Equations 4.10 through 4.9, substituting as the dependent variable separate measures of total purchases of US and foreign assets. Results are reported in Table 34.

As can be seen here, there are minimal differences between the coefficients estimated for foreign vs. US assets, and little difference from the more aggregated model. Coefficients are broadly similar in sign, magnitude, and significance, with

[4] Including both public and private bonds.

Table 33. Regression Results: Finflows Data, Interaction with Direction

	Dependent variable:		
	Log Investment		
	(1)	(2)	(3)
Log Comm. Cost	3.357	3.451*	2.265
	(2.029)	(2.008)	(3.098)
Outflow	2.668	3.450	−2.877
	(4.697)	(4.671)	(3.537)
Log Comm. Cost × Outflow	−1.853	−2.372	1.462
	(2.969)	(2.953)	(2.296)
Language Difference	−1.818	−1.811	
	(1.227)	(1.215)	
Culture Difference	0.115	0.663	
	(5.455)	(5.408)	
Control Variables	All	All	None
Fixed Effects	None	t	p, t
Observations	115	115	174
R^2	0.594	0.606	0.784
Adjusted R^2	0.518	0.528	0.622
Residual Std. Error	1.728 (df = 96)	1.710 (df = 95)	1.577 (df = 99)

*p<0.1; **p<0.05; ***p<0.01

two exceptions: a greater effect of language difference on purchases of foreign assets, and a similar greater effect of cultural difference on purchases of US assets. In both cases, the difference is sufficient to make the effect significant, but does not noticeably extend to the corresponding interaction terms.

Further disaggregating the TIC data, I next re-estimate the same models separately for each of the six categories of asset, with results for language interaction reported in Table 35 and for culture interaction in Table 36.

As before, the model interacting communications cost with language difference produces more significance, and so I shall focus the comparisons here. Estimated coefficients are again broadly similar in sign across the six asset categories, but there is variation in magnitude and significance: most notably, none of these information-friction variables has any significance in the model using US corporate stock investment as the dependent variable, and only log communication cost is significant for US corporate bonds. This could be interpreted as due to the highly diverse nature of these categories, except that there is far more significance in the equally, if not more, diverse foreign bond and stock categories.

Additionally, the positive coefficient on the interaction term is significant only for federal marketable bonds, foreign bonds, and foreign stock. While it remains positive for the other three categories, it is not significant even at the 10% level, indicating a noisier relationship with investment in these categories.

Table 34. Regression Results: Foreign vs. US Assets

	Dependent variable:					
	Log Investment					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Comm. Cost	0.092	−0.374	−8.131***	−5.788**	−1.841	−0.365
	(1.047)	(1.034)	(2.557)	(2.754)	(2.239)	(2.379)
Language Difference			−25.177***	−19.956***	−2.219***	−0.687
			(7.086)	(7.631)	(0.744)	(0.791)
Culture Difference			−5.022	−8.881**	−14.833	−16.308
			(3.363)	(3.622)	(11.395)	(12.107)
Log Comm. Cost × Language Difference			14.354***	12.044**		
			(4.398)	(4.736)		
Log Comm. Cost × Culture Difference					5.277	3.852
					(7.777)	(8.263)
Asset Types	Total Foreign	Total US	Total Foreign	Total US	Total Foreign	Total US
Fixed Effects	ij, t	ij, t	t	t	t	t
Controls	Country-Time	Country-Time	All	All	All	All
Observations	200	200	168	168	168	168
R^2	0.978	0.979	0.645	0.615	0.619	0.598
Adjusted R^2	0.950	0.950	0.567	0.531	0.535	0.510
Residual Std. Error	0.441 (df = 85)	0.435 (df = 85)	1.190 (df = 137)	1.282 (df = 137)	1.234 (df = 137)	1.311 (df = 137)

*p<0.1; **p<0.05; ***p<0.01

Table 35. Regression Results: All Asset Categories, Language Difference Interaction

	Dependent variable:					
	Log Investment					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Comm. Cost	-6.404^{*}	-9.530^{*}	-6.116^{*}	-3.220	-8.691^{***}	-6.831^{**}
	(3.499)	(4.917)	(3.603)	(2.845)	(2.725)	(2.829)
Language Difference	-23.677^{**}	-23.066	-16.125	-12.543	-27.172^{***}	-20.298^{**}
	(9.704)	(14.539)	(10.322)	(7.884)	(7.612)	(7.839)
Culture Difference	-11.833^{**}	-24.251^{***}	-6.206	-1.173	-5.516	-8.256^{**}
	(4.624)	(7.887)	(4.885)	(3.742)	(3.557)	(3.721)
Log Comm. Cost × Language Difference	14.537^{**}	13.267	8.695	6.555	15.878^{***}	11.681^{**}
	(6.020)	(8.780)	(6.353)	(4.893)	(4.712)	(4.865)
Asset Types	Federal Marketable Bond	Fed. Agency Bond	US Corp. Bond	US Corp. Stock	Foreign Bond	Foreign Stock
Fixed Effects	t	t	t	t	t	t
Controls	All	All	All	All	All	All
Observations	167	134	155	168	166	168
R^2	0.519	0.615	0.629	0.658	0.639	0.611
Adjusted R^2	0.413	0.502	0.540	0.583	0.559	0.525
Residual Std. Error	1.627 (df = 136)	2.047 (df = 103)	1.603 (df = 124)	1.324 (df = 137)	1.248 (df = 135)	1.317 (df = 137)

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 36. Regression Results: All Asset Categories, Culture Difference Interaction

	Dependent variable:					
	Log Investment					
	(1)	(2)	(3)	(4)	(5)	(6)
Log Comm. Cost	-0.164	2.104	-0.990	0.562	-1.926	-1.312
	(3.017)	(3.912)	(2.962)	(2.419)	(2.358)	(2.439)
Language Difference	-22.525	-3.074	-4.736	-0.609	-17.497	-14.026
	(15.342)	(19.468)	(15.135)	(12.311)	(12.005)	(12.412)
Culture Difference	-0.424	-0.695	-2.036^{*}	-2.023^{**}	-1.724^{**}	-1.600^{*}
	(1.005)	(1.825)	(1.072)	(0.804)	(0.790)	(0.810)
Log Comm. Cost × Language Difference	5.858	-19.245	-2.435	-1.197	6.527	2.711
	(10.489)	(13.709)	(10.404)	(8.402)	(8.217)	(8.471)
Asset Types	Federal Marketable Bond	Fed. Agency Bond	US Corp. Bond	US Corp. Stock	Foreign Bond	Foreign Stock
Fixed Effects	t	t	t	t	t	t
Controls	All	All	All	All	All	All
Observations	167	134	155	168	166	168
R^2	0.500	0.614	0.624	0.653	0.611	0.594
Adjusted R^2	0.390	0.501	0.533	0.577	0.524	0.506
Residual Std. Error	1.660 (df = 136)	2.050 (df = 103)	1.615 (df = 124)	1.333 (df = 137)	1.296 (df = 135)	1.344 (df = 137)

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

4.5 Conclusion

Based on the results of models with exhaustive fixed effects, it does not appear that communications costs have any significant effects on cross-border portfolio investment, as they appear to capture an aspect of information frictions that varies with origin and destination, but not time. It may be that this is due to the relatively short timeframe in which the necessary data is available to estimate the communication costs, and the small amount of temporal variation in communication costs. However, even if communication costs truly have no effect on portfolio investment, they may still find use in contexts that do not allow for the use of the full set of fixed effects. Employed alongside other information-friction controls, communication costs often have a significant and negative effect on portfolio investment, in keeping with their interpretation as a barrier to investment. Interactions between communication costs and other information frictions show an intriguing pattern that suggests communication costs may be less important between countries with greater linguistic or cultural differences, a pattern that bears further investigation.

CHAPTER V

CONCLUSION

As demonstrated here, technical data on Internet communication represents a previously untapped source of data that allows economists to examine an aspect of information frictions that previously was difficult to capture. This data presents a low-cost and high-frequency option to measure the degree of Internet availability within a region, as well as a simple measure of the quality of that access. These measures perform similarly to prior measures used in the literature, as seen in my adaptations of Freund and Weinhold (2004) and Allen (2014).

More importantly, however, this data can be employed in more sophisticated models to extract more detailed measures of Internet communication cost. My adaptation of Allen and Arkolakis (2019)'s structural model is only one potential application in this vein, and the value of such models is demonstrated by the ability of the estimated costs to explain the patterns of heterogeneity first seen in Keller and Yeaple (2013).

Finally, while these estimated communication costs do not provide any significance in models of cross-border financial flows when employed alongside full complements of fixed effects, they do possess significance and explanatory power in contexts where it is not possible to use the origin-destination fixed effect.

As Internet communication will almost certainly continue to gain importance in the future, the ability to measure the cost and quality of said communication will only become more crucial in international economics. These data and the measures derived from it represent an important step forward in that regard.

APPENDIX A

DATA DESCRIPTIONS

A.1 Raw Routing Data

A small excerpt of relevant fields from the ORVP routing data is provided in Table A.1: the excerpted observations are five distinct routes that the Equinix Chicago facility could use to communicate with a block of devices located physically near Portland, Oregon.

Table A.1. Excerpt from Routing Data (Equinix Chicago, January 1, 2018, 12:00 AM)

N	IP Block	Route
155044	23.206.120.0/22	53828 6939 7922 33490
155045	23.206.120.0/22	23367 6461 7922 33490
155046	23.206.120.0/22	19653 3356 7922 33490
155047	23.206.120.0/22	293 6939 7922 33490
155048	23.206.120.0/22	19016 3257 7922 33490

Taking the first row of Table A.1 as an example, this observation describes a route which allows the collector, in this case the Equinix Chicago IXP, to send information to the 23.206.120.0/22 block of IP addresses. (This notation is a shorthand which is not necessary for the reader to understand; it refers to the block from 23.206.120.0 to 23.206.123.255, containing 1024 addresses total.) This route will, after leaving the device which collected this routing data, pass through the networks with identifying numbers 52828, 6939, 7922, and 33490. These four networks are CTS Telecom, Hurricane Electric, the Comcast network backbone, and Comcast's Portland/Spokane regional network. All four are US-based.

A.2 Raw Trace Data

A small excerpt of the relevant fields in the CAIDA trace data is provided in table A.2. The three relevant fields are the origin and destination IP addresses, which can be geolocated to determine the country of origin and destination of the observed flow, and the packet size, which measures the size of the flow in bytes.

Table A.2. Excerpt from Trace Data (Equinix Chicago, April 6, 2016, 1:00 PM UTC)

Origin IP Address	Destination IP Address	Packet Size
133.87.38.108	3.137.145.218	56
70.42.44.237	65.42.255.211	530
147.73.59.126	29.188.50.86	1474
161.69.48.219	161.69.45.5	1504
137.227.47.182	221.46.221.84	1504

A.2.1 Trace Data Anonymization.

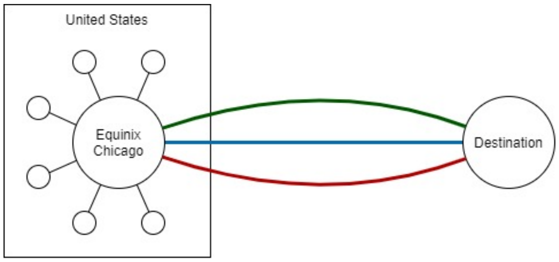
The anonymization referred to in the name of the Anonymized Internet Traces Dataset is a prefix-preserving anonymization algorithm, which slightly perturbs the recorded origin and destination IP addresses to preserve the privacy of the users whose communication is being described. This prevents identifying the exact users who sent or received the packets recorded, but, because the algorithm is prefix-preserving, allows the users' network to be correctly identified. To use an analogy, it is as though the addresses of respondents to a survey were obscured by altering each respondent's recorded address to a random, but still extant, address on the same street or in the same neighborhood. This is sufficiently accurate for the purposes of my model, as it is unnecessary to identify anything beneath the network level.

APPENDIX B

ILLUSTRATED EXAMPLES OF TRAFFIC COMPUTATION

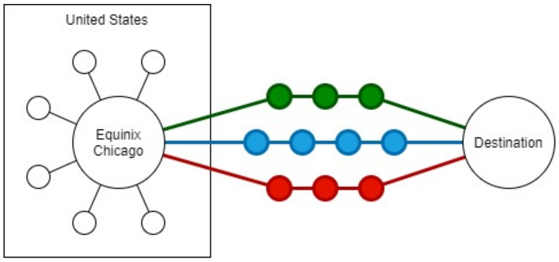
The figures in this section provide toy examples of how the process discussed in Section 3.3.5for computing a measure of link traffic, may be applied.

Figure B.1. Selection of Most Direct Routes



(a) All routes from IXP to an Arbitrary Destination

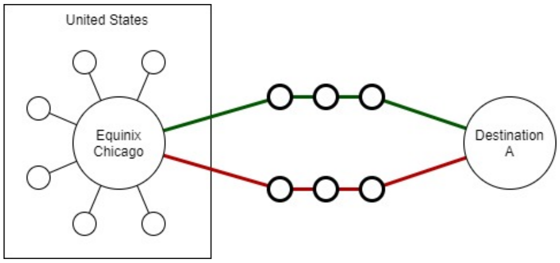
There exist three routes that Equinix Chicago can use to send communication to a particular destination.



(b) Breakdown of Routes into Links

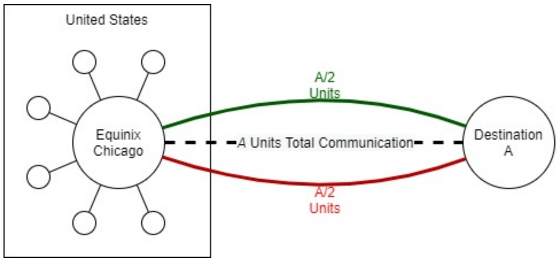
Of these three routes, the green and red routes pass through three intermediaries each (which can be interpreted as networks or countries depending on the scale of the application), while the blue route passes through four.

Figure B.2. Assignment of Communication and Traffic



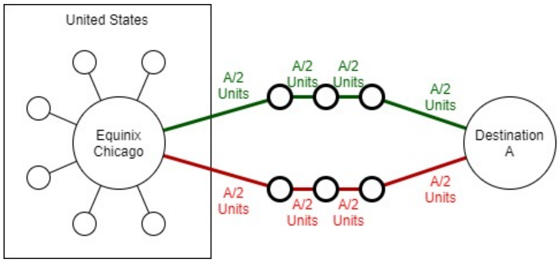
(a) Only Most Direct Routes

In the absence of other distinguishing characteristics, I discard the blue route, as it is less direct than the green or red routes (which are tied).



(b) Assignment of Communication Volumes to Routes

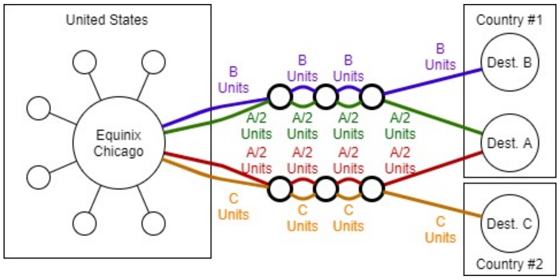
Here, the remaining most-direct routes are coupled with measures of communication, also from Chicago Equinix. There are A units of total communication observed being sent to destination A, and as there are two routes that this communication might take, this volume is divided evenly so that $A/2$ units are assigned to each of the two routes.



(c) Assignment of Traffic Volumes to Links

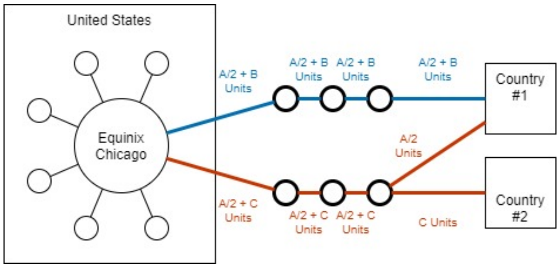
Each of these routes is composed of multiple links between networks. The entire volume of communication flows across each of these links, and so each of the links in the two most-direct routes receives $A/2$ units of traffic.

Figure B.3. Aggregation of Link Traffic



(a) Assignment of Traffic Volumes to Links

This figure shows a toy example that is more complex. Here, there are three distinct destinations, two of which belong to the same country. Communication and traffic have been assigned to routes and links as previously described; some links receive traffic from multiple routes.



(b) Summation of Link Traffic

Traffic across individual links is now summed to generate the measure of total traffic across each link. Destinations within each country are also combined so that the measure captures country-to-country traffic. For simplicity in this toy example, each intermediary represents a distinct country, but this is not necessary in practices.

APPENDIX C

ALTERNATE COMPUTATIONAL METHODOLOGIES

This appendix will describe alternate methods of computing certain measures used in my models.

C.1 Adjustments to Third-Party Communication

Because Equinix Chicago is located on the global Internet backbone, and therefore sees significant numbers of packets which neither originate from or are destined for Chicago and/or the US, it is not entirely unreasonable to treat this communication data as a valid measure of such third-party flows, especially when concerning flows among countries which are physically near the US. However, as discussed in Section 3.3.3.1, it is likely that the monitoring devices at Equinix Chicago do not capture a representative amount of the communication between, e.g., Germany and the Netherlands. Adding additional communication datasets gathered from collectors in different countries would alleviate this problem by providing additional locations from which first-party flows could be measured, but using multiple datasets would also allow for third-party flows to be predicted:

Consider, as an illustrative example, a situation in which there are two communication datasets, from collectors in the US and Canada. From the perspective of the US collector, communication from Canada to France is a third-party communication flow that would not be accurately represented in the US data. But this is a first-party flow which is presumably accurately measured in the Canada data! Using flows to and from Canada, and to and from the US, it becomes possible to create a regression model which predicts the flows which are truly third-party (not to or from either Canada or the US) based on the doubly-observed communication flows. An example for this simple, two-dataset situation is as follows:

$$Comm_{ijt}^{FP} = \beta_0 Comm_{ijt}^{US} \times USTP_{ijt} + \beta_1 Comm_{ijt}^{CA} \times CATP_{ijt} + FE_i + FE_j + FE_t + \epsilon_{ijt} \tag{C.1}$$

Here, $Comm_{ijt}^{FP}$ is the communication observed along link ij at time t by a first-party collector, $Comm_{ijt}^{US}$ and $Comm_{ijt}^{CA}$ are the same flows as measured by the US and Canada collectors, respectively, and $USTP_{ijt}$ and $CATP_{ijt}$ are indicator variables which take the value 1 if link ij is third-party from the perspective of the US or Canada, respectively. This allows the model to be estimated using flows which are first-party to the US but third-party to Canada, and vice versa; the model can then be used to predict truly third-party flows from the imperfect measurements in one or the other of the datasets. Extensions of this approach would include the use of an averaging mechanism, so that the predictions can use the information contained in both datasets, additional covariates, and even expansion to make use of three or more datasets.

C.2 Computationally-Intensive Construction of Link Traffic

The following method of computing link traffic is more detailed, as it associates each IP address observed to receive communication with a set of routes that reach it. However, this increased level of detail makes the procedure impractically complex: using my computational resources, it took weeks instead of hours to process a dataset, when it did not crash due to lack of available memory.

Instead of aggregating to the IP-block level, I instead work at the level of individual IP addresses η. Let $IPComm_{c\eta}$ be the volume of communication observed from the IXP to η, $R_{c\eta}$ the set of routes the IXP would use to reach it, and $R^{min}_{c\eta}$ the most-direct of those routes. Then, define

$$RouteComm_{rc\eta} = \begin{cases} \frac{IPComm_{c\eta}}{|R^{min}_{c\eta}|} & \text{if } r \in R^{min}_{c\eta} \\ 0 & \text{otherwise} \end{cases} \tag{C.2}$$

where $|R^{min}_{c\eta}|$ is the size of $R^{min}_{c\eta}$, or the multiplicity of most-direct routes serving η. Traffic over each route is defined as in the computationally-simpler method, and then the amount of traffic originating from the IXP and present on link kl is given by summing over IP addresses η and routes r:

$$TotalTraffic_{ckl} = \sum_{\eta} \sum_{r} Traffic_{kl}(c, \eta, r) \tag{C.3}$$

C.3 Probabilistic Assignment of Autonomous Systems to Countries

Rasti et al. (2010) details an approach to geolocating networks that, while more detailed than the one I use, adds enough computational complexity to the processing of my communication data that I ultimately chose to leave it as an alternate methodology in this appendix.

My approach to geolocating an autonomous system maps each network to a single country: the one in which the plurality of its IP addresses are located according to the MaxMind geolocation database.

The approach of Rasti et al. (2010) instead maps a network probabilistically[1] to a set of countries with probabilities based on the distribution of its IP addresses. This can then be used to map a flow of Internet communication between networks into a set of flows between countries:

$$Traffic^{rs}_{ij} = p_i p_j Traffic^{rs} \tag{C.4}$$

where p^r_i is the probability of an IP address in network r being located in country i. Likewise p^s_j is the probability of an address in s being located in j, and $Traffic^{rs}$

[1] Or rather, the mapping can be interpreted probabilistically.

is a flow of traffic observed from r to s. $Traffic_{ij}^{rs}$ is then the flow of traffic from country i to j due to the traffic passing from network r to s. More sophisticated mappings are of course possible, for example by placing additional weight on country-pairs that are closer geographically so that a larger proportion of traffic from r to s is mapped to links between closer countries.

However, the difficulty in employing this methodology is that, due to the one-to-many mapping of newtork-pair flows into country-pair flows, the time required to process a dataset rises significantly. In my small-scale test runs I determined that this approach required 1-2 orders of magnitude longer to process a complete set of communication traces, with indications that the effect would be more pronounced in larger datasets. As the processing of a complete set of communication data already took roughly a day, I instead opted for the simpler one-to-one mapping approach.

REFERENCES CITED

Aggarwal, R., Kearney, C., & Lucey, B. (2012). Gravity and culture in foreign portfolio investment. *Journal of Banking & Finance*, *36*(2), 525-538. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0378426611002536` doi: https://doi.org/10.1016/j.jbankfin.2011.08.007

Allen, T. (2014). Information Frictions in Trade. *Econometrica*, *82*(6), 2041-2083. doi: 10.3982/ECTA10984

Allen, T., & Arkolakis, C. (2014). Trade and the Topography of the Spatial Economy [Article]. *Quarterly Journal of Economics*, *129*(3), 1085-1140. doi: {10.1093/qje/qju016}

Allen, T., & Arkolakis, C. (2019, January). The Welfare Effects of Transportation Infrastructure Improvements [Working Paper]. (25487). Retrieved from `http://www.nber.org/papers/w25487` doi: {10.3386/w25487}

Anderson, J. E., & van Wincoop, E. (2003, March). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, *93*(1), 170-192. doi: 10.1257/000282803321455214

Anderson, J. E., & van Wincoop, E. (2004, September). Trade costs. *Journal of Economic Literature*, *42*(3), 691-751. doi: 10.1257/0022051042177649

Bahar, D. (2019). Measuring knowledge intensity in manufacturing industries: a new approach. *Applied Economics Letters*, *26*(3), 187-190. Retrieved from `https://doi.org/10.1080/13504851.2018.1456643` doi: 10.1080/13504851.2018.1456643

Bahar, D., Hausmann, R., & Hidalgo, C. A. (2014). Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion? *Journal of International Economics*, *92*(1), 111-123. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022199613001098`

Behrens, K., & Picard, P. M. (2011). Transportation, freight rates, and economic geography. *Journal of International Economics*, *85*(2), 280–291.

Beneish, M. D., & Yohn, T. L. (2008). Information friction and investor home bias: A perspective on the effect of global ifrs adoption on the extent of equity home bias. *Journal of Accounting and Public Policy*, *27*(6), 433-443. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0278425408000902` (International Financial Reporting Standards) doi: https://doi.org/10.1016/j.jaccpubpol.2008.09.001

Berkel, B. (2007). Institutional determinants of international equity portfolios-a country-level analysis. *The BE Journal of Macroeconomics*, *7*(1).

Blonigen, B. A., Cristea, A., & Lee, D. (2020). Evidence for the effect of monitoring costs on foreign direct investment. *Journal of Economic Behavior & Organization*, *177*, 601 - 617. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0167268120301955` doi: https://doi.org/10.1016/j.jebo.2020.06.008

Blonigen, B. A., & Wilson, W. W. (2008). Port efficiency and trade flows. *Review of international Economics*, *16*(1), 21–36.

Brancaccio, G., Kalouptsidi, M., & Papageorgiou, T. (2020). Geography, transportation, and endogenous trade costs. *Econometrica*, *88*(2), 657-691. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15455` doi: https://doi.org/10.3982/ECTA15455

Brancaccio, G., Kalouptsidi, M., Papageorgiou, T., & Rosaia, N. (2020, June). *Search frictions and efficiency in decentralized transportation markets* (Working Paper No. 27300). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w27300` doi: 10.3386/w27300

Carter, L., Burnett, D., Drew, S., Marle, G., Hagadorn, L., Bartlett-McNeil, D., & Irvine, N. (2009). Submarine cables and the oceans – connecting the world. *UNEP-WCMC Biodiversity Series*, *31*.

Center for Applied Internet Data Analysis. (2016). *The CAIDA Anonymized Internet Traces 2015-2016.* `https://www.caida.org/catalog/datasets/passive_dataset_download/`.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589–8594. doi: 10.1073/pnas.1017031108

Cisco. (2016). *Bgp best path selection algorithm.* `https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html`. ((Accessed on 05/27/2022))

Clark, X., Dollar, D., & Micco, A. (2004). Port efficiency, maritime transport costs, and bilateral trade. *Journal of development economics*, *75*(2), 417–450.

Clarke, G. R. G., & Wallsten, S. J. (2006). Has the internet increased trade? Developed and developing country evidence [Article]. *Economic Inquiry*, *44*(3), 465-484. doi: {10.1093/ei/cbj026}

Coeurdacier, N., & Martin, P. (2007). The geography of asset holdings: Evidence from sweden. *Riksbank Research Paper Series*(202).

Cristea, Anca D. (2015, MAY). The effect of communication costs on trade in headquarter services. *Review of World Economics*, *151*(2), 255-289. doi: {10.1007/s10290-015-0214-0}

Da Lozzo, G., Di Battista, G., & Squarcella, C. (2014). Visual discovery of the correlation between bgp routing and round-trip delay active measurements. *Computing*, *96*(1), 67–77.

Daude, C., & Fratzscher, M. (2008). The pecking order of cross-border investment. *Journal of International Economics*, *74*(1), 94-119. Retrieved from `https://EconPapers.repec.org/RePEc:eee:inecon:v:74:y:2008:i:1:p:94-119`

De Santis, R., & Gérard, B. (2006). *Financial integration, international portfolio choice and the european monetary union* (Working Paper Series No. 626). European Central Bank. Retrieved from `https://EconPapers.repec.org/RePEc:ecb:ecbwps:2006626`

Doan, T. V., Bajpai, V., Ott, J., & Pajevic, L. (2019). Tracing the path to youtube: A quantification of path lengths and latencies toward content caches. *IEEE communications magazine*, *57*(1), 80–86.

Donaldson, D., & Hornbeck, R. (2016, 02). Railroads and American Economic Growth: A "Market Access" Approach *. *The Quarterly Journal of Economics*, *131*(2), 799-858. Retrieved from `https://doi.org/10.1093/qje/qjw002` doi: 10.1093/qje/qjw002

Duranton, G., & Storper, M. (2008). Rising trade costs? agglomeration and trade with endogenous transaction costs. *Canadian Journal of Economics/Revue canadienne d'économique*, *41*(1), 292-319. Retrieved from `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2966.2008.00464.x` doi: https://doi.org/10.1111/j.1365-2966.2008.00464.x

Ejrnæs, M., & Persson, K. G. (2010, 07). The gains from improved market efficiency: trade before and after the transatlantic telegraph. *European Review of Economic History*, *14*(3), 361-381. doi: 10.1017/S1361491610000109

Fan, Q., & Salas Garcia, V. B. (2018). Information access and smallholder farmers' market participation in peru. *Journal of agricultural economics*, *69*(2), 476–494.

Faruqee, H., & Yan, I. K. (2004). The determinants of international portfolio holdings and home bias.

Fernández, A., Klein, M. W., Rebucci, A., Schindler, M., & Uribe, M. (2016). Capital control measures: A new dataset. *IMF Economic Review*, *64*(3), 548–574.

Ferreira, M. A., & Miguel, A. F. (2007). Home equity bias and industry concentration. *Available at SSRN 989341*.

Fink, C., Mattoo, A., & Neagu, I. (2005, DEC). Assessing the impact of communication costs on international trade. *Journal of International Economics*, *67*(2), 428-445. doi: {10.1016/j.jinteco.2004.09.006}

Freund, C. L., & Weinhold, D. (2004, January). The effect of the Internet on international trade. *Journal of International Economics*, *62*(1), 171-189. Retrieved from https://ideas.repec.org/a/eee/inecon/v62y2004i1p171-189.html

Ganapati, S., Wong, W. F., & Ziv, O. (2020). Entrepot: Hubs, Scale, and Trade Costs [CESifo Working Paper].

Gokan, T., Kichko, S., & Thisse, J.-F. (2019, SEP). How do trade and communication costs shape the spatial organization of firms? *Journal of Urban Economics*, *113*. doi: {10.1016/j.jue.2019.103191}

Grimes, A., Ren, C., & Stevens, P. (2012). The need for speed: impacts of internet connectivity on firm productivity. *Journal of productivity analysis.*, *37*(2), 187–201.

Gruber, S., & Marattin, L. (2010). Taxation, infrastructure and endogenous trade costs in new economic geography. *Papers in Regional Science*, *89*(1), 203–222.

Head, K., & Mayer, T. (2014). Chapter 3 - gravity equations: Workhorse,toolkit, and cookbook. In G. Gopinath, E. Helpman, & K. Rogoff (Eds.), *Handbook of international economics* (Vol. 4, p. 131-195). Elsevier. Retrieved from https://www.sciencedirect.com/science/article/pii/B9780444543141000033 doi: https://doi.org/10.1016/B978-0-444-54314-1.00003-3

Huawei. (2019). *Bgp route selection rules.* https://support.huawei.com/enterprise/en/doc/EDOC1100055099/70e83769/bgp-route-selection-rules. ((Accessed on 05/27/2022))

Hummels, D. (2007, September). Transportation costs and international trade in the second era of globalization. *Journal of Economic Perspectives*, *21*(3), 131-154. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/jep.21.3.131` doi: 10.1257/jep.21.3.131

Hummels, D., Lugovskyy, V., & Skiba, A. (2009). The trade reducing effects of market power in international shipping. *Journal of Development Economics*, *89*(1), 84–97.

Jonkeren, O., Demirel, E., van Ommeren, J., & Rietveld, P. (2011). Endogenous transport prices and trade imbalances. *Journal of Economic Geography*, *11*(3), 509–527.

Juniper Networks. (2020). *Understanding bgp path selection.* `https://www.juniper.net/documentation/en_US/junos/topics/reference/general/routing-protocols-address-representation.html`. ((Accessed on 05/27/2022))

Karolyi, G. A. (2016). The gravity of culture for finance. *Journal of Corporate Finance*, *41*, 610–625.

Keller, W., & Yeaple, S. R. (2013). The gravity of knowledge. *American Economic Review*, *103*(4), 1414–44.

Kikuchi, T. (2002, FEB). Interconnectivity of communications networks and international trade. *Canadian Journal of Economics-Revue Canadienne D Economique*, *36*(1), 155-167. doi: {10.1111/1540-5982.00008}

Kleinert, J., & Spies, J. (2011). Endogenous transport costs in international trade. *Working Paper*.

Lane, P. R., & Milesi-Ferretti, G. M. (2008). International investment patterns. *The Review of Economics and Statistics*, *90*(3), 538–549.

Leuven, E., Akerman, A., & Mogstad, M. (2018, Feb). Information Frictions, Internet and the Relationship between Distance and Trade [Memorandum]. (1/2018). Retrieved from `https://ideas.repec.org/p/hhs/osloec/2018_001.html`

Lew, B., & Cater, B. (2006, AUG). The telegraph, co-ordination of tramp shipping, and growth in world trade, 1870-1910. *European Review of Economic History*, *10*(2), 147-173. doi: {10.1017/S1361491606001663}

Limer, E. (2016). *Ship's anchor slices through three undersea internet cables.* `https://www.popularmechanics.com/technology/infrastructure/a24064/anchor-cuts-undersea-internet-cables/`.

Limão, N., & Venables, A. J. (2001). Infrastructure, geographical disadvantage, transport costs, and trade. *The World Bank Economic Review*, *15*(3), 451–479. Retrieved from `http://www.jstor.org/stable/3990110`

Mishra, A. V. (2007). International investment patterns: evidence using a new dataset. *Research in International Business and Finance*, *21*(2), 342–360.

Nagy, D. (2016). City location and economic development [2016 Meeting Papers]. (307). Retrieved from `https://ideas.repec.org/p/red/sed016/307.html`

Nardo, M., Ndacyayisenga, N., Pagano, A., & Zeugner, S. (2017). Finflows: database for bilateral financial investment stocks and flows. (KJ-NA-28833-EN-N). doi: 10.2760/172684

Okawa, Y., & van Wincoop, E. (2012, JUL). Gravity in International Finance. *Journal of International Economics*, *87*(2), 205-215. doi: {10.1016/j.jinteco.2012.01.006}

Oregon Route Views Project. (1997). *Route Views Archive.* `http://www.routeviews.org/routeviews/`.

Poese, I., Uhlig, S., Kaafar, M. A., Donnet, B., & Gueye, B. (2011). Ip geolocation databases: Unreliable? *ACM SIGCOMM Computer Communication Review*, *41*(2), 53–56.

Poulsen, K. (2006). *The backhoe: A real cyberthreat — wired.* `https://www.wired.com/2006/01/the-backhoe-a-real-cyberthreat/`.

Rasti, A., Magharei, N., Rejaie, R., & Willinger, W. (2010, 01). Eyeball ases: from geography to connectivity. In (p. 192-198). doi: 10.1145/1879141.1879165

Rauch, J. E. (1996, June). Networks versus markets in international trade [Working Paper]. (5617). Retrieved from `http://www.nber.org/papers/w5617` doi: 10.3386/w5617

Redding, S. (2016). Goods trade, factor mobility and welfare. *Journal of International Economics*, *101*(C), 148-167. Retrieved from `https://EconPapers.repec.org/RePEc:eee:inecon:v:101:y:2016:i:c:p:148-167`

Redding, S. J., & Turner, M. A. (2014, June). *Transportation costs and the spatial organization of economic activity* (Working Paper No. 20235). National Bureau of Economic Research. Retrieved from `http://www.nber.org/papers/w20235` doi: 10.3386/w20235

Roque, V., & Cortez, M. C. (2014). The determinants of international equity investment: Do they differ between institutional and noninstitutional investors? *Journal of Banking & Finance*, *49*(C), 469-482. Retrieved from `https://ideas.repec.org/a/eee/jbfina/v49y2014icp469-482.html` doi: 10.1016/j.jbankfin.2014.0

Sotelo, S. (2015, May). Domestic Trade Frictions and Agriculture [Working Papers]. (641). Retrieved from `https://ideas.repec.org/p/mie/wpaper/641.html`

Steinwender, C. (2018, Mar). Real Effects of Information Frictions: When the States and the Kingdom Became United. *American Economic Review*, *108*(3), 657-696. doi: {10.1257/aer.20150681}

TeleGeography. (2020). *Undersea cable map.* `https://github.com/telegeography/www.submarinecablemap.com`. GitHub.

United Nations. (2003). *UN Comtrade Database.* Retrieved from `http://comtrade.un.org`

US Treasury. (2022). *US Treasury International Capital Data.* Retrieved from `https://home.treasury.gov/data/treasury-international-capital-tic-system`

Zeileis, A. (2021). *pwt10: Penn world table (version 10.x).* Retrieved from `https://CRAN.R-project.org/package=pwt10` (R package version 10.0-0)