Routing Data as a Measure of Internet Access

Zachary Kiefer

University of Oregon

June 3, 2020

Abstract

I propose a pair of new measures of Internet access which can be computed automatically, across large geographic areas, from publicly-available data which describe how information is routed across the Internet. These measures can also be computed at relatively fine levels of detail. I then demonstrate empirically that these measures perform comparably to measures previously used in the literature, and that the two measures capture largely distinct aspects of Internet access.

Acknowledgements

The author extends his thanks to Dr. Woan Foong Wong of the University of Oregon, for her guidance in focusing and preparing this paper. Special thanks also go to the University of Oregon's Route Views Project in general, and to Eric Smith in particular, for their assistance in accessing and interpreting their data.

1 Introduction

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 an important economic variable, but it is difficult to measure Internet access, due partly to the anonymous and decentralized nature of the Internet. Existing measures of Internet access suffer from notable flaws: survey-based measures suffer from the perennial issue of comparability between countries, while measures derived from objective data have so far been debatably effective at capturing the true extent of Internet access. Additionally, there are few measures which are more detailed than the country-level or which capture factors such as the quality of Internet access.

In this paper, I propose a pair of new measures of Internet access which can be computed from publicly-available data which describe how information is routed across the Internet. These measures can also be computed at the province level with little additional effort. I then demonstrate empirically that these measures perform comparably to measures previously used in the literature, and that the two measures capture largely separate aspects of Internet access.

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, when the survey administrators use different definitions of "Internet user," 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.

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[1] 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 in 2004, 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. This may limit the usefulness of the webhost-counting measure of Internet access.

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, and again relies on a micro-survey approach.

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

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 is not available. Given that this is a similar situation, in which detailed and/or accurate data on Internet access are not always available, a similar approach is justifiable here.

3 Data

This paper 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.

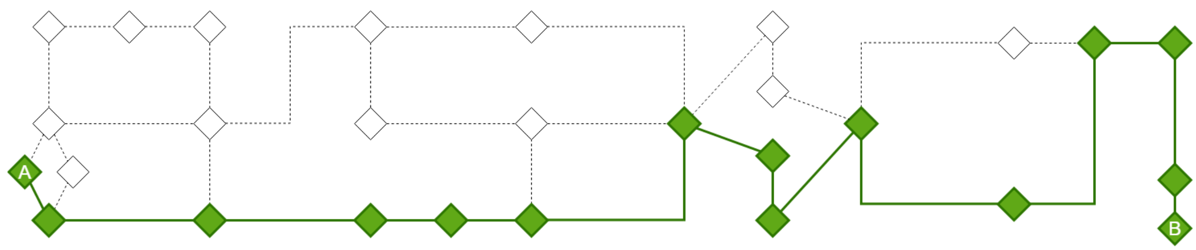
3.1 Internet Routing

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.

Most Internet-connected devices have very few direct connections: they may be wired to other devices in the same household (or accomplish the same thing using WiFi), but if two devices on opposite sides of the world need to communicate with each other, they seldom have a cable which connects them together. Most Internet traffic is therefore forwarded through several intermediary devices to reach its recipient. A *route* is a sequence of IP addresses, starting with a sender and ending with a recipient, describing a chain of intermediary devices that can be used to pass data from the sender to the recipient.

Figure 1 shows a sample network diagram with a sample route from A to B highlighted. In this diagram, diamonds indicate Internet-connected devices (each with its own IP address). Lines indicate direct connections between devices.

Figure 1
A Sample Route from Sender A to Receiver B



Every Internet-connected device also possesses a *routing table* which contains information on the routes which the device can use to send data. When a device needs to send data to another Internet-connected device, it searches its routing table for a route leading to that recipient, and then follows the instructions in that route. The routing table does not necessarily store complete routes: more commonly, the routing table only stores the "next hop" in a route, since all that it needs to know is which device to forward data to next.

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 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, or to a delivery company such as UPS or FedEx, after which it is up to the service or company to get the package to its recipient. The postal service, UPS, and FedEx, on the other hand, need to have detailed information about how to get the package from any given point A to point B.

Extending this analogy, if one were to obtain the complete set of the post office's 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 sorting centers were, or estimate the quality of mail service between towns A and B based on how many steps were in the instructions. The measures of Internet access I propose are based on a similar line of thinking, in that the routing data used by large IXPs conveys information about where Internet users are located and how good their Internet access is.

3.2 Routing Data

Through the Oregon Route Views Project (ORVP), hosted at the University of Oregon, I have gained access to an archive of routing data from major IXPs 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.

The raw data I use consists of observations of the routing tables in devices hosted at IXPs (which I shall refer to as *collectors* from now on), coupled with additional data not normally contained in the routing table.

The unit of observation in this data is the *IP address block*, a group of consecutive IP addresses which are all located at the end of the same route. Each observation contains a list of the IP addresses in the block, the "next hop" in the route to this block, and additional information about the *Autonomous Systems*[2] involved in the route. It is common for there to be multiple simultaneous observations of the same IP address block; these are observations of multiple distinct routes, which provide redundancy in case one is temporarily disrupted.

Observations are taken roughly every two hours, and the collectors which I have focused on for my empirical work[3] have been contributing observations since 2003.

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.

[2] I will address Autonomous Systems more extensively in Section 3.4.

[3] PAIX, the Palo Alto Internet eXchange; EQIX, the Equinix-Ashburn exchange; and LINX, the London InterNet eXchange.

Because the collectors which generate my data are located in highly-connected IXPs, it can be assumed that any IP address observed in their routing table has recently used the Internet.[4] The IP addresses observed in one of these routing tables can therefore be viewed as the set of all IP addresses which have recently been in use.

These IP addresses can be geolocated to the country level—and in many cases, to the province or even city level—using Maxmind commercial geolocation databases.[5] This enables me to construct a variable tracking the number of IP addresses in use, at the country or province level, over the time period from 2003 to 2018.

This variable cannot be interpreted as a count of actual Internet users—because users may be associated with multiple IP addresses—but I propose that it is positively correlated with both the number of Internet users at a location, and the extent of their Internet use. In particular, changes in this IP address count can be interpreted as changes in the number of Internet users, or the extent of Internet use, in a given location.

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.[6]

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 2 shows the growth in median national per-capita

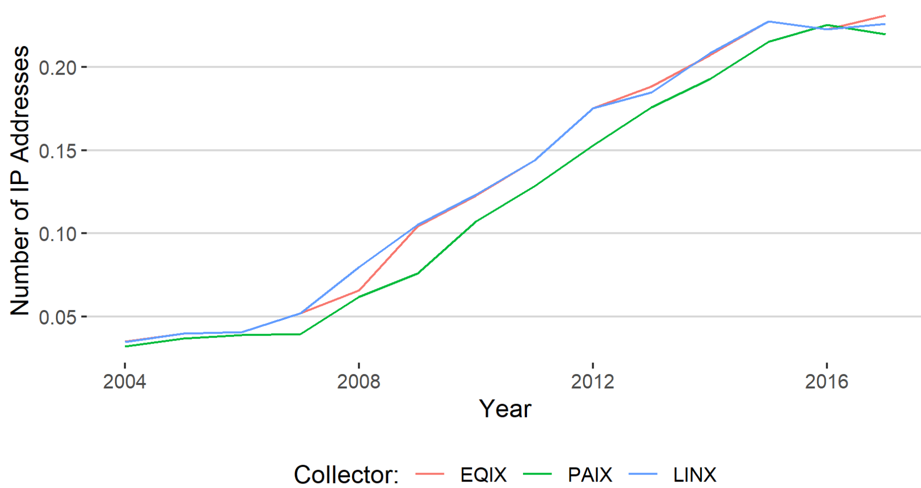
[4] Without going into excessive technical detail, the Border Gateway Protocol ensures that when an IP address block can no longer be reached via a route, that knowledge propagates rapidly.

[5] These are the same databases 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.

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

IP addresses, as observed from three collectors, while Figure 3 shows per-capita IP address counts around the world in 2004, 2010, and 2016.

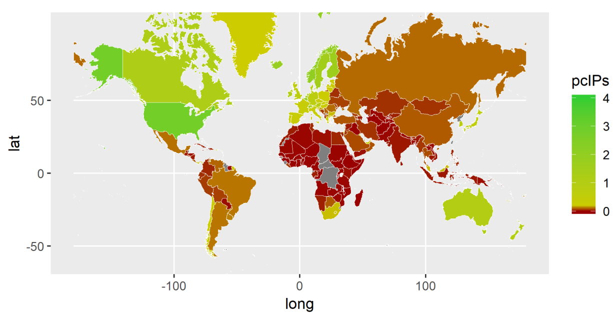
Figure 2
Median National Per-Capita IP Addresses



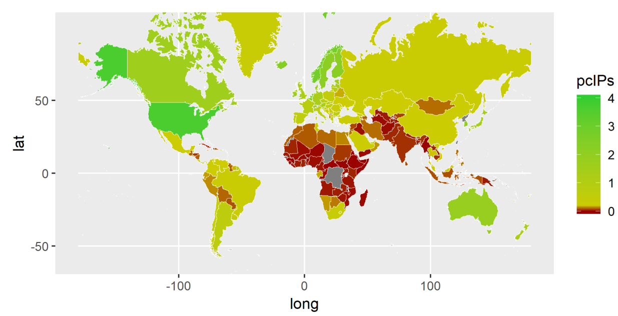
The three collectors which I focus on report highly-correlated IP address counts, as seen in Figure 4. Aside from a small number of extreme outliers (attributable to technical glitches in which one collector briefly observed no IP addresses for a location), there is virtually a one-to-one relationship between the values reported by different collectors.

Figure 3
Per-Capita IP Addresses Around the World

(a) Year: 2004



(b) Year: 2010



(c) Year: 2016

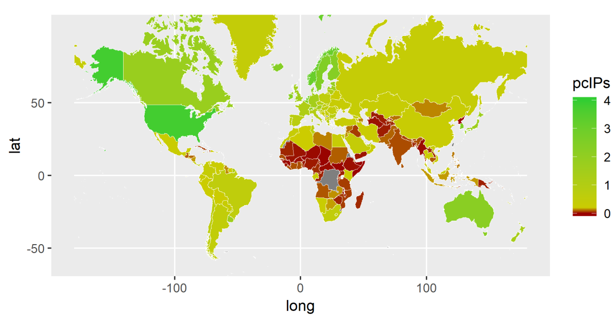
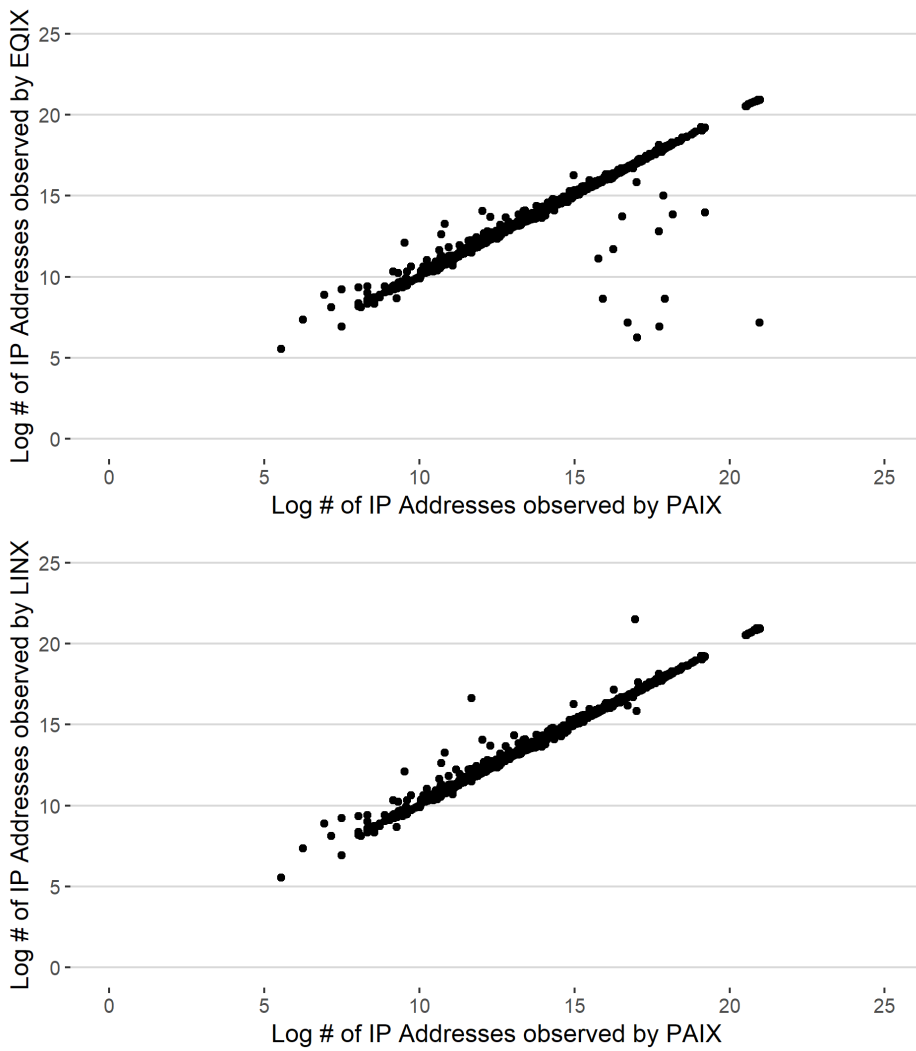


Figure 4
IP Address Count Correlation Among Collectors



3.4 Measuring Route Length

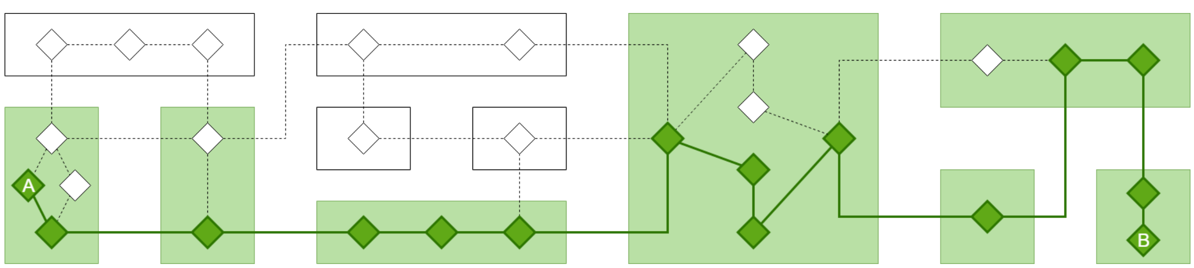
The second of my proposed measures is the *Aggregate-Autonomous System Path Length (Aggregate-ASPL)*. This is a measure of how complicated it is to send data from a major IXP to a target location.

An *Autonomous System (AS)* is a collection of Internet-connected devices, identified by their IP addresses, which are managed by the same organization or entity. ASes include ISPs, IXPs, Internet companies such as Amazon and Google, and others. Each AS has a unique identifying number, much like an IP address.[7]

One of the variables in the routing data is the *Autonomous System Path (ASP)*, an ordered list of all the Autonomous Systems which a route to the target IP address block would pass through. It can be thought of as a list of all the organizations whose cooperation is necessary to use the route.

Figure 5 shows the same route from Figure 1 with the associated ASP marked in light green. As before, diamonds indicate Internet-connected devices and lines indicate direct connections between devices. Rectangular boxes containing one or more devices denote ASes. This particular ASP has a length of 7.

Figure 5
A Sample Autonomous System Path

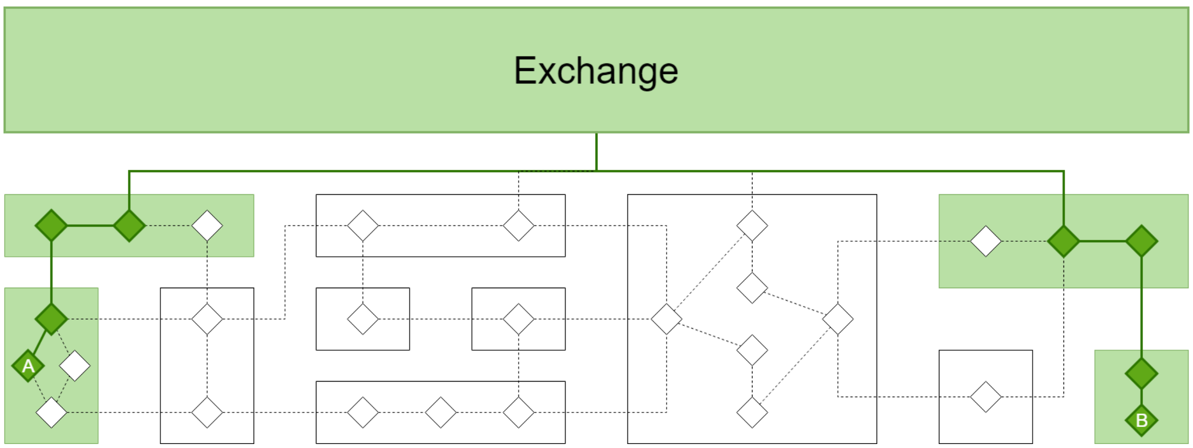


However, I am unable to observe these routes directly, as the routing data collected by ORVP only contains data on routes which originate with one of the collectors contributing to the project. Instead of seeing a route directly from A to B, I am only able to observe routes from a third, outside location to A and B, such as those illustrated in Figure 6. The ASPs from the Exchange to devices A and B shown in this diagram each have length 2, and the ASP connecting A and B via the Exchange has length 5.

[7]The University of Oregon, which acts as its own ISP, is AS #7922, for example.

Because IXPs are highly connected (they have connections to many devices, including some geographically distant ones), and tend to be located at "hubs" of Internet activity, there is a strong tendency for the shortest route from A to B to pass through an IXP naturally, just as seen in Figure 6.

Figure 6
Autonomous System Paths Through an Exchange



Based on interviews with the network engineers who run the ORVP, I make three assumptions about the ASP:

- The length of a route's ASP is positively correlated with latency (the time which it takes data to reach the target IP address block), simply because each additional AS introduces additional computational steps into the routing process.

- The length of a route's ASP is negatively correlated with the route's reliability: each additional AS introduces another point of failure into the route, increasing the frequency of service outages or dropped packets.

- The length of a route's ASP is positively correlated with the cost of Internet access for the end-users: use of the route requires the cooperation of all the ASes involved, and those ASes are rent-seekers who wish to receive some fraction of the end-users' subscription fees.

Based on these three assumptions, the length of the ASP 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 ASP 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 negatively 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 et al., 2019)

I construct the Aggregate-ASPL 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 ASes) to each destination block. My contact at the ORVP states that this is the most common method that ISPs use to choose a route, and as such, is most likely to reflect the actual route used. The length of this route is the Autonomous System Path Length for the IP address block.

- I then aggregate to the province or country level, using a 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 descriptive statistics of the ASPL, weighting by the size of the IP address block so that each individual IP address receives equal weight. This is the Aggregate-ASPL for a given location.

- Finally, when aggregating to the country level at a global scope, I perform this aggregation using routing data from multiple, geographically separated IXPs, and select the median value of the Aggregated ASPL, in order to obtain a measure which is representative of access to the global Internet, as opposed to access to a particular IXP.

In essence, Aggregate-ASPL 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.

3.4.1 Descriptive Statistics

Table 2 presents descriptive statistics for the Aggregate-ASPL measure at the same two levels of aggregation.

Table 2
Descriptive Statistics: Aggregate-ASPL

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

There is again significant variation across time and location, as seen in Tables 7 and 8. However, this Aggregate-ASPL measure contains noticeably more noise, both year-to-year and between collectors. Notably, while PAIX and EQIX report similar median Aggregate-ASPL values, LINX reports a significantly lower median.

Much as with the IP address count, there is a positive correlation between the Aggregate-ASPL values reported by different collectors, as seen in figure 9. However, the correlation is weaker here—as would be expected, since the ASP to a target location would be heavily influenced by the location of the collector. Again, LINX reports lower Aggregate-ASPL than PAIX. This may be due to London's advantageous position at the end of multiple undersea cables, which would allow for shorter ASPs to many distant locations.

Figure 7
Median National Aggregate-ASPL

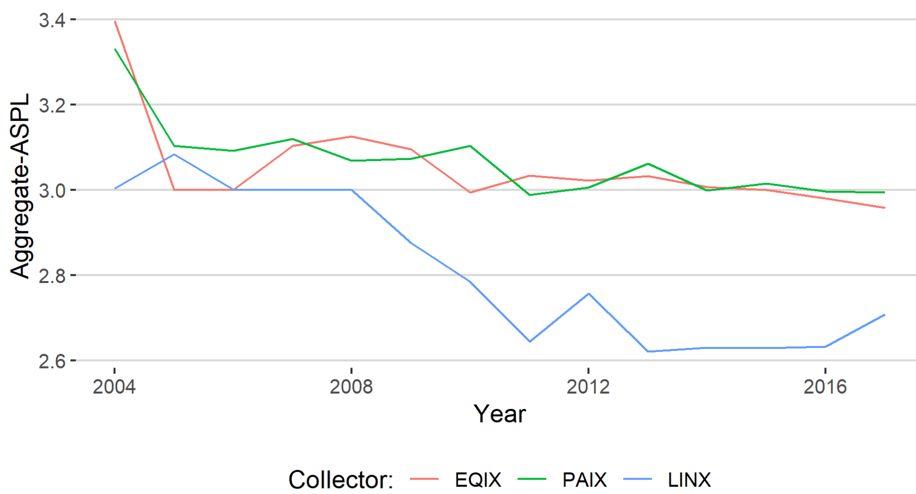
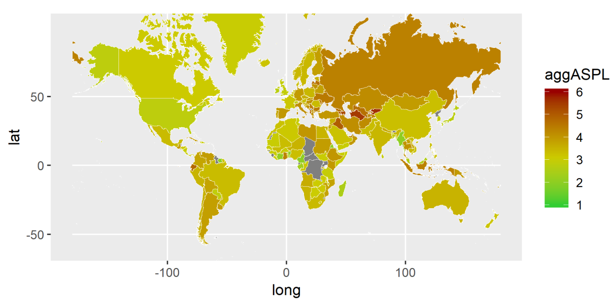
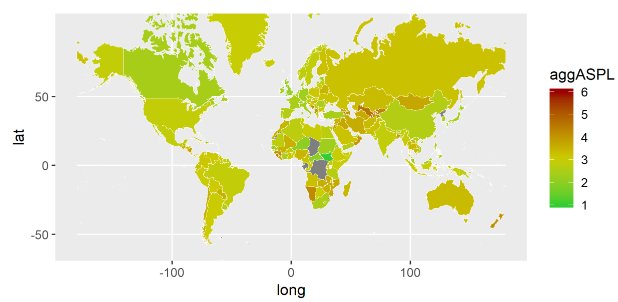


Figure 8
Aggregate-ASPL Around the World

(a) Year: 2004



(b) Year: 2010



(c) Year: 2016

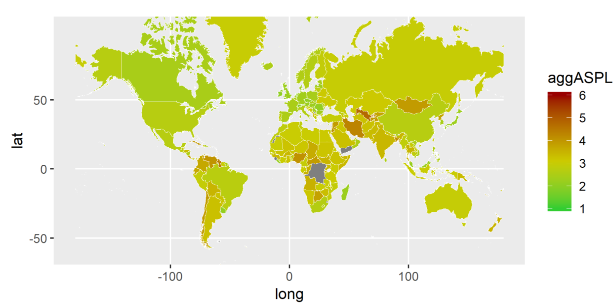
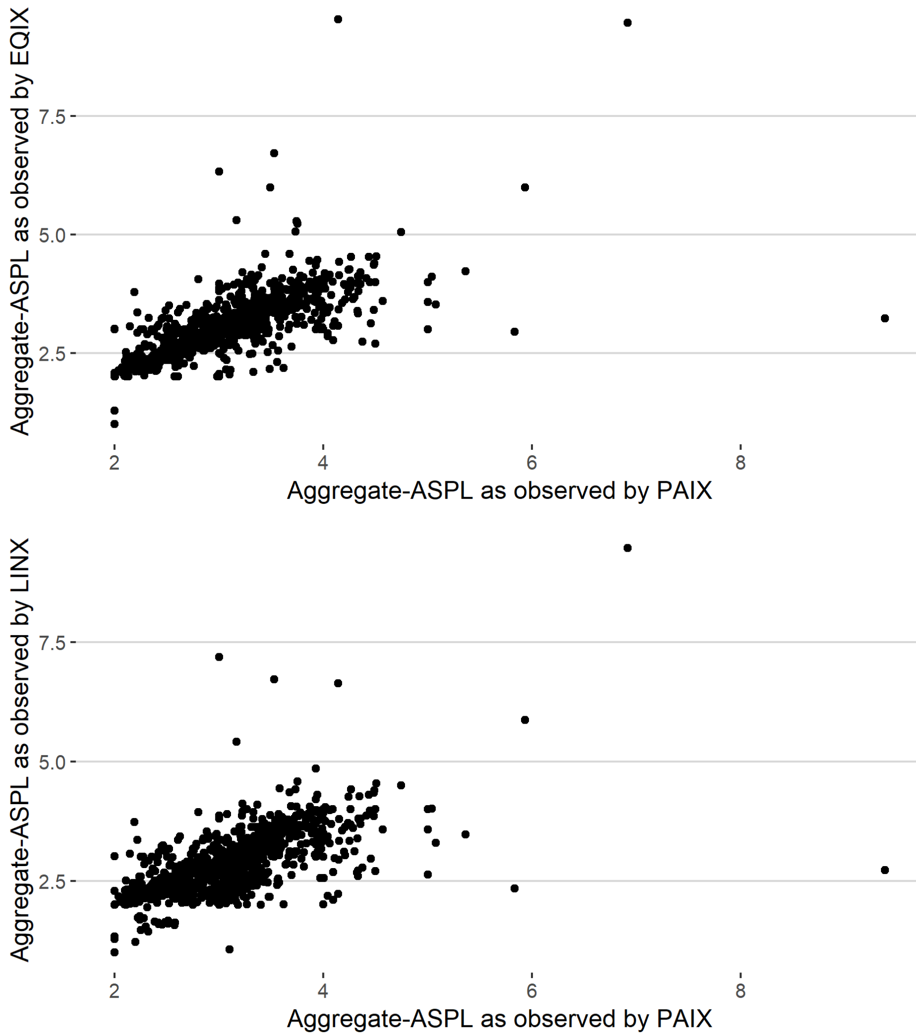


Figure 9
Aggregate-ASPL Correlation Among Collectors



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.

There exists a commercial version of the free datasets which I use in this paper, 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.

3.6 Computation

I compute these measures using a set of R scripts utilizing the *tidyverse* packages. These scripts also use the BGPReader tool for Linux, the Geospatial Data Abstraction Library (GDAL), and other resources. All data-collection scripts are available upon request.

3.7 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.

4 Empirical Results

4.1 Freund and Weinhold Replication

I begin by replicating the trade-growth models from Table 3 of Freund and Weinhold (2004), substituting my proposed measures of Internet access. With the exception of Model (1) of Table 3, which is a baseline model that includes no measures of Internet access, the models of this table are of the form

$$\begin{aligned} gExports_{ijt} =& \beta_0(gNumIPs_i)_{t-1} + \beta_1(gNumIPs_j)_{t-1} + \beta_2 ln(NumIPs_i)_{04} \\ &+ \beta_3 ln(NumIPs_j)_{04} + \beta_4 ln(Export_{ij})_{04} + \beta_5(gGDP_j)_t \\ &+ \beta_6 log(Distance_{ij}) + \beta X_{ijt} + FE_t + \epsilon_{ijt}. \end{aligned}$$

Here, i describes origin (exporting) countries, j describes destination (importing) countries, and t describes year.

$gExports_{ijt}$ is the growth in exports from i to j between years $t-1$ and t. $(gNumIPs_i)_{t-1}$ and $(gNumIPs_j)_{t-1}$ are likewise the growth in the count of IP addresses contained in countries i and j, respectively. $ln(NumIPs_i)_{04}$ and $ln(NumIPs_j)_{04}$ 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. $ln(Export_{ij})_{04}$ is likewise a control for initial exports from i to j. $(gGDP_j)_t$ is the growth in j's GDP and $log(Distance_{ij})$ is the log of the distance between the centroids of i and j; Freund and Weinhold's theoretical model predicts that both of these will influence trade. X_{ijt} is a vector of controls, and FE_t is a year fixed effect. Lastly, ϵ_{ijt} is a heteroskedasticity and autocorrelation-robust error term.

Model (2) is the first model using the IP address count, but includes no controls.

Model (3) is the same model as (2), but is estimated after deleting all observations which had a residual more than four standard deviations from zero in model (2).[8]

Models (4) and (5) introduce control variables, continuing to delete outliers. Model (4) introduces controls for various economic factors, and (5) introduces a lag of the dependent

[8] About 1.3% of the sample.

variable to account for persistent trends.

The data used in these models, as well as in the models of Tables 4 to 7 represent 82 countries, chosen for having complete sets of routing and economic data, over the period of time from 2004 to 2014.

As can be seen in Table 8, the variables using IP address count become less significant with the introduction of control variables: the only measure of Internet penetration to remain significant in (5) is the measure of initial IP addresses in the destination country. Additionally, persistent trends account for far more variation in the growth of exports than any other factor, as seen from the large increase in R^2 from model (4) to (5).

Table 4 repeats the models from Table 3, but substitutes Aggregate-ASPL for the IP address count. Aggregate-ASPL is non-significant at the 5% level in all models, although it should be noted that the signs on the growth of ASPL 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 larger ASPL represents more difficult access.

In both of these tables, which closely follow the models used by Freund and Weinhold, much of the models' explanatory power appears to come from variables which control for initial conditions (e.g. $\text{Log}(\text{EXPORT}_{12})_{95}$) 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. $\text{Log}(\text{NumIPs}_2)_{95}$ in Table 3).

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 ($\text{Log}(\text{ASPL}_2)_{04}$) 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 Aggregate-ASPL. I can draw no direct comparisons between the coefficients on the ASPL 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.

4.2 Freund and Weinhold Adaptation

In Table 5, I modify the original Freund model to use a large number[9] of origin-destination fixed effects as a substitute for many of the control variables. Adapted models (1)-(3) are of the form

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

Here, FE_{ij} is an origin-destination fixed effect, while all other variables are defined as they were previously. The controls included in X_{ijt} are restricted only to those which vary by origin-year and/or destination-year.

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 control for the 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[10]—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 number of IP addresses in both countries combined:

[9] Approximately 5500.

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

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

Here, $(gNumIPsTotal_{ij})_{t-1}$ is the lagged growth of total IP addresses in i and j combined. The controls in X_{ijt} are only those which vary by origin-destination-year. 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 $(gNumIPsTotal_{ij})_{t-1}$.

Model (4) includes no controls; model (5) introduces a lag of the dependent variable.

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 Aggregate-ASPL 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 ASPL 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 Aggregate-ASPL remains the more useful measure.

In all of these models, the coefficients on Aggregate-ASPL 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 larger ASPL represents 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.

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 Aggregate-ASPL 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 Aggregate-ASPL 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 Aggregate-ASPL lie within a relatively small band: a small percentage change in Aggregate-ASPL can therefore have a large impact.

4.3 Allen Replication

4.3.1 Simultaneous Import and Export

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

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

$$ImpExp_{itc} = \beta NetworkAccess_{it} + FE_i + FE_c + \epsilon_{itc}.$$

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.

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 Aggregate-ASPL 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 ASPL has upon this market failure. In this table, all models are of the form

$$ImpExp_{itc} = \beta aggASPL_{it} + FE_i + FE_c + \epsilon_{itc}.$$

Here, $aggASPL_{it}$ is the Aggregate-ASPL from the PAIX exchange in San Francisco[11] 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 larger Aggregate-ASPL makes 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 a long ASP) 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.

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

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

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

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

$$dLogDestPR_{ijt} = \beta_0 dLogOrigPR_{ijt} + \beta_1 dLogOrigPR_{ijt} \times Connection_{ijt} + FE_t + \epsilon_{ijt}.$$

In both forms of the model, i represents origin province, j represents destination province,

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

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.

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

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

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

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

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. Aggregate-ASPL remains non-significant in both provinces, although the signs of both coefficients are as predicted. ASPL 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."

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 Aggregate-ASPL in the 95th, 75th, 50th, and 25th percentiles.[12] In the case of provinces with Aggregate-ASPL in the 5th percentile,[13] it is not possible to reject this hypothesis.

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 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

$$FarmerTraded_{iymc} = \beta_0 logQuantity_{iymc} + FE_{pymc} + \epsilon_{iymc}.$$

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.

[12]It is important to remember that larger ASPL suggests worse Internet access; provinces with Aggregate-ASPL above the 95th percentile are therefore the 5% of provinces with the worst Internet connection by this measure.

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

Model (2) is of the form

$$\begin{aligned} FarmerTraded_{iymc} =& \beta_0 logQuantity_{iymc} + \beta_1 InternetAccess_{pym} \\ &+ \beta_2 logQuantity_{iymc} \times InternetAccesspym + FE_{pyc} + FE_{mc} + \epsilon_{iymc}, \end{aligned}$$

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

$$\begin{aligned} FarmerTraded_{iymc} =& \beta_0 logQuantity_{iymc} + \beta_2 logQuantity_{iymc} \times InternetAccess_{pym} \\ &+ FE_{pymc} + \epsilon_{iymc}. \end{aligned}$$

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

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

In the first three models, I find that Internet access has a stronger effect than Allen (2014) 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 Aggregate-ASPL into this analysis in model (4), it appears to be the provinces where ASPL 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.

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 Aggregate-ASPL 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.

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.

Table 3
Results: Original Model Using Number of IP Addresses

	Dependent variable:				
	Growth of exports from country 1 to country 2				
	(1)	(2)	(3)	(4)	(5)
$(\text{gNumIPs}_i)_{t-1}$		0.074^{***}	0.031	0.028	0.029
		(0.029)	(0.019)	(0.020)	(0.020)
$(\text{gNumIPs}_j)_{t-1}$		0.016	0.032^{*}	0.031^{*}	0.020
		(0.023)	(0.018)	(0.018)	(0.016)
$\ln(\text{NumIPs}_i)_{04}$		0.009^{***}	0.005^{**}	0.007	0.008
		(0.003)	(0.002)	(0.005)	(0.005)
$\ln(\text{NumIPs}_j)_{04}$		0.007^{***}	0.005^{**}	0.007^{*}	0.011^{***}
		(0.003)	(0.002)	(0.004)	(0.004)
$\ln(\text{Export}_{ij})_{04}$	-0.011^{***}	-0.013^{***}	-0.009^{***}	-0.012^{***}	-0.025^{***}
	(0.002)	(0.004)	(0.003)	(0.003)	(0.003)
$(\text{gGDP}_j)_t$	0.300^{***}	0.313^{***}	0.322^{***}	0.250^{***}	0.238^{***}
	(0.064)	(0.067)	(0.053)	(0.056)	(0.054)
$\ln(\text{Distance}_{ij})$	0.002	-0.005	-0.0004	-0.008	-0.021^{***}
	(0.006)	(0.007)	(0.006)	(0.006)	(0.006)
$(\text{gExchRate}_i)_{t-1}$				-0.123^{**}	-0.154^{***}
				(0.060)	(0.059)
$(\text{gExchRate}_j)_{t-1}$				-0.057	-0.141^{***}
				(0.051)	(0.050)
$\ln(\text{GDP}_i)_{04}$				-0.006	0.002
				(0.008)	(0.007)
$\ln(\text{GDP}_j)_{04}$				-0.005	-0.007
				(0.007)	(0.007)
$\ln(\text{Population}_i)_{04}$				0.014^{***}	0.020^{***}
				(0.004)	(0.004)
$\ln(\text{Population}_j)_{04}$				0.012^{***}	0.021^{***}
				(0.004)	(0.004)
$(\text{gExport}_{ij})_{t-1}$					-0.333^{***}
					(0.009)
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
Note:					$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 4
Results: Original Model Using Aggregate ASPL

	Dependent variable:				
	Growth of exports from country 1 to country 2				
	(1)	(2)	(3)	(4)	(5)
$(\text{gAggASPL}_i)_{t-1}$		-0.090	-0.064	-0.049	-0.078^{*}
		(0.059)	(0.043)	(0.043)	(0.041)
$(\text{gAggASPL}_j)_{t-1}$		0.032	0.038	0.051	0.051
		(0.053)	(0.039)	(0.039)	(0.037)
$\ln(\text{NumIPs}_i)_{04}$		-0.018	-0.015	-0.022	-0.006
		(0.031)	(0.024)	(0.025)	(0.024)
$\ln(\text{NumIPs}_j)_{04}$		0.011	-0.001	-0.010	-0.004
		(0.023)	(0.019)	(0.019)	(0.018)
$\ln(\text{Export}_{ij})_{04}$	-0.011^{***}	-0.006^{***}	-0.004^{***}	-0.012^{***}	-0.024^{***}
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
$(\text{gGDP}_j)_t$	0.300^{***}	0.287^{***}	0.316^{***}	0.261^{***}	0.261^{***}
	(0.064)	(0.065)	(0.052)	(0.056)	(0.054)
$\ln(\text{Distance}_{ij})$	0.002	0.009	0.007	-0.005	-0.017^{***}
	(0.006)	(0.006)	(0.005)	(0.006)	(0.005)
$(\text{gExchRate}_i)_{t-1}$				-0.106^{*}	-0.153^{***}
				(0.059)	(0.057)
$(\text{gExchRate}_j)_{t-1}$				-0.053	-0.143^{***}
				(0.051)	(0.050)
$\ln(\text{GDP}_i)_{04}$				0.004	0.011^{***}
				(0.004)	(0.004)
$\ln(\text{GDP}_j)_{04}$				0.004	0.008^{**}
				(0.004)	(0.003)
$\ln(\text{Population}_i)_{04}$				0.012^{***}	0.017^{***}
				(0.003)	(0.003)
$\ln(\text{Population}_j)_{04}$				0.010^{***}	0.016^{***}
				(0.003)	(0.003)
$(\text{gExport}_{ij})_{t-1}$					-0.332^{***}
					(0.009)
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

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

Table 5
Results: Streamlined Model Using Number of IP Addresses

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
$(\text{gNumIPs}_i)_{t-1}$	0.087***	0.086***	0.051		
	(0.029)	(0.029)	(0.032)		
$(\text{gNumIPs}_j)_{t-1}$	0.013	0.009	−0.004		
	(0.030)	(0.029)	(0.030)		
$(\text{gNumIPs}_{Total})_{t-1}$				0.047	0.046
				(0.082)	(0.084)
$(\text{gGDP}_j)_t$		0.155	0.184		
		(0.122)	(0.142)		
$(\text{gExchRate}_i)_{t-1}$		−0.083	−0.136		
		(0.154)	(0.158)		
$(\text{gExchRate}_j)_{t-1}$		−0.063	−0.035		
		(0.120)	(0.135)		
$(\text{gExport}_{ij})_{t-1}$			−0.400***		−0.407***
			(0.011)		(0.011)
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
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 6
Results: Streamlined Model Using Aggregate ASPL

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
$(\text{gAggASPL}_i)_{t-1}$	-0.113	-0.113	-0.126^{**}		
	(0.073)	(0.073)	(0.063)		
$(\text{gAggASPL}_j)_{t-1}$	0.042	0.039	0.058		
	(0.068)	(0.067)	(0.059)		
$(\text{gAggASPL}_{Total})_{t-1}$				1.044	0.598
				(0.980)	(0.712)
$(\text{gGDP}_j)_t$		0.152	0.175		
		(0.122)	(0.141)		
$(\text{gExchRate}_i)_{t-1}$		-0.104	-0.148		
		(0.152)	(0.156)		
$(\text{gExchRate}_j)_{t-1}$		-0.069	-0.041		
		(0.119)	(0.134)		
$(\text{gExport}_{ij})_{t-1}$			-0.400^{***}		-0.407^{***}
			(0.011)		(0.011)
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
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 7
Results: Streamlined Model Using Both Measures

	Dependent variable:				
	Growth in exports				
	(1)	(2)	(3)	(4)	(5)
$(\text{gNumIPs}_i)_{t-1}$	0.089***	0.088***	0.053*		
	(0.027)	(0.028)	(0.030)		
$(\text{gNumIPs}_j)_{t-1}$	0.014	0.010	−0.003		
	(0.029)	(0.028)	(0.029)		
$(\text{gAggASPL}_i)_{t-1}$	−0.120*	−0.120*	−0.130**		
	(0.067)	(0.067)	(0.062)		
$(\text{gAggASPL}_j)_{t-1}$	0.044	0.040	0.057		
	(0.069)	(0.068)	(0.059)		
$(\text{gNumIPs}_{Total})_{t-1}$				0.047	0.046
				(0.082)	(0.084)
$(\text{gAggASPL}_{Total})_{t-1}$				1.042	0.596
				(0.977)	(0.709)
$(\text{gGDP}_j)_t$		0.150	0.176		
		(0.120)	(0.140)		
$(\text{gExchRate}_i)_{t-1}$		−0.083	−0.136		
		(0.156)	(0.161)		
$(\text{gExchRate}_j)_{t-1}$		−0.067	−0.040		
		(0.119)	(0.134)		
$(\text{gExport}_{ij})_{t-1}$			−0.400***		−0.407***
			(0.011)		(0.011)
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
Note:			*p<0.1; **p<0.05; ***p<0.01		

Table 8
Results: Simultaneous Import/Export

	Dependent variable:			
	Simultaneously imported and exported			
	Prov.-prov., annual		Port-port, 4th quarter	
	(1)	(2)	(3)	(4)
$\text{NetworkAccess}_{it}$	-0.036**	-0.064*	-0.020***	-0.061**
	(0.018)	(0.032)	(0.007)	(0.025)
Sample Provinces/Ports	All	Trading	All	Trading
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

Note: *p<0.1; **p<0.05; ***p<0.01

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)
aggASPL_{it}	0.041**	0.047**	0.019***	0.043**
	(0.017)	(0.024)	(0.006)	(0.019)
Sample Provinces/Ports	All	Trading	All	Trading
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

Note: *p<0.1; **p<0.05; ***p<0.01

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
dLogOrigPR$_{ijt}$	0.828***	0.752***	0.831***	0.762***
	(0.053)	(0.117)	(0.053)	(0.113)
dLogOrigPR$_{ijt}\times$ Connection$_{ijt}$			-0.125	-0.110
			(0.188)	(0.188)
Year Fixed Effects	Yes	Yes	Yes	Yes
R-squared	0.645	0.641	0.645	0.643
Observations	229	229	229	229
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 11
Results: Internet Access and Price Pass-through Using Aggregate-ASPL

	Dependent variable:		
	Change in log destination price ratio		
	(1)	(2)	(3)
	2SLS	2SLS	b7
dLogOrigPR$_{ijt}$	0.764***	0.682***	0.710***
	(0.113)	(0.113)	(0.115)
dLogOrigPR$_{ijt}$ $\times$ Connection$_{ijt}$	-0.726		
	(1.996)		
dLogOrigPR$_{ijt}$ $\times$ aggASPL$_{ijt}$	0.105		
	(0.338)		
dLogOrigPR$_{ijt}$ $\times$ Connection$_{it}$		1.483**	0.847
		(0.631)	(0.631)
dLogOrigPR$_{ijt}$ $\times$ aggASPL$_{it}$		-0.271	-0.160
		(0.190)	(0.194)
dLogOrigPR$_{ijt}$ $\times$ Connection$_{jt}$		1.031	
		(1.562)	
dLogOrigPR$_{ijt}$ $\times$ aggASPL$_{jt}$		-0.626	
		(0.569)	
Year Fixed Effects	Yes	Yes	Yes
R-squared	0.643	0.665	0.653
Observations	229	229	229
Note:			*p<0.1; **p<0.05; ***p<0.01

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 ASPL	0.028**	0.039**	0.043**
...with 75th percentile ASPL	0.025**	0.025**	0.028**
...with 50th percentile ASPL	0.027**	0.026**	0.026**
...with 25th percentile ASPL	0.027**	0.026**	0.050**
...with 5th percentile ASPL	0.300	0.179	0.428

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

Table 13
Results: Internet Access and Farmer Search Patterns

	Dependent variable:			
	Farmer seached for trade			
	(1)	(2)	(3)	(4)
$\text{logQuantity}_{iymc}$	0.017***	0.024***	0.028***	0.114***
	(0.001)	(0.003)	(0.002)	(0.044)
$\text{InternetAccess}_{pym}$		0.112***		
		(0.025)		
$\text{logQuantity}_{iymc} \times \text{InternetAccess}_{pym}$		−0.026***	−0.042***	
		(0.004)	(0.003)	
$\text{logQuantity}_{iymc} \times \text{pct05\_24}_{pym}$				−0.081*
				(0.046)
$\text{logQuantity}_{iymc} \times \text{pct25\_49}_{pym}$				−0.128***
				(0.044)
$\text{logQuantity}_{iymc} \times \text{pct50\_74}_{pym}$				−0.109**
				(0.044)
$\text{logQuantity}_{iymc} \times \text{pct75\_94}_{pym}$				−0.146***
				(0.044)
$\text{logQuantity}_{iymc} \times \text{pct95Plus}_{pym}$				−0.129***
				(0.045)
Prov.-Comm.-Year FE	No	Yes	No	No
Commodity-Month FE	No	Yes	No	No
Prov.-Comm.-Year-Month FE	Yes	No	Yes	Yes
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

Note: *p<0.1; **p<0.05; ***p<0.01

References

Allen, Treb. "Information Frictions in Trade." *Econometrica* 82 (2014): 2041–2083.

Chen, Xi and William D. Nordhaus. "Using luminosity data as a proxy for economic statistics." *Proceedings of the National Academy of Sciences* 108 (2011): 8589–8594.

Clarke, George R. G. and Scott J. Wallsten. "Has the internet increased trade? Developed and developing country evidence." *ECONOMIC INQUIRY* 44 (JUL 2006): 465–484.

Da Lozzo, Giordano, Giuseppe Di Battista, and Claudio Squarcella. "Visual discovery of the correlation between BGP routing and round-trip delay active measurements." *Computing* 96 (2014): 67–77.

Doan, Trinh Viet, Vaibhav Bajpai, Jorg Ott, and Ljubica Pajevic. "Tracing the Path to YouTube: A Quantification of Path Lengths and Latencies Toward Content Caches." *IEEE communications magazine* 57 (2019): 80–86.

Fan, Qin and Vania B. Salas Garcia. "Information Access and Smallholder Farmers Market Participation in Peru." *Journal of agricultural economics* 69 (2018): 476–494.

Freund, Caroline L. and Diana Weinhold. "The effect of the Internet on international trade." *Journal of International Economics* 62 (January 2004): 171–189.

Grimes, Arthur, Cleo Ren, and Philip Stevens. "The need for speed: impacts of internet connectivity on firm productivity." *Journal of productivity analysis.* 37 (2012): 187–201.