JRC/IPTS Digital Economy Working Paper 2015-XX

Revised Draft – June 2015

International Trade in Online Services

Georgios Alaveras and Bertin Martens[1]

Abstract:

This paper presents an innovative database of domestic and bilateral online services trade between 39 countries, including the US, the EU and some emerging market economies. It combines monetized and "free" online services in a single measure based on the volume of page views on websites of online service providers. We find that the online services market is geographically very fragmented. Less than 1% of all online service providers export worldwide and account for almost half of worldwide online services trade; they are mostly US-based. In the EU and other regions the share of online services imported from the US is very substantial. Conversely, in the US 32% of its online services providers export and these exports account for nearly twice as much as domestic demand. Application of the well-known gravity model of trade shows that trade frictions from geographical distance are greatly reduced in online services. However, cultural and linguistic borders are reinforced and home bias is stronger online than offline. We explain this paradox in terms of online information cost reduction and consumers' quest to explore the longer tail of online supply, both at home and worldwide. Larger firms export to more markets than smaller firms. It is easier for larger online firms to reduce distance and language related trade costs; however large firms do not reduce strong online home bias. Trade costs and home bias vary considerably across services sectors though, in principle, all online services are fully tradable. The export performance of online firms is driven mostly by the comparative advantages of their home country, more so than by their own competitiveness. We conclude with some suggestions for further research.

[1] Both authors are researchers at the Institute for Prospective Technological Studies (IPTS), one of the Joint Research Centres of the European Commission. The views and opinions expressed in this paper are the authors' and should not be attributed to the European Commission.

1. Introduction

There is a fast-growing volume of research on online trade, mostly focused on e-commerce or monetised online transactions. It covers a wide variety of micro-economic issues, ranging from advertising and ad auctioning mechanisms to online pricing, search ranking, consumer behaviour, etc. Relatively little of that research takes a more aggregate macro-perspective on online trade and tries to go beyond monetised transactions to examine the patterns of domestic and cross-border online services trade, partly because there are as yet few data sets on cross-border online trade.

Online services are often presumed to operate in a seamless and flat global market, a view enshrined in the "death of distance" hypothesis (Cairncross, 1997). However, empirical evidence has invalided this hypothesis. Considerable geographical segmentation exists in online services markets. Research by Blum & Goldfarb (2006), Hortaçu (2010), Lendle et.al. (2013), Cowgill et.al (2013), Gomez et.al. (2014) and Hui & Sundaresan (2015) shows that segmentation factors observed in traditional offline trade, such as geographical distance, consumer home bias and shared borders and language, remain important in online trade. There may be differences in the magnitude of the drivers of segmentation in offline and online market. While distance-related trade costs are generally lower online, cultural proximity and home bias may sometimes induce stronger segmentation effects online, especially for cultural goods like digital media (Gomez et.al., 2014; Aguiar & Waldfogel (2014); Gomez, Martens & Waldfogel (2015) and Gomez & Martens, 2015). Except for Blum & Goldfarb (2006), all these studies use data on monetized online transactions.

In the absence of official statistics on online services trade, most studies use company-specific data sets and focus on particular sub-sectors of online services trade. The present study takes a more comprehensive view. We use commercial data sources on website traffic to reconstruct the volume of domestic and cross-border bilateral web page views as a proxy for online services trade flows[2]. The data should be representative for around 90% of observed online services trade and includes both monetized and free(mium) services. We apply several trade models, including the gravity model and a firm level trade model, to explain the observed trade flows and market segmentation. Our contribution consists of a quantitative measure of the extent of worldwide geographical fragmentation in online services trade and first steps towards understanding the drivers of this fragmentation.

We equate online data exchange between a provider (a website) and a user with online services trade. Websites can only deliver bytes of information to the user. For some types of online services these bytes constitute the final service delivery to the consumer, for instance for digital media, search engines, etc. Other types of online services require physical delivery of the final product, for instance in the case of e-commerce in goods and travel services. You can order your pizza online but you cannot eat it online. We consider any online information flow between a content provider and a user as an online service, whether monetised (paid) or free. Earlier papers also used website traffic as a proxy for trade in online services (Blum & Goldfarb, 2006; Freund & Weinhold, 2000). Free services also have economic value. They contribute to

[2] In this paper we use the terms "websites" and "online services" as equivalent. We also use "categories" and "services (sub-) sectors" as equivalent. We mix internet jargon with more traditional services trade jargon.

consumer surplus (Goolsbee & Klenow, 2006; Pantea & Martens, 2014), generate production cost and possibly ad revenue for producers and advertisers. Limiting online services trade data to monetized exchanges would create a very incomplete picture.

This study focuses in particular on online services trade in the EU. For historical reasons, there are many cultural, linguistic, political, institutional and regulatory differences between EU Member States that result in a geographically segmented market. Moreover, the EU combines this "natural" state with a long-running policy experiment that seeks to overcome this segmentation. From its birth in 1957 the main policy goal of the EU has been to remove these barriers to trade and create a Single Market for goods, services, capital and labour. In recent years, this policy goal has been transposed to online services and the creation of a Digital Single Market.

We find that less than 1% of all service providers serve all country markets but account for almost half of all online services trade; they are mostly US-based. In the EU we find that about 42% all online services consumption is domestic. About 54% of online services consumption is imported from the US; the remainder are imports from the rest of the world (4%). About two thirds of all EU online services suppliers do not operate in more than 4 countries and account for about a third of all trade volume. In contrast to the US, the top-1% EU providers generate only 5% of all online trade. Similar patterns prevail in other regions in the world outside the US. The internet is both local and global: a large number of highly diversified local online services websites generate relatively little trade and a small number of truly global giant services providers account for the bulk of all trade. Demand for online services in larger countries is relative more inward-focused compared to smaller countries – as predicted by (offline) international trade models. The US is the big exception: 32% of its domestic online services providers export and these exports account for nearly twice as much (189%) as domestic demand in the US. This confirms the US position as the dominant supplier of worldwide online services. Apart from the US exception, smaller economies engage relatively more in online cross-border trade than larger economies.

We apply several trade models to the data. The gravity model concludes that the importance of geographical distance is greatly reduced online. However, cultural distance and home bias are stronger online than offline. Following Berthelon & Freund (2004) we explain this paradox as the consequence of information cost reduction and the consumer's quest to explore the longer tail of online supply at lower cost, both at home and abroad. At firm level, we confirm that larger firms and firms that are more competitive at home are also likely to export to more markets. Traditional country level comparative advantage trade models contribute little to explaining the observed trade patterns. Observed tradability varies considerably across online services sectors though, in principle, all online services should be tradable unless legal, regulatory or commercial trade barriers would intervene.

This paper is structured as follows. Section 2 explains the data sources and treatment. Section 3 shows some descriptive statistics. Section 4 presents estimates for the gravity and firm level trade models. Section 5 tentatively concludes and offers some suggestions for further research.

2. The data

Some authors have tried to reconstruct international information flows using IP level data. Mandel (2014) uses Cisco statistics on international data package flows to measure and compare the data intensity of economies. These data are collected at the internet protocol (IP) level. The data trade patterns that they reveal are biased, for several reasons. First, geographic patterns are distorted by the location of server farms and undersea cables. For instance, Canada stands at the top of the list as the most data-intensive country because it is the home of Cisco and many of its server farms and transatlantic undersea cables make landfall in Canada. Second, these traffic data say very little on domestic and cross-border activity on the internet. Data replication between server farms will blur that picture. Because of the architecture of the IP, it is not possible to track end-to-end or origin-destination traffic at this level. The IP splits data in packages that may follow different routes to reach their destination.

Origin-to-destination internet traffic can only be measured at the application level or the online services delivery level. In this research we focus on online services that are delivered through websites that can be accessed via browsers. This excludes online services delivery through apps and other specific software that is not accessible via browsers, for instance corporate data traffic between server farms and cloud computing services. According to Cisco (2014)[3], global IP traffic reached about 62 Petabytes per month in 2014, of which nearly a quarter in Europe. About two thirds of this was fixed internet traffic. The rest is classified as "managed IP", mostly corporate networks, TV/VoD signals and non-browser applications, and mobile traffic (about 3.5%). Our study covers essentially the first category of fixed IP traffic.

Several companies collect internet traffic statistics at the internet application level between internet users and websites. The simplest methods consist of tracking the IP address of incoming traffic by means of in-site tools. Website operators build these tools into their websites to enable them to track the origin of the incoming traffic. The Google Analytics and Amazon Alexa tools for example are widely used. Incoming traffic trackers can be mirrored in out-going traffic trackers build into the browsers of internet users. The Alexa Toolbar for instance can be voluntarily installed by internet users and collects data on their clickstream. Some companies manage online consumer panels of internet users who agree to install an in-browser tracker to collect data on their online user patterns. Many marketing companies operate such panels, for instance Nielsen NetRatings, Comscore and TNS. Besides these software tools installed by users on the supply and demand side, a myriad of cookies enable third parties not directly involved in the exchange between users and websites to track activity on the internet and collect this information, for advertising and other commercial purposes. More sophisticated methods have been developed in recent years that do no longer rely on cookies and user-installed software. Some companies have made arrangements with ISP providers to harvest directly the clickstream generated by ISP clients.

All these data collection methods have pros and cons. Consumer panel data are very detailed and provide a consistent and continuous picture at the user level, together with socio-economic profile data of these users. That makes them suitable for consumer analysis. However, they

[3] Cisco "Visual Networking Index", June 2014, http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf

usually cover no more than a few thousand internet users per country. They are expensive to maintain and therefore limited to the most important internet economies only. Less than half of all EU Member States have such online panel data. Data derived from website analytics, toolbars and cookies are more comprehensive in coverage, though they are still based on samples and do not cover the entire universe of internet activity.

In this study we use Amazon Web Services "Alexa" data that track country rankings of websites, page views by website and country of origin of the users. Page views are defined as the number of page views from country i to a website in country j per million page views by Alexa users over a rolling period of the last 3 months. We downloaded Alexa data for the 28 EU Member States and 11 non-EU countries (Norway, Switzerland, Turkey, Russia, US, Canada, Brazil, Australia, China, Japan and India. The data represent average traffic for the 3-month period from August to October 2014. Altogether we collected data on 651,000 websites[4] accounting for 895,000 PVs[5] per million or nearly 90 per cent of worldwide internet traffic as monitored by the Amazon Alexa survey data. Alexa collects data from a sample of internet users; it does not cover all internet activity. It does not provide any measures of the representativeness of its survey data. There is considerable debate in blogs and user groups on this issue, in particular with regard to country rankings which are the more commercially valuable data. We use the country ranking lists only to detect the use of a website in a country; the rankings as such are of secondary importance in our analysis. Alexa generates data on the number of PVs on a website by country from which the PV is made – based on the IP address of the user. The potential maximum number of cross-border trade observations in our sample of 651k websites and nearly 200 countries could reach some 130 million. The raw Alexa data contain only about 1.6 million non-zero country PV observations, about 1.2 per cent of the potential maximum.

In the offline economy, service trade is often classified according to the WTO GATS classification. For online services there is no officially accepted classification system yet. Some online services could be rolled back into the more traditional GATS categories but many are "new" online services that are not easily identifiable under the GATS classification. Here we use the McAfee categorisation system[6] (see Table 5 in Annex). A major advantage of the McAfee system it that it offers an online tool to classify websites into one of its 32 categories. We reclassified these into 7 more functional categories: commercial, media, news, personal, social, technical and a residual group of other online services.

[4] Obviously, the Alexa data do not cover all existing websites. According to Tekeye, there were 876 mln registered websites in the world in January 2015, a decline from over 1 bln in 2014. However many of these are never used. Still, the number of active websites will be much higher than the 651.000 recorded here by Alexa, though most of them will have hardly any significant volume of traffic. See http://tekeye.biz/2014/how-many-websites-are-there

[5] The sum of all PVs does not add up to 1 million because we collected data for websites used in 39 countries only. The country of origin of some of these websites may be outside the set of 39 and many will also be viewed in countries outside the 39. Our data capture bilateral traffic that either originates or is destined for one of the 39 countries, even if the other point of the bilateral pair is outside the 39 countries. However, they do not capture traffic where both origin and destination is outside the set of 39 countries. The sum of all PVs (895,000) shows that we are missing about 10% of worldwide PVs in our data.

[6] According to McAfee, the categorization of a particular URL uses objective standards and various technologies, including artificial intelligence techniques, such as link crawlers, security forensics, honeypot networks, sophisticated auto-rating tools, and customer logs.

In order to construct a bilateral trade matrix between suppliers (websites) and consumers (internet users), page views need to be allocated to country of origin (CoO) of the websites and the country of destination (CoD) or residence of the user. We define the CoO as the home country of the online service provider. Identification of the CoO of a website is a challenge. The information supply chain (leaving aside locations in the physical delivery chain, if any) may be in different locations: headquarters of the company where the entire system is managed, location of server farms where data are collected, processed and distributed, etc. Moreover, content provider(s) may be located all over the world.

Some authors have approached the CoO question by using the country IP address of the website. This can easily be detected for instance by using the WHOIS tool on the internet. However the IP address is not necessarily a reliable indicator of the CoO of the website. Some websites are hosted on servers in another country, especially now that cloud services become increasingly prevalent. Large websites will use many servers and IP addresses located in a variety of countries. Some websites may also be hosted in different jurisdictions for legal and tax reasons. The correlation between the results of our own CoO determination procedure (explained below) and the CoO according to the IP address of a website is only 62%.

We worked out a procedure to identify the CoO (see Figure 1 and Table 1). The first two steps use the location of the main audience to determine the country of origin of 94 per cent of the 651.000 websites for which we have obtained traffic data from Alexa:

- Websites that receive more than 50 per cent of all PVs from one country are assumed to be local websites in that country. In fact, by far the largest group of website are local online services whereby supplier and consumers live in the same country.
- Websites that receive twice as much PVs from the top-ranked country than from the second ranked country are allocated to the top-ranked country.

For the first two steps we did an additional check: If the country extension (for websites that have a country specific extension) clashes with the country allocated, we give preference to the country allocated over the country designated by the extension. For websites with a country extension we found that the allocated country matched with the country extension in about 90 per cent of all cases. This proves the reliability of the first two steps and justifies their application to website addresses without country extensions, i.e. .com and .org for instance.

We managed to increase this to 98 per cent in three additional steps, using the country extension of the website, the contact address of the website owner and the IP addresses of websites. The remaining 2 per cent of unallocated websites account for about 3 per cent of all PVs (see Table 1). As such, the CoO or location of a website is to some extent an artificial construction; it is the presumed location based mostly on observed traffic patterns and to a much lesser extent on servers and headquarters locations.

The Alexa country data contain a minor bias. Countries that represent less than 0.5 per cent of the total number of PVs on a website are aggregated by Alexa into a category "other countries". These unallocated PVs account for about 30k or 2 per cent of total PVs. Imputation of the unallocated residual to country pairs is done in two steps. First, we use Alexa country ranking

lists to detect whether a website is actually viewed in a particular country. If our PV number for a website in a country is zero while the website appears on the ranking list in that country we classify this as a false zero; if it does not appear in the rankings we assume it is a true zero. We detected about 45k false zeros. All false zeros are re-estimated into a positive number by means of an imputation procedure described in Annex. The imputation problem occurs mainly with large global websites that are used in small countries.

Another issue in the identification of the country of origin of the online service provider is how to deal with country specific subsidiary websites that providers may open in order to better cater to the needs and (language) preferences of users in these countries. For example, Google has many country versions of its search site and Amazon even has separate physical warehouses in some countries. This brings us to modes of delivery in services trade. Under the WTO General Agreement on Trade in Services (GATS) there are four modes[7]:

- Mode 1 is pure cross-border trade delivery of services whereby the producer and consumer remain in their respective countries. The service is brought to the consumer.
- In Mode 2 the consumer moves to the producer's location in order to consume the service. For example tourism.
- Mode 3 involves a commercial establishment in the consumer's country. Sales by that establishment are counted as exports from the service provider's home country to the consumer's country and recorded in Foreign Affiliate Trade Statistics (FATS). An example is the local subsidiary of a foreign bank.
- Mode 4, the temporary movement of workers from the producer to the consumer country to deliver a service. An example is short-term consultancy services.

Modes 2 and 4 are less relevant for online services trade since they involve the physical movement of persons. The nearest online equivalent to Mode 2 could be the use of VPN or proxy servers that enable a consumer to move his IP address to another country in order to access an online service. We do not cover the use of proxy servers here.

The bulk of online services are delivered in Mode 1. However, there is a significant part of Mode 3 subsidiary services delivery as well. According to the GATS definition, an Amazon warehouse located in another country should be counted as a commercial establishment and its sales constitute foreign affiliate trade in services (FATS) by the home country, in this case the US. However, the GATS definition is not applicable for example to the Apple iTunes[8] country stores in each EU Member State; iTunes has no physical presence in these countries. The physical presence requirement in GATS Mode 3 may therefore be a bit outdated in the case of online services. Here we interpret Mode 3 more broadly as an online presence in a country that is clearly distinguishable from the home country website.

Accounting for Mode 3 online services delivery requires aggregation of different country web domains that belong to the same online service provider into a single domain. This is fairly easy

[7] A more detailed description is available on the website of the World Trade Organisation at http://www.wto.org/english/tratop_e/serv_e/cbt_course_e/c1s3p1_e.htm

[8] In any case, Apple iTunes country extensions come after the first slash in the website address. Alexa data only report until the first slash.

for well-known internet giants like Google, YouTube, Yahoo, Amazon and eBay[9]. Beyond that, it becomes complicated. In order to keep the aggregation manageable, we checked the Top-1000 websites used in the EU for subsidiary websites. The Top-1000 accounts for about two thirds of all page views in the EU. We found 114 subsidiary websites that account for 42% of all page views in the Top-1000. However, 39 belong to Google, YouTube, eBay, Yahoo and Amazon and they account for 95% of all page views on subsidiary websites. Moreover, subsidiaries are mainly an issue for top ranking websites. Two thirds of all subsidiaries occur in the upper half of the Top-1000 and they account for 97% of all Mode 3 page views. We conclude from this that accounting for these four firms' subsidiaries covers the bulk of all Mode 3 online services trade though only a minority of Mode 3 websites. It gives us a good approximation at the intensive margin (page views) but will still leave us with a somewhat distorted picture at the extensive margin (websites). Clearly, this is an approximation that can be improved by going further down the long tail. This would however require substantially more work. For the sake of transparency, we present two versions of the world online trade tables in Annex, with and without Mode-3 re-assignment. The descriptive statistics and analytical work in the next sections is based on Mode-3 aggregated data, except when otherwise mentioned.

3. Some descriptive statistics

Table 2 presents a summary worldwide online services trade at the level of regional blocks. Table 3 presents country specific figures for the EU28 and trade between the EU and the US.

As expected, US online services suppliers dominate the scene. About half of worldwide PV traffic originates from US-based websites. Another striking feature is that about two thirds of all online services traffic is domestic within these regions (including US domestic traffic). Apart from US online services there is relatively little extra-regional cross-border traffic. The picture is very similar for the EU28. About 42% all online services consumption in the EU region is domestic and 54% is imported from the US; the remainder are imports from the rest of the world (4%). Less than 1% of all online service providers cover all EU Member States but account for almost half of all online services trade. Domestic service providers still hold a majority share in FR, DE, IT, PL, ES and the UK. The US is a dominant online services provider in nearly all smaller country markets.

Figures 2A and 2B decompose the geographic distribution of online services trade in the EU in an extensive and intensive margin. Fig 2A show the geographical distribution of all websites used and page views made by users in the EU28. The horizontal axis shows the number of Member States where a website is used. The L-shaped blue line (distribution of website use) indicates that the majority of websites are for domestic or neighbouring country use only. Relatively few operate cross-border. The more U-shaped red line (distribution of page views) on the other hand shows that traffic on domestic sites is much weaker than on a few very large EU-wide service providers. About two thirds of all online services used in the EU do not cover more than 4 countries and account for about a third of all page views only. At the other extreme, less

[9] However, we may not want to aggregate across different types of services owned by the same company, for example Amazon shops and Amazon Web Services. In some cases this would actually require further disaggregation of web domains, for example to separate Google Search – search engine services - from the Google Play store – digital media services. This is not possible with the Alexa data.

than 1% of all websites used in the EU are truly Digital Single Market operators; however they account for almost half of all online services traffic.

Figure 2B does the same but only for websites whose country of origin is in the EU28. The difference between Figures 2A and 2B is striking. The EU has very few online service providers that are truly pan-European operators and no big players. About 85% of all EU-based websites operate in no more than 4 Member States and account for nearly three quarters of all page views on EU websites. The vast majority of EU website traffic is mostly local or limited to countries that share a border and/or a language. At the other end of the distribution, the top-1% online service providers in the EU account for about 5% of all traffic. The sharp surge in page views on truly EU-wide service providers on the right hand side of Figure 2A is caused almost entirely by non-EU (i.e. US-based) service providers. Figure 2C extends this to all websites used in all 39 countries covered by our dataset. It repeats the pattern of Fig 2A and indicates that large US online service providers dominate the scene in all these countries. The cross-border internet use pattern in the EU does not diverge significantly from the global pattern.

These figures show that the internet is both local and global. It is local in the diversity of online services offered and global in the concentration of demand on a few giant online service providers.

This overall bilateral matrix can be decomposed in an extensive margin (the number of websites) and an intensive margin (PVs per website) (see Table 4). The table splits the number of websites used in each country into domestic & foreign websites. It also identifies the number of domestic websites that attract foreign page view traffic. EU Member States are ranked according to the intensity of domestic versus foreign website use.

Larger countries have more domestic websites and also use more foreign websites than smaller countries. Supply and demand for variety increases with total demand. However, demand for online services in larger countries is relative more focused on domestic websites compared to smaller countries, except in the US – a very big exception. This holds both at the extensive (#sites) and intensive (#PVs) margin. This seems to confirm a finding from the offline trade literature that larger economies are relatively less external trade oriented than smaller economies – again with the big exception of the US.

Table 4 also shows foreign demand (exports) for domestic online services in each country. The US sits at the top of that ranking with 32% of its domestic online services providers exporting to other countries. Exports account for nearly twice as much (189%) as domestic demand in the US. This confirms the US position as the dominant supplier of worldwide demand for online services. Other major economies such as China, Japan, Russia and India also have a high share of domestic websites attracting foreign demand. However, at the intensive margin this share is considerably weakened.

A standard finding in the offline trade literature is that larger economies are relatively more inward-looking and smaller economies relatively more foreign trade oriented. This finding is replicated on the import side of online trade: the correlation between the size of the online economy (measured by total PVs and websites) and the number of domestic PVs and websites is strongly positive: +0.6 for PVs and +0.62 for websites. On the export side however the picture

is more nuanced: there is a positive correlation between the number of domestic websites and the number of sites that export (+0.42 in the EU28) but a slightly negative correlation between number of domestic and exported PVs (-0.1 in the EU28).

The extent of cross-border traffic is likely to vary by type of service. Table 5 presents the ratio of domestic to cross-border activity by category of online services – for all online services used in the EU28. Ratios are calculated for the number of websites and page views. We also calculate a simple average of both ratios. They are strongly correlated (+0.63). As expected, less traded and more domestic market focused types of online services include restaurants, real estate services, government services, job search and health services. Surprisingly however, online shopping, merchandising, travel and finance & banking are also poorly traded and predominantly domestic services. The more traded online services include games, news, adult content, media, social networking and file sharing. Search engines are heavily skewed towards cross-border tradable services because of the dominance of Google Search in this category. Less traded does not necessarily imply less tradable however. Facebook, Uber and many other online companies have amply demonstrated that intrinsically local services can be traded on global platforms. In principle, the information component of any service can be brought online and traded globally. Many barriers may still stand in the way however of actually doing so.

4. The gravity model

The descriptive statistics above give a first picture about what is going on in online trade in services. We may already deduct intuitively some of the drivers of online trade. However, we may want to ask more precise research questions, such as: What are the drivers and impediments to cross-border trade in online services? What are the characteristics of exporters of online services and what distinguishes them from firms that do not export? What are the sources of trade costs in online services? We need trade models to answer these questions. In this section we apply the well-known gravity model of trade to country and services sector data. In the next section we move to firm level trade models.

The gravity model has already been applied to online trade (Blum & Goldfarb, 2006; Hortaçsu, 2010; Lendl et al., 2012; Cowgil et al, 2013; Gomez et al, 2014). It explains the volume of bilateral trade between two countries as a function of country characteristics and the cost of doing trade between them. We use the Anderson & Van Wincoop (2003) version of the model that was derived from a consumer demand system based on a CES utility function and add country of origin and destination fixed effects following Feenstra (2002), and sector specific fixed effects when we run it for different services categories. The explanatory variables for trade costs include language preference, home bias or preference for home market products and physical distance as a more general measure of trade costs. We apply the gravity model at the extensive (websites) and intensive (page views) margin of trade. The standard gravity equation looks as follows:

$$Q_{ij} = a + b\ ldist_{ij} + c\ comlang_{ij} + d\ home_{ij} + e\ Country\ FE + f\ Category\ FE + error\ term$$

where

Q_{ij} = traffic[10] (number of websites or page views) from country i to country j
$ldist_{ij}$ = log of the geographical distance between countries i and j
$Comlang_{ij}$ = dummy=1 if i and j share a language, or as a continuous variable of language distance
$Home_{ij}$ = dummy variable with value=1 if country and origin are the same
CFE = importer and exporter country and category fixed effects dummies

In the case of physical trade in goods the distance variable is usually interpreted as a proxy for transport costs and other sources of cross-border trade costs such as tariffs and regulatory barriers. While some online services such as e-commerce may involve physical transport costs most online services only transfer electrons, with virtually zero transport costs. Blum & Goldfarb (2006) already demonstrated that even for immaterial online services geographical distance is not dead. They interpret the distance coefficient as a measure of cultural distance between consumers and suppliers that creates trade costs. Here we interpret the distance coefficient as a proxy for the combined effect of these three sources of trade costs: consumer preferences, regulatory barriers and supply side barriers. Language is the most obvious cultural barrier in B2C online trade where consumers have to communicate directly with the supplier. Sharing a language facilitates trade. A simple solution is to introduce a dummy that takes value 1 if the official languages of two trade partners are identical. Melitz & Toubal (2012) provide six continuous measures of language distance between country pairs. We experimented with these but find little difference in the outcomes (Table 9). In line with the methodology applied by Pacchioli (2011), McCallum (1995) and Wolf (2000), we introduce a dummy variable for domestic trade observations in the gravity model. This is an indicator of home bias or consumer preferences for domestic over foreign products. Home bias determines to what extent consumers shop abroad; distance and other trade costs variables determine how far and to which countries they go when they go abroad.

Table 6 presents the OLS estimation[11] results for the country level gravity model. All coefficients are significant and have the expected signs. A first important finding is that consumers prefer to shop at home. There is strong home bias in consumer demand for online services: users are 110 times ($e^{4.704}$) more likely to click on a website in their own country than on a foreign website. Home bias is still high at the extensive margin: a website is 28 times more likely to be used (at least once) at home than in another country. Balta & Delgado (2007) arrive at similar estimates for home bias in trade in services among OECD countries (129), considerably higher than for goods (11.5). They attribute this to the fact that services are intrinsically less tradable than goods. Van der Marel & Shepherd (2013) argue that all services are in principle tradable in at least one of the four GATS modes. However, they find significant differences in the probability of a service being traded by countries and service sectors. Finally, home bias in online services trade is also higher than in online goods trade (e-commerce in goods) where Gomez et.al. (2014) find that users are 16 times more likely to buy at home. For comparison, Pacchioli (2011) estimates that EU consumers are between 7.4 and 24 more likely to

[10] For the Tobit and OLS regressions, the dependent variable is transformed as log(1+Q). For PPML we take the level of Q.
[11] We experimented with Tobit estimators but they produced very similar results since only 0.5% of all 39 x 39 = 1521 bilateral trade observations have a zero value.

buy goods at home than in any other EU Member State; US consumers are between 2.6 and 7 times more likely to buy goods in their home state than in another US state.

When users decide to shop abroad, the distance and language coefficients tell us where and how far they go. The distance effect is negative: a 1% increase in geographical distance between the country of the user and the country of origin of online services supplier decreases traffic volume at the intensive margin by about 0.35%. It reduces the number of consumers (at the extensive margin) interested in clicking on a foreign website by 0.55%. Suppliers in countries that share a language with the user country will see cross-border traffic increase by about 70% ($e^{0.348}$). In an online B2C environment users will only use websites that they understand. Note that 32 of the 39 countries in the sample have their own language.

How does this compare with other online and offline trade in services studies? To the best of our knowledge, the only other study that estimates a gravity model for online services is Blum & Goldfarb (2008). They find a distance effect of -3.25% for cultural taste-dependent goods such as media products, and a zero effect for other online services. The distance effect becomes much weaker for several other specifications of their gravity model. Their home bias effect is very weak (factor 4). This may be due to the fact that the study considers only US demand for online services. The geographical remoteness of the US from most of its trade partners may explain the strong distance effect in this study. Their dataset consists of a single row and not a square bilateral trade matrix. We should therefore be cautious in comparing the results of the present study with those of Blum & Goldfarb. There are also several studies that estimate the distance effect in offline trade in services. Kimura and Hyun (2006) estimate the distance effect for offline services between -0.6 and -0.7, somewhat stronger than for goods (around -0.5). Walsh (2006) finds a similar value of -0.7 for services trade using an OLS estimator. We conclude that our distance estimate is the lowest of all, both for online and offline services trade. The lower value suggests that digital technology has indeed made it easier for consumers to search further away for the services they prefer, once they decide to go cross-border.

How does this compare to the distance coefficient for trade in goods? The distance coefficient for online services is lower than for online trade in goods (e-commerce), estimated at -0.89 for the EU by Gomez et.al. (2014), lower than for offline trade for the same basket of goods (-1.349). Lendle et.al. (2012) find a similar reduction in the distance coefficient between offline and online trade in goods. Bertelon & Freund (2004) find an average distance coefficient around -1.2 to -1.4 across trade in goods in all industries for the period 1985-2000. Disdier & Head (2004) find a weighted mean distance coefficient of -1.1 across more than one thousand gravity studies on trade in goods. We conclude that the distance effect in online services trade is much lower than in both online and offline trade in goods. Digitization of services trade has reduced search and delivery costs and enables consumers to buy much further away.

How can we explain these findings? At first sight, the strong increase in home bias – compared to offline trade – combined with the decline in the distance effect may seem paradoxical: consumers stay more at home in their overall consumption of online services and at the same time go further away in their cross-border consumption of online service. In the offline international trade literature, Bertelon & Freund (2004) note that, despite the rapid growth in

world trade, the effect of distance on (offline) trade in goods has increased. Increased online variety stretches the long tail not only vertically into the product rankings but also horizontally across geographical distance: consumers can go further to find the exact product variety that they are looking for. At the same time, increased price competition online implies that they are likely to find the same product closer to home at lower prices and/or lower transaction costs. This may explain the observed paradox in the online services trade data: home bias increases while cross-border trade costs decline. Consumers find more online variety and price competition in their domestic market – compared to offline search – and therefore consume more at home. However, when they do go abroad online they can go much further than in offline search because of the dramatic reduction in information costs. Still, language and culture differences generate transaction costs that are likely to be lower in neighbouring countries.

Table 7 presents the distance, language and home bias coefficients by service category that have been produced by separate runs of the OLS gravity model regressions for each category, ranked by the descending value of the home bias coefficient. We used importer and exporter country fixed effects in the regressions. Despite the fairly high number of zero observations (61% of 79k observations), coefficients and category rankings do not change significantly when we switched from OLS to PPML (Santos & Tenreyro, 2006) and Tobit estimators. We therefore stick to the OLS results.

We are interested in relative differences in the distance, language and home bias coefficients across categories of online services. The home bias coefficient is positively correlated with the language coefficient (Spearman rank correlation +0.48, Pearson ordinary correlation +0.52). This indicates that online service categories with a high consumer preference for home markets are also sensitive to language preferences. The rank correlation between the home bias and distance coefficients is much lower (Spearman and Pearson between +0.1 and +0.2).

As expected, we find online services that are more focused on local markets in the top group: restaurants, classified ads, health, job search, real estate and banking, politics. At the other end, gambling, games, software services, adult services, file sharing, travel, etc. are the more globally used services. News, media streaming and shopping are somewhat in the middle of this ranking. Some online services appear to be inherently more difficult to trade across borders than others, despite the fact that digital technology makes information about these services available anywhere in the world. Several factors may play a role in tradability. Some online services require physically delivery of services: restaurants, classified, job search, real estate. You can order you pizza online but you cannot eat it online. Others are linked to local culture and language environments: social networking, politics. As noted by Blum & Goldfarb (2008), only culturally neutral online services travel more easily across borders, provided there are no other obstacles. Van der Marel & Shepherd (2013) argue that the tradability of services is affected by several factors of which trade costs is only one. Technology and the more Ricardian and Hecksher-Ohlin drivers of trade such as productivity and factor endowments will also play a role (see below).

Gravity models have an important limitation: they allow only one country specific variable, usually a dummy for country fixed effects. Disaggregation of country fixed effects may produce more information. That needs to be done outside the gravity model. Table 4 shows that there is

considerable variation across countries in the supply (number) of domestic websites, domestic demand (PVs) for domestic and foreign websites and foreign demand for domestic websites. In line with the findings from the offline traditional trade literature, small countries have less variety of supply than larger countries. Consumers in smaller countries with a preference for variety will tend to look relatively more at foreign websites, though language barriers are a source of trade costs. Consumers in larger countries will more domestic variety of supply will be more inward looking. Language barriers may amplify this. The result of all this is that smaller economies trade relatively more with the outside world than larger countries. In order to quantify these effects we ran some country level regressions on the data from Table 4 – see results in Table 8. The dependent variable in the regression is the ratio of foreign to domestic website use, either in terms of number of websites (extensive margin in the first column) or PVs (intensive margin in the second column). As predicted, the coefficient on the population variable is negative: larger countries use relatively less foreign websites. A 1% increase in population reduces demand for foreign online services by 0.6%. The intensity of use of foreign websites increases slightly with the percentage of English speakers in a country. Somewhat more puzzling is the negative coefficient at the intensive margin for the percentage of internet users in a country. As the internet become more widely used users become more inward looking and less interested in foreign websites. More intensive internet use may increase the domestic supply of online services and thereby induce a relative decline in demand for foreign websites. Income (GDP per capita) has no impact on the outward orientation of consumers. Since we have only 39 observations in this regression we could not stretch the analysis much further.

5. The trade performance of online firms

The Alexa dataset contains cross-border trade information by website or online firm: the number of export markets served by an online firm/website and the intensity of trade as approximated by page views. We use this information to examine how export performance varies by firm size and to estimate the contribution of different factors to the export performance of an online firm.

Figure 3 shows how export performance at the extensive margin (number of markets) varies by firm size. We classify online firms or websites in size percentiles (in 5% intervals except for the top 1%), based on the total number of PVs that the websites receive. Export market reach is measured as the average number of countries that firms in a size percentile export to. If all firms in a size category would export to all 39 markets in the sample than the reach indicator would be 100%. As Fig 3 shows, export market reach remains below 5% up to the 80th percentile and climbs to 30% for the last percentile only. This outcome is probably not very different from export market reach of offline firms. It indeed it is a well-known fact that very few offline firms export and only the larger tend to do so because they can amortize the fixed costs that come with exploring and setting up a business in new markets. Online firms also face foreign market set-up costs: translating websites into the language of the destination country and possibly adapting the (cultural) content to that market, adapting to regulatory requirements in that new market, setting up a physical delivery network if need be, etc.

In Table 10 we run the gravity model separately for 7 size groups of firms as measured by their total worldwide PVs. The dependent variable is the volume of exports (PVs in foreign markets) by firm (the intensive margin of trade). We apply importer, exporter and category fixed effects.

We observe a gradual decline in the distance and language coefficients across firm size until they becomes essentially zero for the upper percentiles. Larger online firms have less problems to overcome the trade costs related to geographical and cultural distance. They can more easily adapt to consumer preferences and regulatory requirements in foreign markets. The home bias coefficient declines in the lower percentiles but remains more or less stables across the rest of the size distribution. Even larger firms have to face the consumer preferences for home market products and cannot change these preferences.

The export performance of online services firms can be driven by several factors. First, in line we the findings from "new" firm trade theory (See for example Melitz, 2003 and Melitz & Redding, 2012) the firm's size and productivity are important drivers of export performance. Larger and more productive firm are more likely to be exporters and variations in trade performance are to a large extent taking place at the extensive margin (Eaton et al, 2004; Bernard et al, 2011), i.e. the number of firms exporting rather than the volume of trade by firm. "Old" trade theory puts emphasis on the overall comparative advantages of a country in a particular product or services sector. This overall advantage may also play a role at firm level since firms are often part of a network or cluster of firms that produce complementary services for a particular sector. Finally, sector specific issues may play a role. Some services are inherently more tradable than others (Van Der Marel & Shepherd, 2013; Gervais & Bradford Jensen, 2013). Digital technology and the internet may reduce trade costs for some services sectors more than for others. The hypothesis that we want to test is to what extent a firm's online export performance is affected by the country of origin's comparative advantage and the tradability of the services that it produces. We apply the following model to test the explanatory value of each of these trade theories in an online services setting:

$$\log N_i = a + b\ log RCA_{cj} + c\ \log MarketShare_{ij} + d\ Category\ FE + error\ term$$

$$RCA = (Exports_{kc} / Exports_{c}) / (Exports_{kw} / Exports_{w})$$

Where:
N = the number of export markets serviced by website i in service category k
RCA = the revealed comparative advantage for service category k in home country c
MarketShare = website i 's market share in its home country market in service category k
Category FE = a dummy variable for category k
The subscripts k, c and w stand for categories, countries and worldwide (all countries)

A priori we expect to get positive signs on the RCA and market share variables: both a country's comparative advantage in a sector and the firm's own competitive position in the domestic market should help to drive its export performance. The coefficient on the category fixed effects variable will tell us something about the tradability of services. The sign may vary across categories. We have aggregated the original 52 categories into 7 groups for this regression model: commercial, media, news, personal, social, technical and "other" services. Commercial services are the reference point for the category fixed effects.

Comparative advantage is usually measured in terms of production factors, for instance by means of capital and labour intensity indicators for each sector. In the case of online services, we could also add ICT or digital skills and infrastructure intensity indicators. However, we do not have production factor intensity indicators for each of the services sectors in our dataset. Instead, we apply revealed comparative advantage indicators at the sector output level (Balassa & Noland, 1989). Two types of revealed comparative advantage measures can be constructed. The first is the ratio of a country's share in exports for a particular services sector to its share in total exports. The second is the ratio of a country's net exports in a particular sector (exports minus imports) over total trade (exports plus imports) in that sector. Both ratios have their strengths and weaknesses. The net exports index provides a more comprehensive picture because it takes into account both exports and imports. However, it is subject to variations in a country's overall trade balance. Since the US is by far the largest exporter of online services, the US trade balance in online services will affect the estimations. Even a small country can have a comparative advantage over the US in a particular online services sector although it exports far less in that sector (in volume terms) than the US.

The exports ratio is calculated at the intensive (#PVs) and the extensive (#websites) margin; the next exports ratio only at the intensive margin. Market share is a proxy variable for "new" firm level trade theories. A firm with a larger market share in its home market is likely to be more productive and competitive than its competitors and therefore more likely to export. Category fixed effects measure the tradability of a particular type of online services. The gravity model applied to service categories already revealed that home bias and distance-related trade costs vary considerably across categories. The category dummy is expected to produce similar results in this model.

The results of the estimation are shown in Table 11. The dependent variable in all cases is the number of export markets by firm. In the first panel of Table 11 we use the exports only index of revealed comparative advantage, calculated in terms of page views. In the first column, the sign of the RCA coefficient is negative when we run the regression for the full set of 39 countries; it remains negative in the second column where we take the US as reference. It turns positive in column 4 when we leave out the US or in column 3 where we restrict the regression to the EU28 countries. A negative coefficient is hard to explain. It would imply that a country's comparative advantage in a sector makes it harder for a firm in that sector to export. Clearly, the overwhelming dominance of the US distorts the picture so much at the intensive margin (page views) that the RCA coefficient turns negative. Columns 5 and 6 experiment with the export index model but applied to the number of exporting websites rather than the number of page views. This also produces the expected positive sign on the RCA variable. Column 7 uses the net exports index of revealed comparative advantage, again based on page views. That produces a positive sign and a considerably improved fit for the entire model.

A firm's domestic market share variable shows the expected positive signs in all versions: a firm's competitiveness in the domestic market drivers of the firm's export performance. Columns 4, 6 and 7 show a larger coefficient on the RCA variable than on the firm's own competitive position in the domestic market. This would lead to the conclusion that the economic environment in

the country of origin is a more forceful driver of export performance than the firm's own position.

The category fixed effects coefficients show that some online services are more tradable than others, though the ranking in terms of tradability is not very stable across the different regression models. In column 7, social and technical services are the most tradable, followed by media services. News and personal services are less tradable. Commercial services include most of the monetized services that require, in most cases, a physical counterpart to the online transaction. Distance-related trade costs will be much higher for this type of online services, which explains why they are less tradable than purely digital services. In principle, all online services are fully tradable, irrespective of distances. The Uber taxi app has shown that the purely information part of inherently local physical services such as taxi rides can be organised online on a cross-border scale; of course, the taxi ride itself will remain a very local service.

6. Conclusions

This paper presents an innovative database of domestic and bilateral online services trade between 39 countries, including the EU28 and other European countries, the US and the largest emerging market economies. It does not attach a monetary value to these trade flows but combines monetized and "free" online services in a single measure based on the volume of page views on the websites of online service providers. Services are classified by type, following standard categorizations available on the internet.

We find that about 42% all online services trade in the EU (in volume terms) is domestic and 54% comes from the US; the remaining 4% is cross-border trade between EU Member States. Two thirds of all EU online services suppliers do not cover more than 4 countries. Less than 1% of all service online export to all EU Member States; however they account for almost half of all online services trade. The top-1% EU providers generate only 5% of all trade. The dominant pan-European providers in the EU are mostly US-based. The patterns observed in the EU market do not diverge significantly from those observed at global level. The internet is both local and global: a large number of highly diversified local online services websites attract relatively little trade and a small number of truly global giant services providers account for the bulk of all trade. Demand for online services in larger countries is relative more inward-focused compared to smaller countries – as predicted by (offline) international trade models. The US is the big exception: 32% of its domestic online services providers export and these exports account for nearly twice as much (189%) as domestic demand in the US. This confirms the US position as the dominant supplier of worldwide online services.

Using the well-known gravity model of trade at country, sector and firm size level, we find differences in the relative magnitude of the drivers of online and offline services trade. Geographical distance causes considerably less trade costs online than offline. On the other hand, consumer preferences for the home market are much stronger online, though they vary across types of services. Once consumers decide to go outside their home markets, language becomes an important obstacle. Cultural and linguistic borders reinforce home bias. This is understandable in a B2C trading environment where consumers have to communicate directly with the service provider, compared to an offline international trade environment that is usually

routed through B2B wholesale channels. Following Berthelon & Freund (2004) we explain the paradox between declining trade costs and increased home bias as the consequence of information cost reduction and the consumer's quest to explore the long tail of supply, both vertically (in the home market) and horizontally (geographically). Online services are at the same time more local and more global than offline services. The findings from traditional trade theory hold: smaller economies engage relatively more in cross-border trade than larger economies. We also confirm the finding from the "new" firm level trade literature that larger firms are more likely to export than smaller firms. Larger firms manage to overcome the trade costs associated with distance and language though they are still confronted with consumers' home bias. Firms' export performance is driven mainly by sector-specific comparative advantages at country level, more so than by their own competitiveness. Export performance is also related to the tradability of services. Commercial services that often require a physical delivery as a counterpart to the online transaction are faced with high distance-related trade costs and therefore turn out to be less tradable than purely digital services. However, in principle, all online services are fully tradable.

Trade policy makers usually focus on trade costs caused by regulatory and other types of barriers. The US' stellar online services export performance cannot simply be explained in terms of trade costs. Firms' individual productivity and competitiveness are important but the overall comparative advantages of countries seem to be even more important. Hecksher-Ohlin endowment and comparative advantage effects remain important in online services trade. Low trade costs are a necessary condition to break out of small domestic markets and generate network effects but are not a sufficient condition to benefit from productivity and endowment effects.

This brings us to some further research suggestions. An obvious next step would be to compare the EU DSM with the US "single market" at the level of US States. There are no language barriers in the US market and very low regulatory trade costs between US States, not zero but probably the lowest achievable. How much more cross-border trade in online services and scale effects could the EU achieve if it could bring trade costs down to the level of the US internal market and what would be the potential economic impact? Other suggestions for further research include extending the Mode 3 aggregation deeper into the long tail of websites and collect more detailed firm level information from other data sources to match with the website data. This would enable us to run more sophisticated firm level trade models on the dataset.

Bibliography

Anderson, J. (1979) "A theoretical foundation for the gravity equation", American Economic Review, vol 69, pp 106-116.

Anderson, J. and Van Wincoop, E. (2003) "Gravity with gravitas: a solution to the border puzzle", American Economic Review, vol 93:1, pp170-192.

Anderson, J and Van Wincoop, E. (2004) "Trade costs", Journal of Economic Literature, vol 42:3, pp 691-751

Balassa, B and Noland, M (1989) "Revealed comparative advantage in Japan and the US", Journal of international economics, vol 4(2), autumn 1989.

Balta, N., Delgado J. (2007) "Home Bias and Market Integration in the EU", paper presented at the CESifo Venice summer institute in 2007, mimeo.

Berthelon, M. and Freund, C. (2004) "On the conservation of distance in international trade", World Bank working paper 3293.

Blum, B and Goldfarb, A. (2006) "Does the internet defy the law of gravity", Journal of international economics, vol 70, pp 384-405.

Brynjolfsson, E, Smith, M and Hu, Y. (2003) "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," Management Science, Vol 49, No. 11

Brynjolfsson, E., A. Dick and M. Smith (2010) "A nearly perfect market", Quantitative Marketing and Economics, Springer, vol. 8(1), pages 1-33, March 2010

Cairncross, F. (1997) "The death of distance; how the communications revolution will change our lives", London, Orion Business Books.

Cowgill, B, Dorobantu, C. and Martens, B. (2013) "Does online trade live up to the promise of a borderless world? Evidence from the EU Digital Single Market", JRC/IPTS Digital Economy working paper nr 2013-08.

Coughlin, C. and D. Novy (2009) "Is the international border effect larger than the domestic border effect: evidence from US trade", Research paper series nr 2009/29, Leverhulme Centre, Nottingham University

CEPII (2010) "Gravity data set", available at http://www.cepii.fr/anglaisgraph/bdd/gravity.htm

Crafts, N., Klein A. (2014) "Geography and intra-national home bias: U.S. domestic trade in 1949 and 2007", Journal of economic geography, April 2014.

Deardorff, A. (1995) "Determinants of bilateral trade: does gravity work in a neo-classical world", in Frenkel ed "Regionalisation of the world economy", Chicago University Press.

Disdier, A. and K. Head (2008) "The puzzling persistence of the distance effect on bilateral trade", The Review of Economics and Statistics, MIT Press, vol. 90(1), pages 37-48, 09

Feenstra, RC (2002) "Border effects and the gravity equation: consistent methods for estimation", Scottish Journal of Political Economy 49:5, pp. 491-506

Frankel, J. and Wei Shang-Jin (1995) "Trading blocks and the Americas", Journal of development economics, vol 47:1, pp 61-96

Freund, C. and Weinhold, D. (2000) "The effect of the internet on international trade", Journal of international economics, vol 62:1, pp 171-189

Gervais, A. and J. Bradford Jensen (2013), "The tradability of services: geographical concentration and trade costs", NBER WP nr 19759, December 2013.

Gomez, E., Martens, B and Turlea, G. (2014) "The drivers and impediments for cross-border e-commerce in the EU", Information Economics and Policy, Volume 28, September 2014.

Gomez, E. and Martens B (2015) "Language, copyright and geographic segmentation in the EU Digital Single Market for music and film", JRC/IPTS Digital Economy working paper, forthcoming.

Hortaçsu, A., Martinez-Jerez F. and Douglas, J. (2009) "The geography of trade in online transactions: evidence from eBay and Mercado Libre", American Economic Journal: Microeconomics, vol 1, pp 53-74.

Kimura, F., Hyun-Hoon Lee (2006) "The Gravity Equation in International Trade in Services", Review of World Economics, April 2006, Volume 142, Issue 1, pp 92-121

Lendle, Andreas, Marcelo Olarreaga, Simon Schropp and Pierre-Louis Vezina (2012), "There goes gravity: how eBay reduces trade costs", CEPR discussion paper, London

McCallum, J. (1995) "National borders matter: Canada-US regional trade patterns", American Economic Review, vol 85, pp 615-23.

Melitz, M. and Redding, S. (2012) "Heterogeneous firms and trade", CEP Discussion paper nr 1183, December 2012.

Pacchioli, C. (2011) "Is the EU internal market suffering from an integration deficit? Estimating the home bias effect", CEPS working document nr 348, May 2011.

Rauch, J. (1999) "Networks versus Markets in International Trade" Journal of International Economics, vol 48, 7-35

Van der Marel, E. and Shepherd, B. (2013) "International tradability indices for services", World Bank Policy Research Working Paper nr 6712, November 2013.

Walsh, K. (2008), "Trade in services: does gravity hold?",IIIS Discussion Paper

Wolf, H. (2000) "Intranational home bias in trade", Review of economics and statistics, vol 84:4, pp 555-563

World Bank (2012) "Logistics performance index".

Statistical Annexes:

Fig 1: Step-wise procedure to identify the CoO of the online services

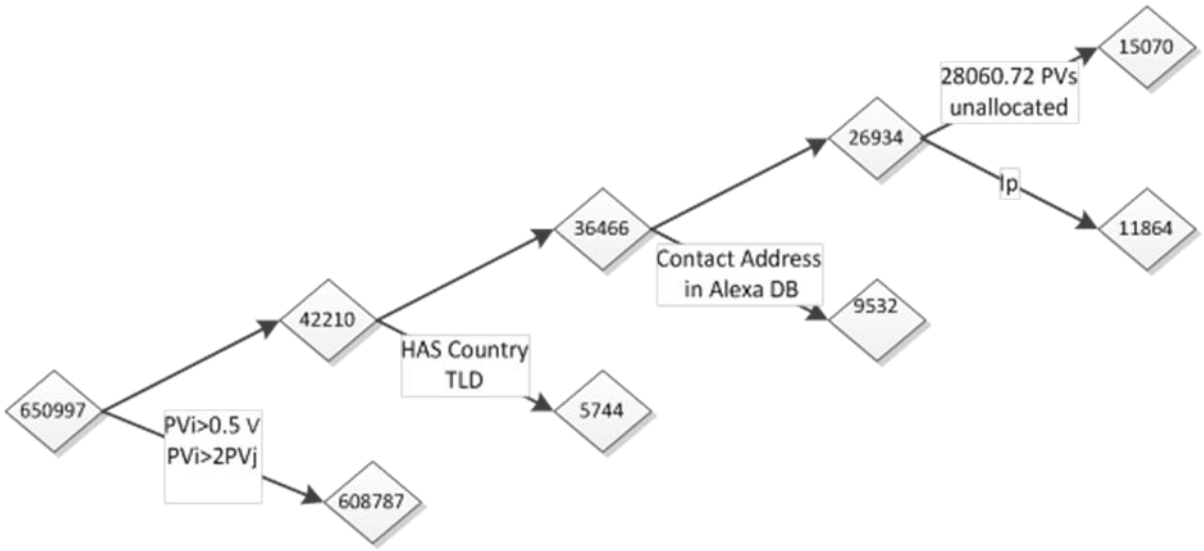


Table 1: Country of Origin classification of websites

Steps in the procedure	WebSites Assigned	%	PageViews Assigned	%
50% of PVs from 1 country and PVs A > 2x PVs B	608,787	94%	751,513	84%
Country extension	614,531	94%	758,423	85%
Contact address in Alexa	624,063	96%	847644	95%
Country IP address	635,927	98%	867080	97%
Unassigned	*15,070*	*2%*	*28060*	*3%*
TOTALS	650,997	100%	895,141	100%

Source: Amazon Alexa and authors' calculations

Table 2: Worldwide online trade in services (with Mode3 aggregation)

Table 2A: Page view traffic between regions (PVs per million)

	Users:							
Origin:	**US**	**EU28**	**RoEUR**	**RoAM**	**Asia**	**RoW**		**TOTAL**
US	156,173	105,914	22,678	50,350	113,102	20,463		468,679
EU28	2,530	81,143	1,864	2,307	3,526	1,304		92,674
RoEUR	810	1,881	47,735	304	3,276	222		54,228
RoAm	867	1,255	186	18,570	1,029	257		22,163
Asia	4,155	3,444	1,466	1,337	164,000	1,468		175,871
RoW	659	932	214	365	1,529	6,672		10,370
TOTAL	165,194	194,569	74,142	73,235	286,461	30,386		823,986

Table 2B: Percentage of incoming page views

	Users:							
Origin:	**US**	**EU28**	**RoEUR**	**RoAM**	**Asia**	**RoW**		**TOTAL**
US	94.5%	54.4%	30.6%	68.8%	39.5%	67.3%		56.9%
EU28	1.5%	41.7%	2.5%	3.2%	1.2%	4.3%		11.2%
RoEUR	0.5%	1.0%	64.4%	0.4%	1.1%	0.7%		6.6%
RoAm	0.5%	0.6%	0.3%	25.4%	0.4%	0.8%		2.7%
Asia	2.5%	1.8%	2.0%	1.8%	57.3%	4.8%		21.3%
RoW	0.4%	0.5%	0.3%	0.5%	0.5%	22.0%		1.3%
TOTAL	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%		100.0%

Table 2C: Percentage of outgoing page views

	Users:							
Origin:	**US**	**EU28**	**RoEUR**	**RoAM**	**Asia**	**RoW**		**TOTAL**
US	33.3%	22.6%	4.8%	10.7%	24.1%	4.4%		100.0%
EU28	2.7%	87.6%	2.0%	2.5%	3.8%	1.4%		100.0%
RoEUR	1.5%	3.5%	88.0%	0.6%	6.0%	0.4%		100.0%
RoAm	3.9%	5.7%	0.8%	83.8%	4.6%	1.2%		100.0%
Asia	2.4%	2.0%	0.8%	0.8%	93.3%	0.8%		100.0%
RoW	6.4%	9.0%	2.1%	3.5%	14.7%	64.3%		100.0%
TOTAL	20.0%	23.6%	9.0%	8.9%	34.8%	3.7%		100.0%

Source: Amazon Alexa and authors' calculations

Table 3: Intra-EU and EU-US online services trade (page views per million, with Mode3 aggregation)

	US	AT	BE	BG	HR	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IE	IT	LV	LT	LU	MT	NL	PL	PT	RO	SK	SI	ES	SE	GB	TOTAL
US	**156,172.8**	**2,124.8**	**2,574.7**	**1,456.3**	**1,443.7**	**1,216.3**	**1,795.3**	**1,848.1**	**1,191.2**	**1,778.2**	**10,800.0**	**13,569.7**	**3,013.9**	**1,580.2**	**1,879.6**	**9,788.0**	**1,246.2**	**1,327.4**	**1,125.4**	**1,119.6**	**4,602.3**	**3,984.6**	**2,227.3**	**2,633.0**	**1,491.6**	**1,325.1**	**9,595.7**	**2,320.1**	**16,855.2**	**262,086.4**
AT	16.9	1,162.0	0.4	1.9	1.0	0.1	2.5	0.1	0.0	0.3	4.3	82.9	2.1	4.6	0.3	11.7	0.2	0.1	0.0	0.0	4.9	4.9	0.9	1.4	2.4	1.4	4.7	0.1	4.9	**1,316.9**
BE	21.5	2.5	951.7	1.3	0.6	0.1	0.2	0.2	0.1	0.0	38.9	15.7	1.6	1.2	0.4	9.2	0.1	0.0	2.1	0.1	32.4	0.6	1.5	1.7	0.2	0.4	10.5	2.5	9.0	**1,106.6**
BG	13.7	2.9	4.4	642.5	0.3	0.4	0.4	1.0	0.0	0.0	1.6	11.0	1.9	0.7	0.1	2.8	0.3	0.0	-	0.0	1.9	1.4	0.9	1.7	0.1	0.1	3.7	1.8	11.8	**707.5**
HR	4.7	2.3	0.1	0.0	348.6	0.0	0.1	0.6	0.0	0.0	0.1	6.1	0.2	0.4	0.0	1.7	0.0	0.0	0.0	0.0	0.3	0.2	0.0	0.1	0.1	5.3	0.4	2.9	0.8	**375.1**
CY	300.4	19.7	25.2	8.2	9.3	68.3	13.0	7.2	7.5	8.6	90.0	215.8	26.2	9.3	8.0	72.5	7.4	7.3	6.6	6.5	40.9	21.3	9.1	14.5	8.4	8.2	41.2	29.1	110.4	**1,200.0**
CZ	39.3	3.8	2.4	0.3	1.2	0.9	1,926.8	3.1	0.2	3.3	14.7	31.2	4.9	1.6	2.6	6.2	1.2	1.2	0.0	0.1	4.9	8.6	3.1	3.9	89.1	1.7	9.4	3.5	22.2	**2,191.2**
DK	29.3	1.2	3.8	0.9	0.3	0.1	1.3	671.1	0.3	0.4	3.9	12.0	1.1	0.2	0.4	4.3	0.3	0.5	0.8	0.3	5.2	1.8	0.3	1.0	0.2	0.2	8.2	15.2	18.0	**782.7**
EE	31.1	3.4	8.9	3.2	3.3	3.1	3.6	3.3	161.7	16.8	31.0	9.8	4.2	3.8	3.2	18.1	4.1	3.6	3.0	3.2	4.4	9.8	3.2	7.1	3.2	3.2	13.6	4.9	10.2	**382.0**
FI	11.9	1.8	0.4	-	0.1	-	0.7	0.8	1.8	735.5	2.8	4.7	0.6	0.1	0.1	1.2	0.1	0.0	0.0	0.2	0.6	1.3	0.1	0.2	0.0	0.1	1.9	3.3	3.4	**773.9**
FR	173.0	11.0	294.0	8.4	2.3	1.4	6.9	5.5	2.1	3.9	11,087.4	101.3	12.9	5.1	7.5	93.5	1.2	2.3	20.5	3.3	18.6	26.2	20.3	10.5	3.7	1.6	115.2	8.1	86.8	**12,134.4**
DE	371.6	446.1	39.5	5.7	13.3	6.2	29.4	25.6	4.5	21.8	193.7	15,473.0	47.4	33.6	9.3	151.5	7.3	10.6	18.2	2.0	105.0	118.2	30.6	43.3	10.8	9.4	172.6	27.9	140.4	**17,568.3**
GR	15.7	1.1	6.4	1.1	0.2	15.2	0.7	3.8	0.2	0.9	2.3	22.2	2,175.1	1.0	0.1	3.5	0.1	0.3	0.1	0.4	7.4	0.3	0.9	1.7	0.1	0.2	2.9	2.2	13.9	**2,280.0**
HU	20.7	4.4	1.6	0.5	1.4	0.5	1.1	2.6	0.4	1.9	3.9	16.6	1.9	970.1	4.1	4.8	0.5	0.7	1.0	0.1	3.4	3.2	1.1	15.9	11.4	0.8	4.8	1.5	11.9	**1,092.8**
IE	23.1	0.4	2.0	0.0	0.2	0.2	0.3	0.4	0.3	0.3	4.1	3.8	0.5	0.3	461.7	1.4	0.1	0.4	0.1	0.1	1.1	2.0	0.4	0.8	0.1	0.1	2.8	0.4	22.1	**529.3**
IT	75.5	4.0	5.5	1.1	1.4	0.5	3.7	1.6	0.9	0.9	34.9	44.5	6.5	2.0	2.0	6,700.3	0.8	0.8	2.7	1.4	13.3	5.8	4.4	6.6	2.6	2.2	33.1	2.8	36.6	**6,998.4**
LV	7.5	0.8	0.1	0.0	0.1	0.0	0.1	6.1	0.9	0.5	0.9	3.3	0.5	0.0	7.7	1.2	324.4	1.5	-	0.0	0.9	0.4	0.4	0.6	0.1	0.1	0.8	1.5	11.6	**372.0**
LT	3.4	0.0	0.1	0.2	-	0.2	0.6	0.7	0.4	1.2	0.9	2.2	0.2	0.0	2.9	0.3	1.4	395.9	-	-	1.2	0.5	0.4	0.1	0.1	-	0.3	2.2	16.8	**432.2**
LU	9.5	0.2	1.5	0.1	0.2	0.0	0.5	-	-	-	5.1	4.1	0.7	0.3	-	1.2	-	0.3	72.9	-	1.0	0.5	0.7	0.8	0.2	0.0	1.9	0.4	2.7	**105.2**
MT	1.6	0.7	0.3	0.2	0.1	0.3	0.1	0.1	0.0	1.4	0.2	2.9	0.6	0.4	0.2	1.3	0.0	0.1	0.0	20.5	3.4	1.3	0.4	0.3	0.0	0.1	0.5	5.4	2.5	**45.0**
NL	164.3	18.6	61.2	6.1	5.4	3.8	8.1	9.8	3.8	5.5	59.8	145.4	24.9	9.6	8.9	73.7	4.6	4.6	5.9	3.2	2,594.8	34.0	18.8	18.1	4.5	4.6	61.3	10.7	80.9	**3,455.2**
PL	47.6	2.7	7.9	0.5	1.2	0.1	4.1	1.5	0.4	0.4	13.0	85.3	2.1	2.6	20.0	6.5	0.9	1.4	0.6	0.3	9.3	6,301.1	2.0	3.8	3.1	0.6	4.3	3.3	98.0	**6,624.4**
PT	48.6	5.2	6.8	4.1	5.7	5.2	8.5	6.9	0.1	5.4	16.6	25.9	21.2	6.6	0.3	23.6	5.5	5.6	2.2	0.0	5.0	7.3	1,013.7	27.6	6.5	5.7	41.2	6.0	16.2	**1,333.2**
RO	35.9	4.1	6.3	1.0	1.0	0.9	2.2	5.9	0.6	1.0	10.4	15.5	9.7	2.4	0.7	21.9	0.4	0.4	0.0	0.0	8.3	3.6	1.6	1,428.8	1.0	0.5	11.1	7.3	21.9	**1,604.3**
SK	4.2	7.5	0.1	0.0	0.2	0.0	18.2	0.1	0.1	0.0	2.2	4.5	0.4	0.7	2.9	2.5	0.0	0.1	0.2	-	0.8	1.9	1.3	0.4	720.7	0.3	1.2	0.4	7.7	**778.7**
SI	3.7	1.3	0.0	0.0	1.6	-	0.2	0.0	0.1	0.0	0.3	1.2	0.3	0.2	0.0	1.8	0.2	0.0	0.0	-	0.8	0.6	0.0	0.3	0.1	229.4	0.5	0.1	1.0	**243.6**
ES	123.2	5.1	8.1	1.1	3.0	1.3	2.1	3.0	1.5	4.2	52.6	55.2	9.9	3.1	4.4	58.9	1.0	1.8	0.0	1.9	25.1	10.6	31.0	6.6	1.4	1.6	6,834.7	6.2	53.3	**7,311.8**
SE	68.4	2.4	2.2	0.9	1.7	0.5	1.3	17.0	0.5	14.0	15.4	24.3	7.1	1.5	2.5	13.0	0.7	1.3	0.2	0.4	11.4	8.9	2.8	4.9	0.5	0.4	14.2	1,399.5	24.0	**1,641.6**
GB	863.9	23.3	32.3	14.6	13.2	12.2	27.0	22.6	9.7	23.8	172.5	139.2	40.1	20.4	102.1	127.4	10.3	11.1	3.8	9.0	76.1	59.0	44.3	37.1	12.1	10.6	142.0	39.7	8,187.5	**10,286.7**
TOTAL	**158,702.9**	**3,863.5**	**4,047.8**	**2,160.2**	**1,860.5**	**1,337.9**	**3,858.8**	**2,648.5**	**1,389.1**	**2,630.1**	**22,663.7**	**30,129.3**	**5,418.7**	**2,662.2**	**2,532.0**	**17,204.1**	**1,619.4**	**1,779.2**	**1,266.5**	**1,172.4**	**7,584.7**	**10,619.8**	**3,421.4**	**4,273.0**	**2,374.3**	**1,613.9**	**17,134.9**	**3,909.0**	**25,881.5**	**345,759.4**
%US	98%	55%	64%	67%	78%	91%	47%	70%	86%	68%	48%	45%	56%	59%	74%	57%	77%	75%	89%	95%	61%	38%	65%	62%	63%	82%	56%	59%	65%	

Source: Amazon Alexa data and authors' calculations

Table 4: Domestic and foreign online trade broken down in extensive (#websites) and intensive (#pageviews) margins

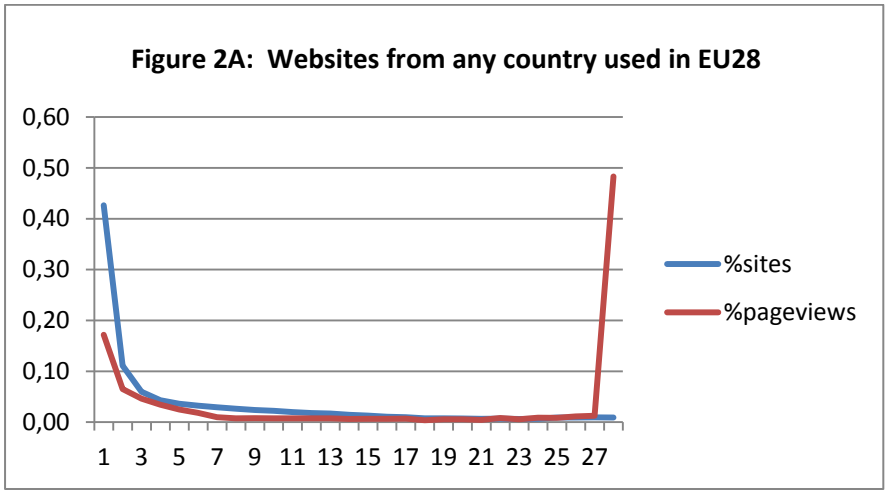
	Extensive margin (#websites)						Intensive margin (#page views per million)					
CC	#sites used	#sites Dom	#sites Foreign	#sites Export	%Dom	%Exp	#PVs Total	#PVs Dom	#PVs Foreign	#PVs Export	%Dom	%Exp
PL	26,207	15,475	10,732	2,458	59%	9%	10,926	6,301	4,625	437	58%	4%
DE	98,894	54,171	44,723	9,694	55%	10%	31,217	15,473	15,744	3,189	50%	10%
FR	54,068	28,062	26,006	8,609	52%	16%	23,373	11,087	12,285	2,048	47%	9%
ES	57,302	27,346	29,956	4,927	48%	9%	17,888	6,835	11,053	1,534	38%	9%
HU	9,453	4,405	5,048	856	47%	9%	2,766	970	1,796	202	35%	7%
GR	26,345	12,105	14,240	1,387	46%	5%	5,633	2,175	3,458	139	39%	2%
IT	38,177	17,256	20,921	3,271	45%	9%	17,774	6,700	11,074	506	38%	3%
CZ	9,415	4,013	5,402	1,054	43%	11%	4,017	1,927	2,090	398	48%	10%
RO	20,419	7,600	12,819	1,335	37%	7%	4,504	1,429	3,076	290	32%	6%
SK	5,364	1,823	3,541	397	34%	7%	2,452	721	1,732	78	29%	3%
NL	20,335	6,768	13,567	2,257	33%	11%	8,042	2,593	5,449	1,511	32%	19%
GB	79,104	25,663	53,441	11,471	32%	15%	26,919	8,187	18,731	3,880	30%	14%
LT	4,634	1,466	3,168	258	32%	6%	1,871	396	1,476	61	21%	3%
HR	6,226	1,856	4,370	448	30%	7%	1,950	349	1,601	66	18%	3%
SE	10,454	3,015	7,439	1,134	29%	11%	4,139	1,400	2,739	436	34%	11%
SI	4,208	1,198	3,010	191	28%	5%	1,675	229	1,446	23	14%	1%
DK	6,152	1,732	4,420	548	28%	9%	2,764	671	2,093	176	24%	6%
BG	2,553	655	1,898	385	26%	15%	2,268	643	1,625	113	28%	5%
EE	2,951	728	2,223	140	25%	5%	1,465	162	1,304	633	11%	43%
LV	3,680	903	2,777	221	25%	6%	1,743	324	1,419	72	19%	4%
FI	3,987	952	3,035	362	24%	9%	2,744	736	2,009	61	27%	2%
PT	9,710	2,301	7,409	571	24%	6%	3,655	1,014	2,642	1,385	28%	38%
AT	14,282	3,366	10,916	1,378	24%	10%	4,011	1,162	2,849	187	29%	5%
BE	12,099	2,099	10,000	832	17%	7%	4,227	952	3,275	210	23%	5%
MT	1,585	268	1,317	65	17%	4%	1,204	20	1,184	40	2%	3%
IE	8,219	1,273	6,946	619	15%	8%	2,646	462	2,184	119	17%	5%
CY	3,966	614	3,352	212	15%	5%	1,401	68	1,333	1,764	5%	126%
LU	1,247	135	1,112	83	11%	7%	1,295	73	1,222	56	6%	4%
EU AV	**19,323**	**8,116**	**11,207**	**1,970**	**32%**	**8%**	**6,949**	**2,609**	**4,340**	**700**	**28%**	**13%**
US	257,208	160,921	96,287	62,692	63%	24%	165,196	156,173	9,024	312,506	95%	189%
RU	45,416	33,362	12,054	12,784	73%	28%	43,136	31,696	11,440	7,115	73%	16%
BR	16,736	9,395	7,341	2,816	56%	17%	26,264	9,370	16,893	795	36%	3%
CN	31,358	25,970	5,388	8,805	83%	28%	91,596	86,052	5,545	6,112	94%	7%
JP	26,434	16,477	9,957	5,546	62%	21%	42,093	20,593	21,500	1,188	49%	3%
IN	126,155	74,281	51,874	23,973	59%	19%	71,324	27,923	43,401	6,602	39%	9%

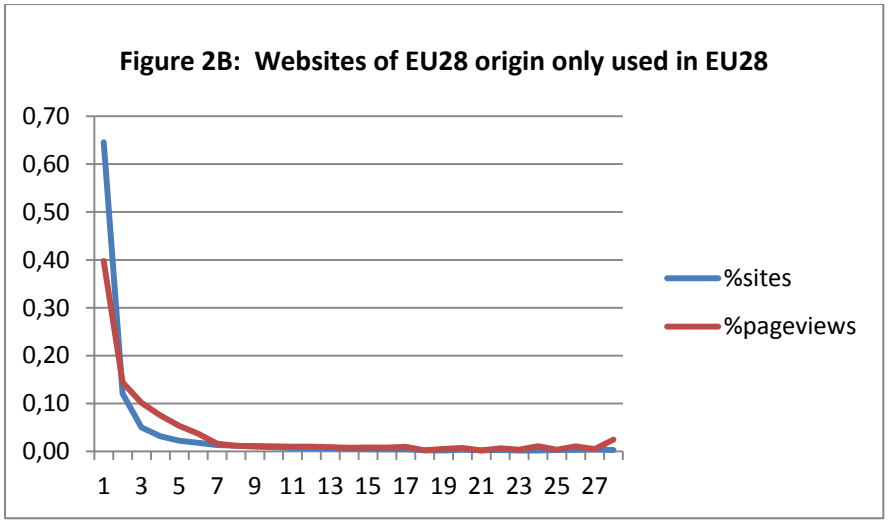
Source: Amazon Alexa and authors' calculations

Number of PVs and WS by category			
Category	**#WS**	**#PVs**	**%PVs**
Search Engines	1,045	197,616	22.2%
Social Networking	1,626	79,301	8.9%
Portal Sites	8,875	62,398	7.0%
Online Shopping	46,134	53,207	6.0%
Business	63,987	41,011	4.6%
Internet Services	40,393	39,444	4.4%
Streaming/Downloading Mec	5,006	33,900	3.8%
Blogs/Wiki	50,300	33,701	3.8%
General News	17,559	32,022	3.6%
Entertainment	27,796	27,903	3.1%
Pornography	18,794	26,166	2.9%
Marketing/Merchandising	55,157	25,202	2.8%
Auctions/Classifieds	3,041	24,843	2.8%
Education/Reference	25,703	18,731	2.1%
Finance/Banking	14,628	18,675	2.1%
Software/Hardware	14,929	16,038	1.8%
Forum/Bulletin Boards	9,747	13,809	1.6%
Games	16,978	13,125	1.5%
Travel	18,671	10,881	1.2%
Sports	15,051	10,485	1.2%
Technical Information	9,683	10,427	1.2%
Media Sharing	2,674	8,846	1.0%
Public Information	13,293	8,206	0.9%
Illegal Software	3,896	7,549	0.8%
Fashion/Beauty	11,497	6,926	0.8%
Web Applications	1,845	6,078	0.7%
Real Estate	9,302	5,760	0.6%
Job Search	5,563	5,356	0.6%
Health	12,481	5,322	0.6%
Web Ads	886	4,510	0.5%
Government/Military	6,186	4,127	0.5%
Recreation/Hobbies	9,645	3,708	0.4%
Dating/Personals	1,636	3,265	0.4%
Motor Vehicles	3,729	3,130	0.4%
Gambling Related	3,697	2,689	0.3%
Malicious Sites	4,151	2,688	0.3%
Parked Domain	4,789	2,539	0.3%
Others	35,130	19,081	2.1%

Source: Alexa data, McAfee categorisations and authors' calculations

Figure 2: Geographic distribution of online services traffic





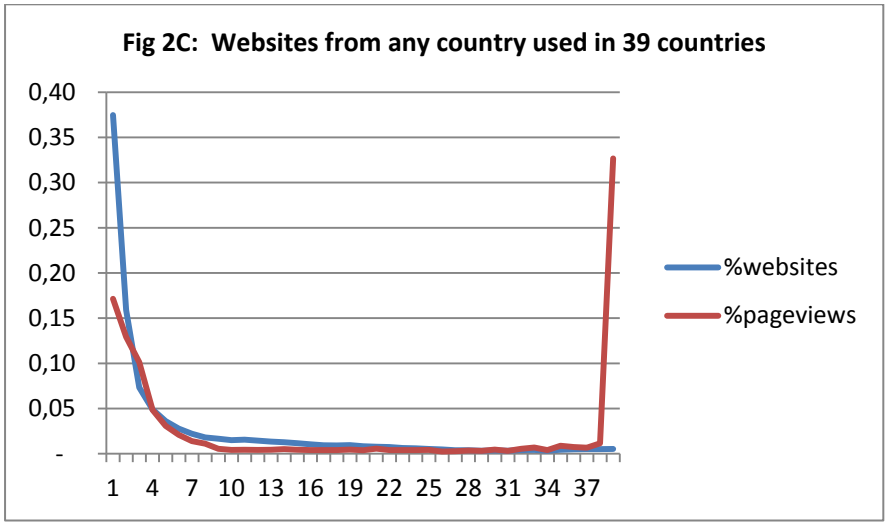


Table 5: Ratio of domestic to cross-border activity by category of online services

		All countries by origin and EU28 destinations only						
		Domestic sites		Cross-border sites		Ratio of domestic to		
	Category	#PVs	#sites	#PVs	#sites	#PVs	#sites	Comb
1	Alcohol/Tobacco	144	861	29	129	4.89	6.67	5.78
42	Restaurants	177	1,441	73	158	2.42	9.12	5.77
39	Real Estate	546	3,348	1,052	514	0.52	6.51	3.52
30	Non-Profit/Advocacy/NGO	200	2,115	94	512	2.13	4.13	3.13
31	Online Shopping	4,313	19,790	7,839	3,798	0.55	5.21	2.88
26	Marketing/Merchandising	2,912	18,942	2,433	4,722	1.20	4.01	2.60
38	Public Information	1,038	6,076	1,389	1,445	0.75	4.20	2.48
18	Government/Military	398	1,855	314	513	1.27	3.62	2.44
20	Health	516	4,427	348	1,302	1.48	3.40	2.44
29	Motor Vehicles	212	1,354	278	350	0.76	3.87	2.32
52	Travel	1,247	8,546	2,107	2,188	0.59	3.91	2.25
53	Weapons	38	220	29	73	1.30	3.01	2.16
40	Recreation/Hobbies	487	3,804	509	1,151	0.96	3.30	2.13
24	Job Search	452	1,895	653	547	0.69	3.46	2.08
12	Fashion/Beauty	807	4,345	827	1,428	0.98	3.04	2.01
13	Finance/Banking	1,274	5,095	2,782	1,560	0.46	3.27	1.86
3	Art/Culture/Heritage	180	1,304	303	448	0.59	2.91	1.75
6	Business	3,058	20,029	4,932	7,027	0.62	2.85	1.74
14	Forum/Bulletin Boards	656	3,305	1,063	1,236	0.62	2.67	1.65
56	Web Mail	61	330	172	113	0.35	2.92	1.64
5	Blogs/Wiki	1,438	15,059	4,455	5,202	0.32	2.89	1.61
47	Spam URLs	60	463	47	244	1.27	1.90	1.59
48	Sports	856	5,265	1,857	2,036	0.46	2.59	1.52
15	Gambling Related	354	1,431	664	625	0.53	2.29	1.41
41	Religion/Ideology	59	603	52	367	1.14	1.64	1.39
35	Politics/Opinion	59	630	88	316	0.67	1.99	1.33
4	Auctions/Classifieds	555	1,025	7,205	396	0.08	2.59	1.33
33	Parked Domain	167	900	157	572	1.06	1.57	1.32
34	Personal Pages	148	1,197	325	557	0.46	2.15	1.30
9	Dating/Personals	249	587	581	296	0.43	1.98	1.21
37	Portal Sites	669	2,869	6,941	1,266	0.10	2.27	1.18
23	Internet Services	2,646	12,014	5,847	6,605	0.45	1.82	1.14
44	Sexual Materials	61	248	142	138	0.43	1.80	1.11
28	Mobile Phone	48	278	58	243	0.82	1.14	0.98
10	Education/Reference	832	6,314	3,189	3,842	0.26	1.64	0.95
8	Criminal Activities	25	136	35	123	0.70	1.11	0.90
16	Games	829	4,526	2,200	3,380	0.38	1.34	0.86
11	Entertainment	1,147	7,688	4,594	6,014	0.25	1.28	0.76
49	Stock Trading	55	342	205	285	0.27	1.20	0.73
19	Gruesome Content	13	52	18	73	0.72	0.71	0.72
7	Chat	44	267	207	232	0.21	1.15	0.68
17	General News	817	5,058	6,065	4,142	0.13	1.22	0.68
36	Pornography	1,154	5,206	5,348	4,788	0.22	1.09	0.65
46	Software/Hardware	571	3,683	1,980	3,812	0.29	0.97	0.63
25	Malicious Sites	77	745	233	811	0.33	0.92	0.62
51	Technical Information	286	2,390	1,443	2,539	0.20	0.94	0.57
45	Social Networking	94	413	18,957	378	0.00	1.09	0.55
27	Media Sharing	136	600	1,444	668	0.09	0.90	0.50
22	Internet Radio/TV	37	313	247	378	0.15	0.83	0.49
2	Anonymizers	28	84	79	157	0.35	0.54	0.44
50	Streaming/Downloading Media	72	859	7,808	1,058	0.01	0.81	0.41

Table 6: Gravity regressions at country level

	Without Mode 3 aggregation				With Mode 3 aggregation			
	Dep = Log #pageviews		Dep = Log #websites		Dep = Log #pageviews		Dep = Log #websites	
	OLS	TOBIT	OLS	TOBIT	OLS	TOBIT	OLS	TOBIT
(Intercept)	13.387 ***	13.457 ***	16.448 ***	16.479 ***	13.504 ***	13.574 ***	16.457 ***	16.487 ***
Distance	-0.355 ***	-0.362 ***	-0.555 ***	-0.558 ***	-0.345 ***	-0.352 ***	-0.556 ***	-0.559 ***
Home bias	4.947 ***	4.941 ***	3.324 ***	3.321 ***	4.704 ***	4.698 ***	3.324 ***	3.321 ***
Com language	0.545 ***	0.559 ***	0.570 ***	0.576 ***	0.532 ***	0.546 ***	0.570 ***	0.577 ***
Border effect	140.8	139.9	27.8	27.7	110.4	109.7	27.8	27.7
Language eff	1.72	1.75	1.77	1.78	1.70	1.73	1.77	1.78

Source: Amazon Alexa data and authors' estimations.

Table 7: Gravity regression coefficients by category (OLS only)			
Category	Distance	Language	HomeBias
Portal Sites	0.10	2.66	7.12
Restaurants	-0.27	2.64	6.47
Auctions/Classifieds	-0.17	1.62	6.47
Alcohol/Tobacco	-0.18	2.04	6.17
Health	-0.26	2.99	5.97
Social Networking	0.01	3.54	5.81
Public Information	-0.22	2.36	5.78
Job Search	-0.25	2.00	5.73
Real Estate	-0.48	1.80	5.70
Government/Military	-0.40	1.75	5.65
Religion/Ideology	-0.16	2.18	5.58
Finance/Banking	-0.33	1.59	5.56
Marketing/Merchandising	-0.18	2.13	5.52
Forum/Bulletin Boards	-0.16	2.18	5.48
Weapons	-0.42	1.51	5.46
Recreation/Hobbies	-0.27	2.41	5.41
Motor Vehicles	-0.35	1.71	5.38
Politics/Opinion	-0.26	2.15	5.35
Mobile Phone	-0.22	3.00	5.31
Non-Profit/Advocacy/NGO	-0.13	1.65	5.21
Blogs/Wiki	-0.13	1.74	5.17
General News	-0.18	2.56	5.12
Sexual Materials	-0.28	2.39	5.09
Spam URLs	-0.36	1.73	5.07
Parked Domain	-0.16	1.61	5.01
Streaming/Downloading Mec	-0.06	2.46	4.98
Web Mail	-0.61	2.35	4.97
Education/Reference	-0.11	1.63	4.92
Online Shopping	-0.20	1.13	4.90
Gambling Related	-0.35	1.68	4.87
Fashion/Beauty	-0.25	1.73	4.82
Technical Information	0.10	2.51	4.78
Chat	-0.08	2.18	4.78
Search Engines	-0.25	2.49	4.71
Internet Services	-0.13	2.14	4.70
Business	-0.14	1.31	4.70
Entertainment	-0.29	2.18	4.69
Sports	-0.25	1.75	4.62
Criminal Activities	-0.19	1.28	4.46
Malicious Sites	-0.18	1.66	4.44
Illegal Software	-0.07	2.26	4.41
Travel	-0.36	1.40	4.37
Personal Pages	-0.22	1.70	4.29
Dating/Personals	-0.55	2.09	4.22
Art/Culture/Heritage	-0.31	1.41	4.22
Software/Hardware	-0.13	1.98	4.16
Internet Radio/TV	-0.48	1.64	4.13

Table 8: Regressions with dependent variable = ratio of foreign / domestic WS and PVs

	Websites		Page views	
(Intercept)	3.412		12.977	***
logGDP/capita	0.281		-0.098	
percentint	-1.125		-2.647	***
%Englishspeakers	0.009	*	0.012	***
log(Population)	-0.325	***	-0.598	***

Table 9: Gravity regressions with the Melitz-Toubal (2012) continuous language distance variables

	(1)		(2)		(3)		(4)		(5)		(6)	
(Intercept)	13.690	***	12.629	***	13.398	***	12.520	***	12.455	***	12.517	***
ldist	-0.366	***	-0.315	***	-0.340	***	-0.295	***	-0.290	***	-0.295	***
HomeBias	4.670	***	5.129	***	4.738	***	5.088	***	5.113	***	5.089	***
col	0.613	***					0.087		0.095		0.088	
csl			1.140	***			0.761	***	0.727	***	0.759	***
cnl					1.656	***	1.058	***	1.117	***	1.060	***
lp1									0.015			
lp2											0.001	

Note: COL= common official language; CSL = common spoken language; CNL = common native language; LP1 and LP2 are linguistic proximity indexes.

Table 10: The gravity model at the extensive margin (number of export markets) by firm size quantile

Quantiles	10%		25%		50%		75%		90%		95%		99%	
(Intercept)	-5.180	***	-3.796	***	-2.834	***	-1.739	***	-0.315	**	1.123	***	4.859	***
distance	-0.080	***	-0.084	***	-0.069	***	-0.050	***	-0.042	***	-0.042	***	0.009	
HomeBias	2.408	***	1.517	***	1.337	***	1.334	***	1.361	***	1.336	***	1.485	***
Common lang	0.868	***	0.298	***	0.121	***	0.046	**	0.028		0.034		0.180	
Imp/Exp FE	x		x		x		x		x		x		x	
Category FE	x		x		x		x		x		x		x	

Source: Amazon Alexa data and authors' estimations

Table 11: The drivers of online firms' export performance (dep var = log of #export markets)

	Export index of RCA (pageviews)								Export index of RCA (websites)				Net exports index of RCA (page views)	
	(1)		(2)		(3)		(4)		(5)		(6)		(7)	
(Intercept)	1.438	***	1.478	***	1.365	***	1.489	***	1.261	***	1.329	***	2.368	***
Log RCA	-0.114	***	-0.114	***	0.095	***	0.124	***	0.064	***	0.129	***	0.587	***
Log Firm share	0.078	***	0.078	***	0.098	***	0.098	***	0.073	***	0.073	***	0.167	***
Category FE:														
Cat Media	0.005		-0.044	***	-0.143	***	-0.120	***	0.095	***	0.094	***	0.065	***
Cat News	0.000		-0.014		-0.012		-0.028	**	-0.015	*	-0.014		-0.083	***
Cat Other	-0.011		0.023		-0.107	***	0.044	***	-0.004		-0.004		-0.171	***
Cat Personal	0.033	***	0.022	***	-0.019		-0.017	**	0.072	***	0.072	***	0.040	***
Cat Social_net	-0.067	***	-0.123	***	-0.066	***	-0.008		0.086	***	0.086	***	0.277	***
Cat Technical	0.266	***	0.245	***	-0.012		0.129	***	0.280	***	0.278	***	0.184	***
R2	0.053		0.053		0.073		0.065		0.046		0.046		0.208	
# Obs	159,995		159,995		29,918		56,080		159,995		159,995		159,995	
Countries	39		39 (Ref = US)		EU28		38 excl US		39		39		39	

Source: Amazon Alexa data and authors' calculations.

Notes: Reference category = commercial services. Firm share = ratio of a website's domestic page views in total domestic page view in the same category.

Annex II: Construction of the Data Set

The purpose of this annex is to outline the methodology for the construction of the dataset. The objective is to construct a bilateral information exchange matrix at the country level of the following form:

$$\begin{matrix} PV_{11} & PV_{12} & \cdots & PV_{1n} \\ PV_{21} & PV_{22} & & \\ \vdots & & \ddots & \\ PV_{k1} & \dots & & PV_{kn} \end{matrix} \quad (1)$$

PV_{ij} represents Pageviews per million Pageviews made in the web over a three month average[12] from users in country *j* to web sites hosted in country *i*. Therefore, rows represent host countries and columns user countries. The matrix is proportional as absolute values are not available at least from the data source considered in this paper.

In what follows a ***Pageview*** as is defined as "*the total number of Alexa user URL requests for a site. Multiple requests for the same URL on the same day by the same user are counted as a single Pageview*"[13]. Below the roadmap for constructing this matrix is given.

Getting the Top websites of each country and Querying AWIS

The first step for the construction of the data set is to obtain the most popular websites for a pre-defined set of countries. This is done by querying the Amazon Top Sites (henceforth **ATS**) **API** of Amazon. A brief overview of the Top Sites can be found already in http://www.alexa.com/, where the top 500 web sites are displayed for each country. According to Alexa the ranking of each web site is determined using a combination of the Unique Visitors and Pageviews over a rolling period of 3 months. The top sites were obtained for the **eu-28** plus 11 more countries; **Switzerland. Norway, Turkey, Russia, Japan, China, India, Brazil, United States, Canada** and **Australia**. In **Table 1** the precise number of websites obtained for each country in the analysis can be found. In order to obtain information per website a query was made to the API of the Amazon Web Information Services (henceforth **AWIS**).

[12] The data Collection process took place in September – October 2014.

[13] https://alexa.zendesk.com/hc/en-us/articles/200449744-How-are-Alexa-s-traffic-rankings-determined- . Note that Amazon offers a second variable Reach which is instead of Pageviews shows the *Users* out of 1e6 Users in the internet that visited a website. Pageviews were used because they are bound to sum to a million which is not true for the Reach variable. However, the same exercise can be done for the unique users.

Table 1: Number of Websites obtained for each country in the sample [14]

CountryCode	WebSites	CountryCode	WebSites
AT	10764	**LV**	3513
BE	9385	**MT**	1771
BG	1331	**NL**	19781
CH	9905	**NO**	4823
CY	4106	**PL**	30671
CZ	9076	**PT**	8647
DE	98714	**RO**	20419
DK	5737	**SE**	10031
EE	3072	**SI**	4115
ES	60240	**SK**	4875
FI	3119	**TR**	74791
FR	54930	**AU**	27923
GB	66054	**BR**	15241
GR	29281	**CA**	45023
HR	6904	**CN**	33116
HU	9661	**IN**	125726
IE	6931	**JP**	22212
IT	34781	**RU**	49136
LT	4778	**US**	179263
LU	790		

From the AWIS query the following set of information were obtained relevant to the construction of the web traffic matrix:

- ***Percentage of Pageviews ranked by country***. This is a vector of percentages giving the Pageview distribution of users for website WS_i broken down by country.

- ***Pageviews per million*** is the number of times a website was viewed in a predefined period, normalized over one million. AWIS provides different periods; 1 month and 3 months. The one used here is a 3 month period such that it is homogenous with the percentage of Pageviews and Top Websites.

- ***Physical Address:*** is the contact address AWIS reports for a website[15]

From the first two pieces of information the total Pageviews made from country ***c*** to website ***i*** out of one million Pageviews in the internet can be pinned down. For instance, if a Website has **x** Pageviews over the last 3 months, and the fraction of them made by country **c** is $\mathbf{y}_c$ then

[14] The number of Websites obtained for the US was 212932. The difference between this number and the one in the table is that the Alexa dataset for US was composed by a lot of duplicates. This is not the number of Websites used in the final dataset, as per country information could not be retrieved for all of them.

[15] The physical address is reported scarcely. An additional obstacle is that a significant part is in the language of the website.

simply $x_c = x \times y_c$. These two variables were not always present. In fact after merging the data the remaining sample is **650997** websites.

Hence the user countries and their proportional use of a Website are obtained. The next step for the construction of (1) is to determine the host countries.

Defining the Host Country of a Website.

Perhaps one of the most challenging and novel parts of the paper is to determine the origin of the website in the borderless internet world. The problem is less technical and mostly conceptual; what does exactly an 'Origin' of Website mean? Does it mean the physical location of its servers or does it mean its target users. In this paper the host country of a website is assumed to be the country where the bulk of its users come from. In order to classify a website to host country the process depicted in the Tree Diagram in **Figure 1.**

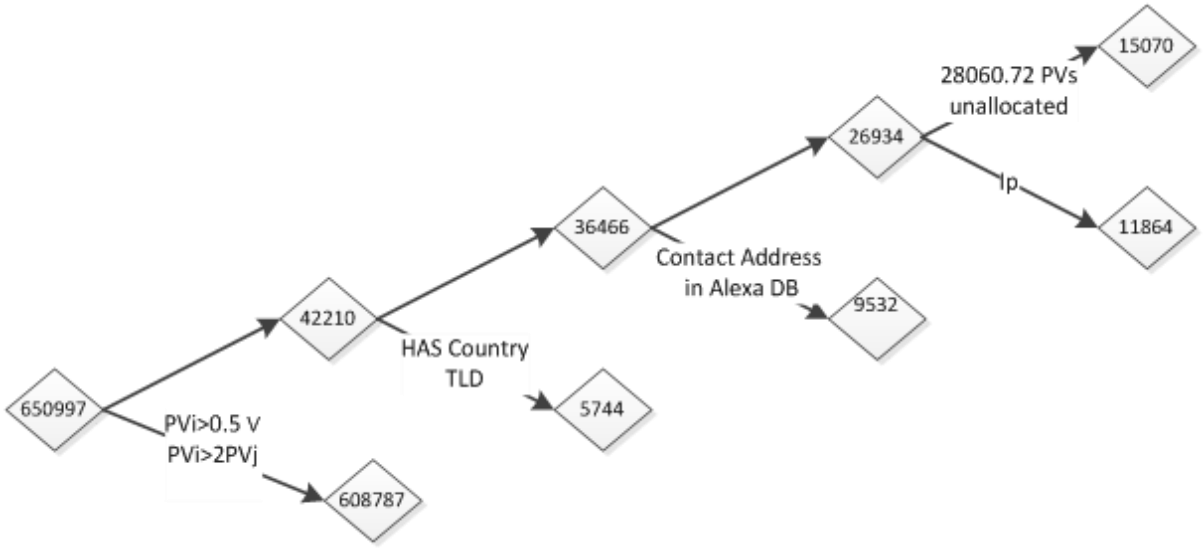


Figure 1: Website classification process to Host countries

The numbers in the nodes represent number of websites, and in the branches are the rules used to classify the websites. The origin country of a website was determined primarily based on the users of the website. Since the user countries of the website is known from the AWIS data, **when the audience exceeds a certain threshold t_i or the audience of the country with the most Pageviews is double the one with the second then the website is assumed to be hosted in that country**. The threshold was set to 50% percent. Approximately 98% of websites were allocated using this rule. The rest were allocated by their country Top Level Domain (ccTLD) if it exists, the information obtained from the Alexa Database and finally some of them using the IP. The geographic location of the IP was found by two different locations and only the WebSites who's IP where the same from both locations were classified.

A potential check to see if the rule in the first branch of the tree works is to compare it with the ccTLD that may be considered as an accurate definition for the host country of a website. For the urls where both definitions apply **91%** of the times the two definitions give the same results. This enables us to confidently apply to websites with no ccTLD domain. Note further that with this classification scheme Global Platforms Amazon, Google, Yahoo, Youtube and ebay were aggregated.

After the classification process only 0.25% remained unallocated accounting for 0.032% of the Pageviews. **635324** Websites Remain for the analysis with **794767.7** PageViews accounted for. Merging the original data with the host country and aggregating over it enables us to get the bilateral matrix given in **(1)**.

Missing Values

A significant problem with the Data was encountered with concern to smaller countries. When a country accounts for less than **0.5%** of the viewing audience for a particular website AWIS rolls it up to the category "Others". This creates a significant caveat in the construction of the dataset. *Facebook.com* for instance appears almost always in the first three places on the top list of any country. Since countries with a small population relative to others like Austria or Cyprus always account for less than 0.5% of the viewing audience of *Facebook* no data are available for these countries. This is not usually a problem for local websites but mostly for websites that are global. Therefore, while for big countries like Germany, UK or US data are always going to be available and reliable for smaller countries numbers will be understated as only local web sites are accounted for. In order to solve this issue a two-step procedure is followed. First, a way is needed in order to identify the countries for which Pageviews are missing. After doing so the missing values were imputed.

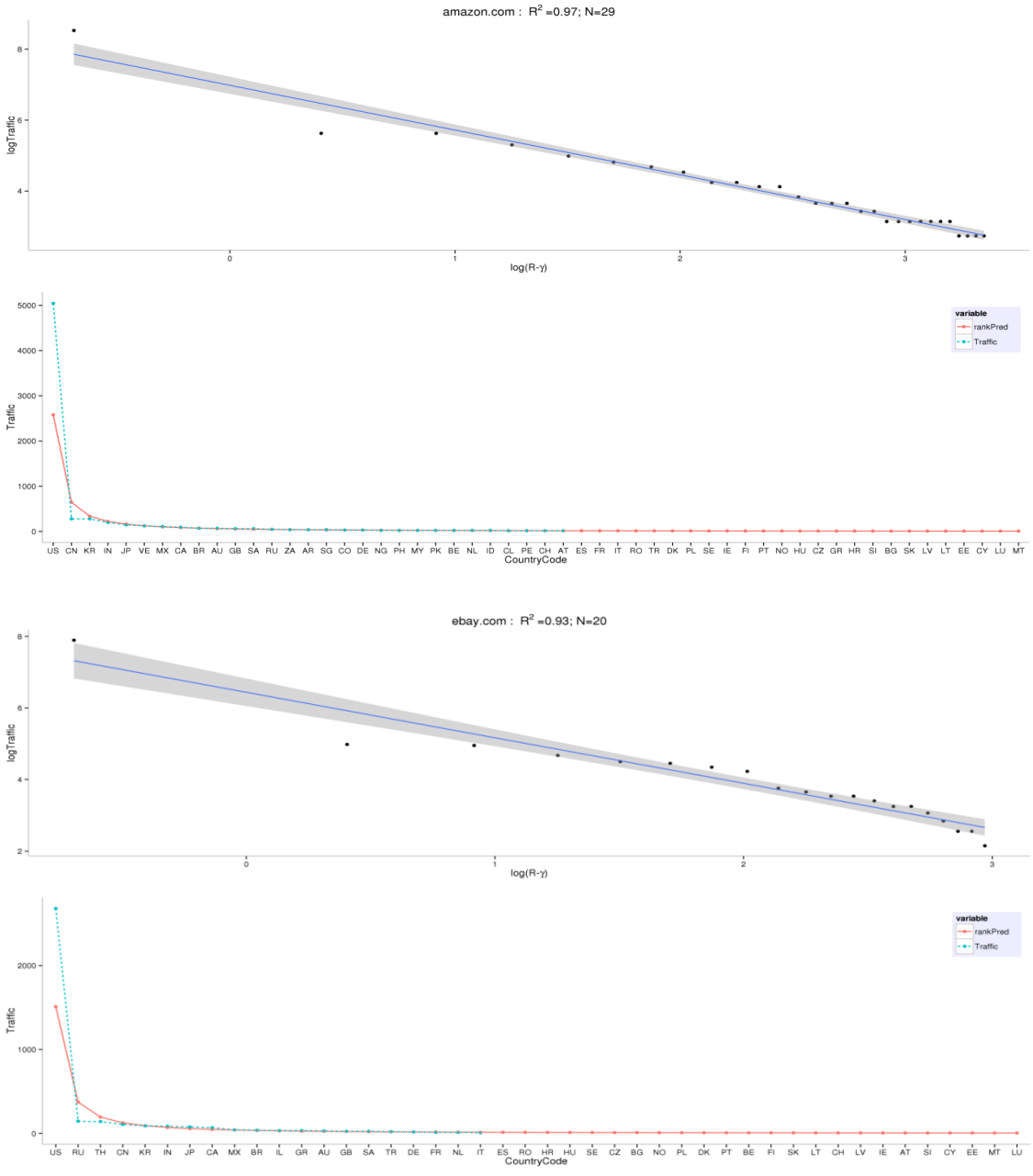
I. Identifying Actual Missing Values

The first step is to find a way to identify if a Website should have Pageviews in a certain country. It might be that users in a particular country do not actually view the website. If all countries are just assumed to be viewers for all websites, then Pageviews will be falsely estimated for a subset of countries. Take again *Facebook* as an example that appears in the second place in the top list for Austria but is censored in China.

A natural way to do this is to identify for which countries a website does not have Pageviews and at the same time appears in the list of top websites in that country obtained from ATS. In that case the website-country combination is identified as missing. If there is no presence of that Website in that country, then it is assumed that the contribution of that country in the audience of the web site is 0 or negligible. Naturally, the "bigger" is the website the largest is going to be the set of missing countries (Correlation between the two is 0.49).

II. Imputation

After identifying the countries for which Pageviews need to be estimated the missing elements were imputed in the matrix. The initial observation for constructing the imputation methodology is that per Website Pageviews follow a power-law distribution and therefore Zipf's law may apply. This is shown in the upper panels of **Figure 2** for some major Websites.



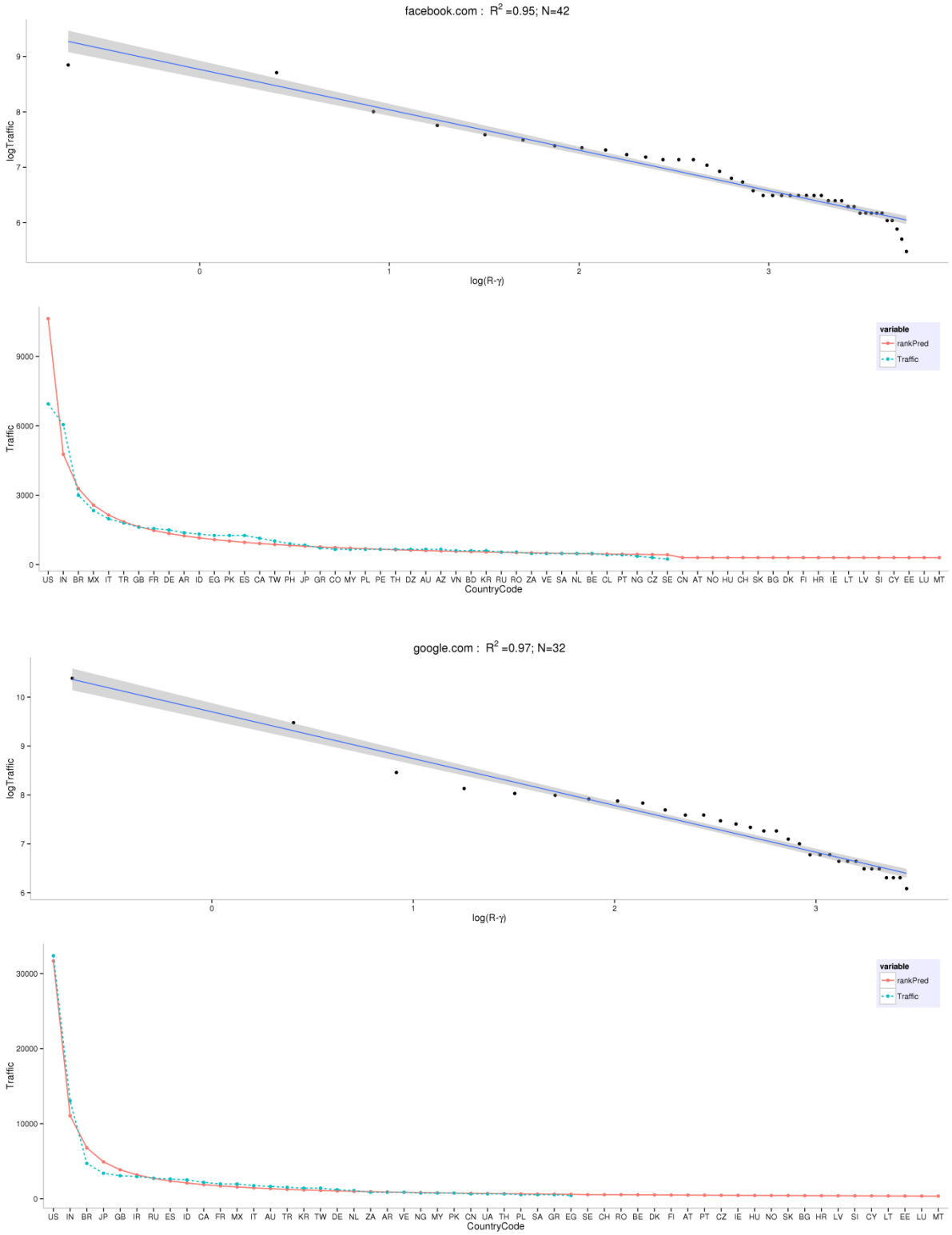


Figure 2 Imputation Plots.

As it can be seen there the log-Traffic and the log of the rank have an almost linear relationship. Therefore the following model was estimated[16]:

$$\log(PV_{\mathrm{WSi}}) = c + \delta\log(R_{\mathrm{WSi}-}\gamma) + \varepsilon \qquad (2)$$

Where:

- *PV* is a nx1 vector of Pageviews where n is the number of countries for which data are available.
- *R* is the by country Rank in terms of Pageviews for Website *I* of the same dimension
- ε is the error term.

The estimator of δ is referred as the *QQ-estimator* by Kratz and Resnick (1996). $\gamma=1/2$ is a constant parameter used to correct the bias in small samples (as in this case) suggested by Gabaix and Ibragimov (2009). In the training set the countries were ranked according to their traffic; since traffic is not available, for the prediction set the countries were ranked according to their effective internet Population weighed by the rank of the Website: $wIU_c = \frac{IU_c}{RankWS_{c,i}}$. This weighting scheme was used in order to account for the popularity of the web site in a country; otherwise the model would falsely predict high Pageviews for countries with very large population. The Spearman correlation between Traffic and that variable is **0.84** indicating it is an adequate Ranking Variable.

Table **2** gives an overview of the imputation Regression. The minimum length of the training dataset allowed was 10; Furthermore, given that the ceiling of the target is known from before when a prediction exceeded this value, then Traffic was Top-coded to 0.5% of the Websites total traffic:

Table 2:**Imputation Statistics**

Median R^2	**Median RMSE**	**WebSites Imputed**	**Constraint Violation**	**PageViews Imputed**	**Observations Imputed**	**Individual constraint Violation**
0.93	0.13	12112	67	29236.68	45022	14553

As it can be seen from some statistics the model performs well in terms of goodness of fit and RMSE, at least in the training set. The performance of the model is average in the prediction set. The problem is that 32.3% of the time the individual constraint was violated; $\widehat{PV_{\mathrm{WSi,c}}} > 0.005 \times Traffic_{\mathrm{WSi}}$. In that case as mention above the Traffic for country **c** was top-coded to

[16] We would like to thank Duch-Brown Nestor for pointing us to this direction. Models with other exogenous regressors were implemented as well. Given that the data are not missing at random this model was preferred.

the ceiling value. On the other hand the total constraint was violated only 0.005% of the times. After the imputation the total traffic accounted for by our data is 824004.4 PageViews.

URL classification

AWIS offers some categorization based on the Open Directory Project. The major problem with this categorization is that a very small amount of Websites in the sample is categorized and an even smaller amount is in English. An attempt was made to go to the source Data Base and try to automatically translate Categories in to English via the Bing API. This attempt was abandoned as the sample fraction that would be categorized would be potentially small (around 15%), while it was quite difficult to have uniform categorization in the remaining websites after translating for more than 30 Languages.

McAfee's categorization was finally used. The advantage of this is that the coverage is more than 90% of the sample, while the categories are well documented[17]. The con is that this categorization is security oriented. Nevertheless, after some merging the following picture emerges with concern to the categorization of the data. Table 4b presents the classification of the urls into categories.

Data Sources:

- Top Sites: The top sites for each country were obtained from the Amazon Top Sites (**ATS**) API.
  http://aws.amazon.com/alexa-top-sites/

- Web Site Metrics: Information about each website was obtained from the Amazon Web Information Services (**AWIS**) API:
  http://aws.amazon.com/de/awis/

- Internet usage and broadband Penetration were obtained from the International Telecommunication Union (**ITU**) for 2013:
  http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

- Language and Distance Variables were obtained from the Centre d'Etudes Prospectives et d'Informations Internationales (**CEPII**):
  http://www.cepii.fr/cepii/en/bdd_modele/bdd.asp
- World Population was obtained from the World Bank. (**WB**):
  http://databank.worldbank.org/data/home.aspx

- Top Level Domains were obtained from the Internet Assigned Numbers Authority (**IANA**):

[17] http://www.trustedsource.org/download/ts_wd_reference_guide.pdf

http://www.iana.org/

- Web Page Classification and Riskiness was obtained from MacAfee threat intelligence centre:
  http://www.mcafee.com/threat-intelligence/domain/popular.aspx

Bibliography

Gabaix, Xavier, and Rustam Ibragimov. "Rank− 1/2: a simple way to improve the OLS estimation of tail exponents." *Journal of Business & Economic Statistics* 29, no. 1 (2011): 24-39.

Gelman, Andrew, and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006.

Kratz, Marie, and Sidney I. Resnick. "The QQ-estimator and heavy tails." *Stochastic Models* 12, no. 4 (1996): 699-724.

MacKinnon, James G., and Halbert White. "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *Journal of Econometrics* 29, no. 3 (1985): 305-325.

Software Packages[18]:

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Python Software Foundation. Python Language Reference, version 2.6.6. Available at http://www.python.org

Auguie B (2012). gridExtra: functions in Grid graphics. R package version 0.9.1. http://CRAN.R-project.org/package=gridExtra

Behnel, Stefan, Martijn Faassen, and Ian Bicking. "lxml: XML and HTML with Python." (2005). http://lxml.de/

Brown C. (2012). dummies: Create dummy/indicator variables flexibly and efficiently. R package version 1.5.6. http://CRAN.R-project.org/package=dummies

Dowle M., Short T., Lianoglou S., Srinivasan A. with contributions from R Saporta and E Antonyan (2014). data.table: Extension of data.frame. R package version 1.9.4. http://CRAN.R-project.org/package=data.table

[18] Only the packages used are cited; for dependencies please look at the packages.

Gelman A. and Su Yu-Sung (2014). arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R package version 1.7-07. http://CRAN.R-project.org/package=arm

Henningsen A. (2013). censReg: Censored Regression (Tobit) Models. package version 0.5-20. http://CRAN.R-project.org/package=censReg

Wes McKinney. Data structures for statistical computing in python. In Stefan van der Walt and Jarrod Millman, editors, Proceedings of the 9th Python in Science Conference, pages 51-56, 2010.

Warnes Gregory R., Bolker B., Bonebakker L., Gentleman R. , Liaw W.H.A, Lumley T., Maechler M., Magnusson A., Moeller S., Schwartz S. and Venables B. (2015). gplots: Various R Programming Tools for Plotting Data. R package version 2.16.0. http://CRAN.R-project.org/package=gplots

Wickham H. (2007). Reshaping Data with the reshape Package. Journal of Statistical Software, 21(12), 1-20. http://www.jstatsoft.org/v21/i12/.

Wickham H. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

Wickham H. (2012). stringr: Make it easier to work with strings. R package version 0.6.2. http://CRAN.R-project.org/package=stringr

Wickham, H. (2014). scales: Scale functions for graphics.. R package version 0.2.4. http://CRAN.R-project.org/package=scales

Wickham H. and Francois R. (2015). dplyr: A Grammar of DataManipulation. R package version 0.4.0. http://CRAN.R-project.org/package=dplyr

Zeileis A. and Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. http://CRAN.R-project.org/doc/Rnews/

Zeileis A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. Journal of Statistical Software 11(10), 1-17. URL http://www.jstatsoft.org/v11/i10/.