GRAVITY AND BORDERS IN ONLINE COMMERCE

BO COWGILL AND COSMINA DOROBANTU

ABSTRACT. We use a proprietary data set from Google to study the geographical patterns of cross-country transactions done over the internet. Covering online transactions over the course of four years, the data set allows us to shed light on this fast growing, but largely neglected area of trade. Although we present results for a worldwide level gravity equation, the focus of our paper is on the US-Canada border effect, a topic that has never been studied using online trade data. We find distance to have a smaller effect on trade between US states and Canadian provinces than the one estimated using traditional trade figures. More importantly, we find borders to have a large and significant impact on trade, indicating that the US-Canada border continues to hamper trade even in the online environment. A sector-by-sector analysis reveals that the border between the US and Canada has a larger effect on services that are heavily regulated, like finance, than on services that are less regulated, like professional, scientific and technical services.

1. INTRODUCTION

In 2011, $194 billion worth of products were sold online to internet users in the United States. These sales represented 5% of the country's total retail product sales.[1] In Canada, online sales reached $18 billion last year, with approximately half of the products being ordered from retailers outside of Canada.[2] The online retail market is large and it is one of the top growing segments, with double digit annual growth rates predicted for both the United States and Canada over the next five years. Yet, data availability limits our knowledge of this market and academic papers studying cross-border trade are rare.

The current article adds to the literature by using a proprietary data set from Google to study the US-Canada border effect. This is the first paper to use online trade data in order to shed new light on this well-known puzzle in international trade. Like previous studies that used online trade data in a gravity setting, we find distance to matter less in the virtual world. The estimated distance elasticity

Date: November 22, 2012.

[1] The figures come from the U.S. Department of Commerce's *Quarterly Retail E-commerce Sales Report.*

[2] The figures come from Statistics Canada and eMarketer's *Canada Retail E-commerce Forecast.*

of trade in the online environment is about 10% of that estimated in the traditional setting. If distance matters less, the US-Canada border effect shows no signs of abating in the virtual world. In our analysis, we show that the border between US and Canada has a large and negative impact on trade flows, even in this new, online environment. Estimates for the average border effect range from 3.8 in 2009 to 9.6 in 2011, indicating that depending on the year studied, intranational trade was 3.8 to 9.6 times higher than international trade. By comparison, Feenstra (2002) estimated an average border effect using traditional trade figures from 1993 of 4.7.

To get a better understanding of the border effect, we classify our data by NAICS sectors[3]. We conduct the only other analysis besides Anderson and Yotov (2012) that examines the US-Canada border effect for individual service sectors. The results show the border effect to be much larger for services that are heavily regulated, like finance, than for services that are less regulated, like professional, scientific and technical services.

2. Previous Literature

2.1. **Literature on Effects of Borders on Trade.**

2.1.1. *Original US/Canada Empirical Specification.*

The effects of the US-Canada border on trade flows is a well researched topic in international economics. A large literature about the effect of borders on international trade has focused on the US/Canada case. Scholars have focused on the US/Canada because it is the area where one would be least expect an international border to effect trade – the two countries are bound by a free trade agreement and have many common cultural, legal and economic traditions.

McCallum (1995) is the first paper to ask whether trade between Canadian provinces is higher or lower than trade between Canadian provinces and US states. He estimates a simple gravity model, where bilateral trade between two provences/states depends on the two regions' GDPs, the symmetric distance between them and a dummy variable taking the value of 1 for trade between two Canadian provinces and 0 for trade between a Canadian province and a US state. Surprisingly, the regression results return a very large coefficient on the dummy variable measuring

[3] North American Industry Classification System (NAICS), an industry classification system developed by statistical agencies from the US, Canada and Mexico to classify economic activity.

the border effect. Trade between two provinces is 22 times larger than trade between a province and a state. This result generated a vast literature in search of an answer to the border puzzle.

Anderson and van Wincoop (2003) was a major breakthrough in this search. They derive a gravity equation from formal economic theory, basing their model on differentiated goods and constant elasticity of substitution preferences. The model shows that bilateral trade between two regions depends not only on the regions' GDPs and the distance between them, but also on their implicit price indexes. The previous studies' failure to account for these indexes of multilateral resistance has two implications. First, it biases the Canada border effect upwards. Second, it masks the fact that border effects affect countries differently, and in particular, they have a larger effect on smaller countries. Anderson and van Wincoop (2003)'s theoretically grounded estimates of the impact of border barriers on intranational versus international trade paint a more plausible picture than McCallum (1995). They find that trade between two Canadian provinces is 10.5 times larger than trade between a province and a state. Trade between two US states, on the other hand, is only 2.56 times larger than trade between a province and a state, reflecting the fact that the US is an economically bigger country than Canada.

The methodology proposed by Anderson and van Wincoop (2003) has been criticized on two main grounds. First, the multilateral resistance terms are unobserved. Although they can be calculated, the available data allow only for the construction of crude approximations. Second, the price indices are endogenous, leading to estimation biases. In response to these criticisms, Feenstra (2002) suggests that rather than calculating and including the multilateral resistance terms, researchers include fixed effects for both the importing and the exporting regions. Because this approach gives consistent estimates of the average border effects and avoids the two criticisms mentioned above, it has become the standard in the literature.

2.1.2. *Improvements to the Basic Specification.*

Most studies assume that contiguity has a symmetric effect on trade. Brown and Anderson (2002) hypothesize that contiguity between two US states and contiguity between a province and a state might have a different effect on trade. Their regressions are on a per-industry basis. They find that contiguity between two US states has a positive and significant effect on trade for all industries considered, while contiguity between a US state and a Canadian province does not have a statistically significant effect for every industry. Examining this issue from a Canadian prospective, Anderson and Yotov (2010) find that at the aggregate level, contiguity between two Canadian provinces does not have a statistically significant effect on trade, while contiguity between a Canadian province and a US state has a positive and significant effect. Although the literature has not reached a consensus on the direction of these effects, the results of the papers mentioned above point to the importance of allowing trade costs to affect trade asymmetrically.

2.1.3. *Importance of Sectoral Differentiation.*

A recent set of papers written by Anderson, Milot, and Yotov point to the importance of examining the US-Canada border effect on data broken down by sector of activity. Aggregation biases gravity estimates. Anderson and Yotov (2010) focus on the manufacturing sectors and show that the aggregate border effects are lower than the average border effects estimated with commodity level data. Anderson, Milot and Yotov (2012), on the other hand, underline the need to estimate gravity models for each of the nine service sectors separately by showing that the coefficients on the border effect dummies vary significantly across these sectors.

2.2. **Internet Trade and Gravity Equations.**

A small literature at the intersection of computer science, economics and quantitative marketing has studied the relationship between economic geography and the internet. This literature has primarily focused on consumers' substitution between online and offline purchasing channels (for example, see Forman, Ghose and Goldfarb (2009), Brynjolfsson, Hu and Rahman (2009) or Goolsbee (2001)).

To our knowledge, no study has examined the Canada-US border effect using data relating to online trade. All studies published to date have used traditional trade measures, such as the value of exports between the Canadian provinces and the US states. Indeed, very few studies have used online transaction data to address any trade-related questions. Those that have addressed these questions have faced data limitations, some of which we are able to improve upon.

Lendle et. al (2012), Hortacsu, Asis Martinez-Jerez and Douglas (2009), and Blum and Goldfarb (2006) are the most convincing studies to estimate gravity equations on data relating to online activities. Lendle et. al (2012) work with eBay data and they have access to the most comprehensive data set out of the three papers mentioned above. The authors perform a gravity analysis focusing on 62 countries and find the effect of distance to be 65% smaller on eBay's online platform than offine.

Hortacsu, Asis Martinez-Jerez and Douglas (2009) look at geographical trade patterns using data from eBay and MercadoLibre, a South American website dedicated to e-commerce and online auctions. The eBay data cover US buyers and sellers over a four-month time frame and the location of the buyer is known for 27% of the transactions. This limits the authors to examine trade patterns only within the United States and only on a quarter of the transactions registered in their data set. The MercadoLibre data is more complete, but it only allows for an analysis of trade flows between nine South American countries. Two of their findings are relevant to the current study. First, the results show that distance has a negative effect on trade, but the value of the coefficient is only about 10% of that obtained by studies looking at the offline world. Second, the results also show that there is a positive and significant home state bias.

Blum and Goldfarb (2006) use data from a defunct ISP company to look at the internet activities of 2,654 American internet users, who visited websites from 46 countries over the span of three months. Their data is also limiting. They only cover page views, the country of origin of the website visited is difficult to determine and there are questions about the representativeness of the sample, given that there were over 100 million internet users in the United States in the year examined by the authors. They find that distance matters more for taste-dependent differentiated products and less for homogenous products.

3. Empirical Approach

3.1. **Empirical Specification.**

Using the theoretical framework proposed by Anderson and van Wincoop (2003), the following gravity equation can be easily derived:

$$x_{ij} = \frac{y_i y_j}{y_w} \left(\frac{t_{ij}}{P_i P_j} \right)^{1-\sigma} \tag{1}$$

where:

x_{ij} = region i's exports to region j
y_i = GDP of region i
y_j = GDP of region j
y_w = world GDP
t_{ij} = transport costs between i and j.
P_i = consumer price index of region i, also called a multilateral resistance term
P_j = consumer price index of region j
σ = elasticity of substitution between all goods

Taking the log of each side of equation (1), we obtain:

$$\ln x_{ij} = \ln y_i + \ln y_j - \ln y_w + (1-\sigma)\ln t_{ij} - (1-\sigma)\ln P_i - (1-\sigma)\ln P_j \tag{2}$$

The transport costs, t_{ij}, are hypothesized to be a loglinear function of the bilateral distance between the two regions, d_{ij}, as well as all the observable intranational and international border barriers, $b_{ij}(h)$, where h indexes these barriers.

$$\ln t_{ij} = \rho \ln d_{ij} + \sum_h \gamma_h b_{ij}(h) \tag{3}$$

Using (3) to substitute for $\ln t_{ij}$ in (2), realizing that $\ln y_w$ is a constant term, and including importer and exporter fixed effects to capture the effects of the regions' GDPs and of the unobserved multilateral resistance terms, we obtain the following operational econometric model:

$$\begin{aligned}\ln x_{ij} = {} & \beta_0 + \beta_1 \ln d_{ij} + \\ & + \beta_2 \text{ Contig CA-US} + \beta_3 \text{ Contig CA} + \beta_4 \text{ Contig US} + \beta_5 \text{ Internal} + \\ & + \beta_6 \text{ Border CA}\leftrightarrow\text{US} + \lambda_i + \chi_j + \varepsilon_{ij}\end{aligned}$$

where Contig CA-US, Contig CA, Contig US, Internal, Border CA→US and Border US→CA capture the observable intranational and international border barriers. In keeping with previous studies, we allow for asymmetric contiguity and border effects. We define the dummy variables as follows:

- Contig CA-US: 1 if a province and a state share a land border and 0 otherwise
- Contig CA: 1 if two provinces share a land border and 0 otherwise
- Contig US: 1 if two provinces share a land border and 0 otherwise

- Internal: 1 for a province's trade with itself and 0 otherwise
- Border CA↔US: 1 for exports between states and provinces and 0 otherwise

In the econometric model, λ_i and χ_j are exporter and importer fixed effects, respectively, and ε_{ij} is a the error term.

3.2. **Estimation Method.**

Gravity equations have traditionally been estimated using OLS, but this methodology has encountered heavy criticism for failing to properly account for zero trade flows. The lack of any trade between some regions is prevalent in the data, yet by taking the log of export flows, researchers are excluding these observations from their analyses. Two alternative estimation methods have gained popularity recently, the first being proposed by Helpman, Melitz and Rubinstein (2008) and the second being proposed by Santos and Tenreyro (2006).

Helpman, Melitz and Rubinstein (2008) propose a two-step estimation procedure to correct for sample selection. The first stage is a probit regression that yields a prediction for the probability that region i exports to region j. The second stage is a gravity equation estimated on the subset of positive trade flows. Santos and Tenreyro (2006) argue that the main drawback of the Helpman, Melitz and Rubinstein (2008) paper is the fact that the validity of the estimation procedure proposed rests on the assumption that all random components of the model are homoskedastic. Econometric tests presented in Santos and Tenreyro (2009) show that this assumption is unrealistic. To consistently estimate a gravity equation in the presence of heteroskedasticiy and zero trade flows, Santos and Tenreyro (2006) propose a Poisson pseudo maximum likelihood estimator (PPML). Monte Carlo simulations show that this estimator produces results that are robust to the patterns of heteroskedasticity prevalent in trade data, while log-linearized OLS regressions or estimations performed following the recommendations of Helpman, Melitz and Rubinstein (2008), do not.

Unlike the estimation strategy proposed by Santos and Tenreyro (2006), the method proposed by Helpman, Melitz and Rubinstein (2008) has the advantage of being firmly grounded in formal trade theory. Unfortunately, it relies on the existence of an exclusion restriction that affects the probability that two locations trade, but not the volume of trade between them. In their paper, the authors show that either the cost of starting a business or religion makes for a reasonable exclusion restriction. While these variables work on a gravity analysis at the world level, there is too little variation between the US states and the Canadian provinces to find suitable exclusion restrictions.

As data constraints rule out the possibility of using the method proposed by Helpman, Melitz and Rubinstein (2008), we test the standard OLS and PPML estimators to ascertain which one is best suited for our analysis of US-Canada border effects. With the right data, the PPML estimator has two advantages over standard OLS: it can deal with zero trade flows better than OLS and it does a better job at handling heteroskedasticity. Regarding the first advantage, our data set contains very few zero trade flows: they represent between 0.2% and 1.7% of the observations, depending on the year for which the data are reported. By comparison, in the worldwide data sets traditionally used for gravity analyses, zero trade flows represent over 50% of the observations. Our data, therefore, does not suffer from the same problem relating to zero trade flows as the data sets for which the PPML estimator was designed. Regarding the second advantage, we perform the RESET test proposed by Santos and Tenreyro (2006) to ascertain whether the OLS or the PPML estimator does a better job at addressing potential heteroskedasticity problems. We find that the estimators perform equally well. As there is no evidence that the PPML estimator is better suited for our data than standard OLS, we choose to use the latter for our analysis.

4. Data

4.1. **Google Data.**

4.1.1. *Description.*

The data we use to study the Canada-US border effect come from Google's online advertising platforms. In order to provide the clients of its advertising program, AdWords, with useful intelligence about their ads, Google offers free "conversion tracking" software. This software tracks users anonymously from viewing an ad to purchasing, signing up or other forms of online "conversions" that are valuable to businesses. It allows clients to measure the return on their investment in advertising in terms of dollars, transaction counts or signups. The transaction data we use in this study come from the records generated by the conversion tracking software. As a robustness check, in later sections we analyze subsets of the data for specific types of conversions.

We are confident that our data set is the most comprehensive one used to date to examine trade related questions in the digital world. However, like any other data set, the Google data we analyze have some limitations. We consider them to be minor compared to the limitations faced by the other studies that looked at online trade, yet is is important to clearly spell them out:

- *Counts only*: The conversion tracking code reliably generates conversion counts. Counts represent the number of times users reach the sections of the sites where the advertisers placed the code. For most advertisers this constitutes a transaction or sale – in later sections, we discuss restrictions on this data to ensure that it represents sales, as well as use of conversion values (sale amounts).

- *AdWords ads only*: The code will only track conversions generated as a result of users clicking on an ad placed on Google or on one of Google's partner sites. Although Google partners with millions of sites around the world, there are online transactions which will not be recorded in this data set. For example, if a user types www.ebay.com into their browser and purchases an object directly from www.ebay.com, that transaction will not be recored in the data set, as it happened outside of Google's advertising program.

- *30 day limit*: The code generates a conversion count if a user clicks on an AdWords ad and reaches the page where the code is placed within 30 days of the initial click on the ad. If the user reaches the page after 30 days, the system will no longer record the conversion.

- *Optional tool*: The conversion tracking tool is optional, so advertisers are not required to enable it. However, the adoption rate is high, as it is the easiest way for advertisers to assess whether their online advertising campaign are effective.

The location data associated with our transactions comes from two sources. For buyers, we use estimates based on IP (Internet Protocol) address. For sellers, we use the self-reported address data required of businesses who sign up to be Google advertisers.

Although we believe that this information allows us to place most buyers and sellers in the correct region, it is possible that we are attributing a small part of the buyers or sellers to the wrong location. On the buyers' side, the IP address might indicate an incorrect physical location if the user is accessing the internet using a virtual private network (VPN) connection or if the user is, for whatever reason, using software to mask his or her actual IP address. While we cannot identify these users, we believe that they represent a small enough percentage of the total internet users to make our location identification on the buyers' side reliable. On the seller side, the address field is self-reported, and it is possible that some sellers type in the incorrect or unrepresentative address. Although Google requests a general mailing address, a firm may choose to provide the address of its advertising office or billing department. This may be unrepresentative of the firm's general location. This issue we face on the sellers' side is not specific to our study. Indeed, any paper that studies the behavior of multinational enterprises faces similar problems when trying to pinpoint the geographical location of the seller.

To ensure that the data behave as expected, we also run a regression on worldwide conversion counts. We use the same specification as Helpman, Melitz and Rubinstein (2008) and we compare the results we generate using Google's conversion count data to the results the other researchers obtain using traditional trade flows. They differ in expected ways, indicating that our data capture the complexity of trade flows as well as the traditional figures used in worldwide gravity analyses. In the Appendix, we present the results we obtain on Google's data by replicating the same methodology that Helpman, Melitz and Rubinstein (2008) use to generate the results for Table 2 on page 463 of their paper. We highlight where our results differ.

4.1.2. *Coverage.*

In 2007, Google announced that AdWords, its advertising program, passed the one million mark in terms of number of advertising accounts. This is the last publicly available figure. It is meaningful in that it indicates that five years ago, there were already more than one million businesses opting for Google's advertising platform. The data set used here covers all conversion counts recorded over a period of four years, from 2008 to 2011. The data are available at the daily level, although we aggregate it to the yearly level to generate our main results. Positive conversion counts are reported for advertisers and sellers based in all 50 US states, the District of Columbia, the 10 Canadian provinces and the three Canadian territories. Although we cannot provide precise figures, the overall numbers of recorded conversion counts are well above 10 billion.

4.2. **Other Variables.**

Besides the dependent variable, the only other variable that requires explanation is our distance measure, d_{ij}. Like the majority of studies that use gravity equations, we measure the distance between region i and region j in kilometers and we calculate it using the Great Circle Distance Formula. For same state or same province trade, we calculate the internal distance using the methodology proposed in Head and Mayer (2002). Thus, we calculate the internal distance of region i using the formula $d_{ii} = 0.67\sqrt{\text{internal area}_i/\pi}$. All data on latitude and longitude come from the World Gazetteer web page. Data on the internal area of the US states come from the United States Census Bureau, while data on the internal area of the Canadian provinces and territories come from Statistics Canada.

5. Results

5.1. **Main Specification.**

We ran the regressions on data for each of the four years in our sample: 2008, 2009, 2010 and 2011. In terms of regions, we include all interstate, interprovincial and state-province conversions. We also include same state and same province conversions in order to be able to estimate the home market effect. All regressions have importer and exporter fixed effects. We report robust standard errors, clustered by region pair. Table 1 presents the OLS estimates, which use the log of Google's conversion counts as the dependent variable.

Table 1. Results for the Main US-CA Border Effects Specification

	Conversion counts (in logs)			
	2008	2009	2010	2011
$\ln d_{ij}$	-0.134***	-0.099***	-0.145***	-0.218***
	(0.015)	(0.013)	(0.018)	(0.023)
Contig CA-US	0.037	0.075	0.125	0.339***
	(0.075)	(0.066)	(0.091)	(0.129)
Contig CA	-0.347	0.353*	0.151	-0.245
	(0.260)	(0.197)	(0.231)	(0.276)
Contig US	0.084**	0.122***	0.134***	0.040
	(0.034)	(0.028)	(0.033)	(0.042)
Internal	1.220***	1.359***	1.504***	1.435***
	(0.158)	(0.137)	(0.183)	(0.203)
Border CA↔US	-1.423***	-1.332***	-1.963***	-2.265***
	(0.054)	(0.049)	(0.058)	(0.086)
Border average	4.2	3.8	7.1	9.6
Obs.	4,026	4,022	3,963	3,971
R^2	0.974	0.977	0.977	0.970

Notes: Importer and exporter fixed effects included.
Robust standard errors (clustering by region pair).
*** Significant at 1%, ** significant at 5%, * significant at 10%

In studies that analyze trade flows in a gravity setting, distance has an unequivocally negative effect. In our US-Canada analysis, this is still the case, though the effect of distance is much smaller than that estimated by researchers working with traditional trade data. The magnitude of the coefficient estimated by us using Google data is similar to that estimated by Hortacsu, Martinez-Jerez and Douglas (2009). Like us, they find that for eBay transactions conducted between US states, the coefficient on distance is about 10% of that estimated on offline trade flows.

If distance has a negative effect on trade, the home bias influences it positively. The coefficient of the *Internal* dummy variable is economically and statistically significant. This result is in line with previous work. Like us, Hortacsu, Martinez-Jerez and Douglas (2009) find that the home bias effect is very large and persistent. Contiguity has a positive and significant effect for trade between US states, but not for trade between Canadian provinces or between a US state and a Canadian province for all years in our sample except 2011. The results we obtain for 2008, 2009 and 2010 support those of Brown and Anderson (2002) who also find that sharing a land border matters for trade between American states, but not for trade between states and provinces. Anderson and Yotov (2010), on the other hand, find that contiguity has a positive and significant effect on trade between states and provinces. We obtain the same result only for 2011. Like us, they also find that sharing a land border makes no difference for trade between Canadian provinces. Our results, as well as those reported by the other authors, underline the importance of allowing contiguity to have asymmetric effects.

The Canada-US border continues to impact trade flows, even in the digital age. We ran our regressions with symmetric border effects at first. The coefficient of the CA↔US dummy variable is large, negative and statistically significant. Taking the exponent of the coefficient of the CA↔US dummy gives the average border effect, which ranges between 3.8 in 2009 to 9.6 in 2011. By comparison, the average border effect estimated by Feenstra (2002) using traditional trade figures from 1993 was 4.7. If distance matters less for trade between US states and Canadian provinces in the online environment, the border effect appears to hinder trade just as much, if not more, than it does in the offline world.

5.2. **Robustness Checks for the Main Specification.**

5.2.1. *Limiting the Data Set to 30 States and 10 Provinces.*

Most studies that examine border effects, such as McCallum (1995), Anderson and van Wincoop (2003), and Feenstra (2002) do not consider all US states and Canadian provinces and territories in their analyses, and they also exclude same-state trades. The authors look solely at the trades that take place between the top 30 most populous US states and the 10 Canadian provinces. To ensure that the estimates we obtain for the border effects are not a result of us considering all states, provinces and territories, we run our regressions on a subset of our data that includes only the 30 US states and the 10 Canadian provinces that the other authors studied and that excludes internal trade. We present our results in Table 2. The signs and the magnitudes of the border effect coefficients are almost the same as the ones we obtain when working with the complete data set. The size of the coefficient on CA↔US dummy variable ranges between -1.34 and -2.19 for this smaller data set, while that of the coefficient on the CA↔US dummy variable estimated using the complete data set ranges between -1.33 and -2.27.

Table 2. Subset of 30 States and 10 Provinces

	Conversion counts (in logs)			
	2008	2009	2010	2011
$\ln d_{ij}$	-0.179***	-0.116***	-0.191***	-0.267***
	(0.027)	(0.019)	(0.034)	(0.050)
Contig CA-US	-0.061	0.005	0.022	0.167
	(0.068)	(0.069)	(0.090)	(0.140)
Contig CA	-0.067	0.230	0.172	-0.274
	(0.247)	(0.185)	(0.275)	(0.289)
Contig US	0.018	0.096**	0.092	-0.025
	(0.064)	(0.041)	(0.057)	(0.079)
Border CA↔US	-1.353***	-1.342***	-1.900***	-2.187***
	(0.066)	(0.042)	(0.067)	(0.103)
Border average	3.9	3.8	6.7	8.9
Obs.	1,560	1,560	1,557	1,560
R^2	0.971	0.980	0.978	0.965

Notes: Importer and exporter fixed effects included.
Robust standard errors (clustering by region pair).
*** Significant at 1%, ** significant at 5%, * significant at 10%

5.2.2. *Conversion Values.*

Traditional gravity equations in trade are estimated using the value of exports and imports, while the current study focuses on the number of transactions. This might be misleading, as we inevitably assign the same importance to a small transaction as we do to a large one. Google does have data regarding conversion values, although the figures are not quite as reliable as the ones reported for the conversion counts. When setting up the conversion tracking code, advertisers can specify the value of each conversion. This value then gets multiplied by the number of conversion counts and reports are generated within the advertisers' Google AdWords accounts which compare their total advertising spend to their total profit. Although the conversion values are advertiser-reported, there is reason to believe that they closely match reality, as it is on the basis of these values that the advertisers can make the proper calculations as to how much they are spending and how much they are making from their online ads.

Table 4 reports the results we obtain when running our main specification on our complete data set of conversion values. The results are similar to the ones reported in Table 1. The coefficients estimated for distance are approximately twice as large as those reported in Table 1, but still much lower than the ones obtained by researchers using traditional trade figures. The home bias, as well as the CA-US border effect, have a large and significant effect on trade, confirming the findings reported in Table 1.

Table 3. Conversion Values

	Conversion values (in logs and US dollars)			
	2008	2009	2010	2011
$\ln d_{ij}$	-0.221*** (0.044)	-0.204*** (0.039)	-0.242*** (0.044)	-0.333*** (0.047)
Contig CA-US	0.100 (0.278)	0.156 (0.466)	-0.418 (0.280)	-0.401 (0.356)
Contig CA	0.064 (0.491)	1.041** (0.486)	0.845** (0.386)	0.140 (0.426)
Contig US	0.106 (0.097)	0.265*** (0.083)	0.290*** (0.085)	0.040 (0.096)
Internal	2.629*** (0.330)	2.901*** (0.297)	3.006*** (0.333)	2.748*** (0.317)
CA↔US	-2.256*** (0.095)	-1.804*** (0.097)	-2.317*** (0.100)	-2.717*** (0.105)
Border average	9.5	6.1	10.1	15.1
Obs.	4,026	4,022	3,945	3,903
R^2	0.931	0.939	0.937	0.916

Notes: Importer and exporter fixed effects included.
Robust standard errors (clustering by region pair).
*** Significant at 1%, ** significant at 5%, * significant at 10%
Border dummy variable definition:
CA↔US = 1 for all state-province trade and 0 otherwise

5.2.3. *Conversion Types.*

As previously mentioned, online conversions can constitute a variety of advertiser-specified events. Many advertisers choose to measure conversions that do not immediately lead to money changing hands. For example, they may track mailing list signups. Although they may not lead immediately to purchases, we believe all conversions tracked have economic value to the advertisers. Conversions often lead to future sales and advertisers spend money to promote their sites on Google's platform in order to generate these conversions.

As a robustness check, we also study the aforementioned regression models on a sample of purchase-related data only. Google allows its advertisers to select the type of conversion they are tracking. Selecting a type is optional and the selection must fall into one of the categories listed below:

TABLE 4. Conversion types

Conversion type	Conversion is due to:
Purchase	A purchase, sales, or "order placed" event
Default	An unspecified category
Signup	A sign-up user action
Lead	A lead-generating action
Page view	A user visiting a page
Login	A user logging into an pre-existing account
Shopping-cart post	An item put into a shopping cart
Order charged	A purchase or order that was successfully charged for
Install	A software install action
Download	A software download action
Referral	A user bonus for referrals of new customers

Studies conducted using offline trade data look only at sales from exporting countries to importing countries. To ensure that the results generated in Tables 1 and 2 above are not just an artifact, we will restrict our analysis to conversions that fall into one of the three categories above that most likely describe a sale: purchase, shopping-cart post and order charged. Table 5 presents the results, and the estimated average border effect is even higher than we reported in Table 1, confirming, once more, that the border effect is alive and well in the virtual world.

TABLE 5. Conversion Types

	Conversion counts, purchases only (in logs)			
	2008	2009	2010	2011
$\ln d_{ij}$	-0.129*** (0.017)	-0.145*** (0.017)	-0.181*** (0.020)	-0.207*** (0.019)
Contig CA-US	-0.217* (0.118)	-0.215** (0.095)	-0.125 (0.129)	-0.083 (0.127)
Contig CA	-0.302 (0.221)	-0.069 (0.249)	-0.090 (0.257)	-0.114 (0.291)
Contig US	0.037 (0.034)	0.035 (0.031)	0.028 (0.036)	-0.027 (0.035)
Internal	0.879*** (0.139)	0.942*** (0.156)	0.936*** (0.190)	0.998*** (0.176)
CA↔US	-1.583*** (0.055)	-1.820*** (0.058)	-2.110*** (0.072)	-2.152*** (0.063)
Border average	4.9	6.2	8.3	8.6
Obs.	3,947	3,955	3,929	3,871
R^2	0.971	0.973	0.972	0.970

Notes: Importer and exporter fixed effects included.
Robust standard errors (clustering by region pair).
*** Significant at 1%, ** significant at 5%, * significant at 10%

6. Sector Level Analysis

6.1. **Data Broken Down by NAICS2 Sectors.**

Google has automatic mechanisms in place that classify the ads running on its online platforms into verticals. Google's verticals are structured like NAICS sectors. There are 27 top level categories, which are similar to the 2-digit NAICS sectors. These are the broadest classifications available. Under these top level verticals, there are 241 second and third level categories, which are similar to the 4-digit and 6-digit NAICS classifications. An algorithm assigns each search to the relevant verticals and subverticals. For example, a search for [car tires] would be classified under the third level category 'Vehicle Tires,' the second level category 'Auto Parts' and the top level vertical 'Automotive.'

To map Google's verticals to NAICS sectors, we use a classification scheme designed by Google Cheif Economist Hal Varian in Choi and Varian (2012). Where Choi and Varian (2012) did not provide a NAICS/vertical mapping, we assigned an encoding ourselves. A few examples of how the mapping is done are provided below:

Google Vertical		**NAICS Sector**	
ID	Title	ID	Title
47	Automotive	441	Motor vehicle and parts dealers
5	Computers and Electronics	443	Electronics and appliance stores
1868	Apparel	448	Clothing and clothing accessories stores
...	...	...	...

We manage to successfully assign a 2-digit NAICS classification to over 98.5% of the conversions in our data set. In other words, less than 1.5% of our data remain unclassified. Only 1.5% of the conversions reflected in our data set fall within the Agriculture, Mining, Construction and Manufacturing sectors, leaving the vast majority of the conversions, 97%, reflected within one of the service sectors[4]. We run our main specification for each 2-digit NAICS sector.[5] Detailed results are available upon request.

The first three columns of Table 6 present the coefficient for distance obtained

[4] 32% of our data are classified as "Retail Trade" or "Wholesale Trade." 99% of this fell into "Retail Trade" alone.

[5] The only service sector that we cannot map any conversions to is 'Management of Companies and Enterprises.'

by applying our main specification to each NAICS sector. The next three columns report the exponent of the coefficient of the same-state dummy variable, which gives the home bias effect. Finally, the last three columns contain the exponent of the coefficient of the Canada-US border dummy, which gives the average border effect. Most coefficients are significant at the 1% level. We write the coefficients that are not significant at the 1% level in italics.

The results are intuitive. Some services are bound to a particular location. Real Estate, for example, has the highest elasticity of distance out of all the sectors and by far the largest home bias. Other services, such as Public Administration, are designed for national consumption and will travel very poorly across borders, as the estimated average border effect shows. The interesting results are for sectors such as Finance and Insurance, which have a small estimated home bias, indicating that they travel well outside of the region where they are based, but a large border effect, indicating that they do not travel well internationally. Differences in regulation are the most likely culprit. By comparison, a sector that is not heavily regulated, like Retail Trade, has a similar elasticity of distance and home bias to Finance and Insurance, but a much lower average border effect.

Table 6. Results at the NAICS 2-digit level

NAICS Sector	**Distance**				**Internal**				**Border CA↔US**			
OLS estimates for:	2008	2009	2010	2011	2008	2009	2010	2011	2008	2009	2010	2011
Wholesale trade	-0.1	-0.1	-0.1	-0.1	1.8	1.7	1.6	2.0	6.3	5.9	5.4	6.0
Retail trade	-0.1	-0.1	-0.1	-0.1	2.4	2.8	3.3	3.5	2.7	3.7	5.1	4.9
Transp., warehousing	-0.1	-0.2	-0.2	-0.1	3.1	4.1	4.7	6.8	9.8	11.0	10.7	14.8
Utilities	-0.3	-0.2	-0.2	-0.3	3.0	3.2	8.6	8.5	2.2	1.9	2.8	6.1
Information, cultural ind.	-0.2	-0.2	-0.2	-0.3	4.0	6.1	6.6	5.8	2.9	3.7	4.6	6.2
Finance, insurance	-0.1	-0.1	-0.1	*-0.1*	2.5	2.6	3.1	4.6	8.7	11.6	12.6	17.1
Real estate, rental, leasing	-0.3	-0.3	-0.4	-0.5	20.1	14.8	17.2	18.8	3.7	8.0	8.3	6.4
Prof, sci, and tech services	-0.1	-0.1	-0.1	-0.1	2.8	3.5	4.8	5.4	4.0	4.1	6.0	7.1
Administrative, support	-0.3	-0.2	-0.1	-0.2	2.6	3.6	4.1	3.9	10.0	7.5	16.8	18.0
Education services	-0.1	-0.1	-0.1	-0.2	3.9	5.0	5.4	5.9	9.0	9.9	10.4	13.8
Health care, social assist	-0.1	-0.1	-0.1	-0.2	3.7	4.7	4.0	4.5	3.9	6.2	8.0	14.1
Arts, entertain, recreation	-0.2	-0.2	-0.2	-0.3	3.2	3.3	3.7	3.7	2.5	2.9	3.8	7.9
Accommodation, food	-0.3	-0.2	-0.3	-0.3	3.8	8.0	6.4	6.5	5.7	6.5	9.5	11.6
Other services	-0.1	-0.1	-0.1	-0.2	3.3	3.6	4.0	4.1	2.5	3.1	3.3	5.5
Public administration	-0.1	*-0.0*	*-0.0*	*-0.0*	2.5	3.3	3.1	4.4	12.3	16.0	19.5	14.7

The sector-by-sector analysis might also shed some light on the increase in the border coefficient over the past 4 years. There are several sectors for which the border coefficient increased consistently from year to year and it doubled, or more than doubled, in size between 2008 and 2011. These sectors are Arts, Entertainment and Recreation, Accommodation and Food Services, Finance and Insurance, Health Care and Social Assistance, and Information and Cultural Industries. It is possible that the economic crisis, started by the collapse of Lehman Brothers in September 2008, contributed to the increase in the border effect. There is no doubt that the financial sector became more national in character as a result of the crisis. There is little surprise, also, that the demand for cross border entertainment, tourism or health care would go down, as evidenced by the increasing average border effect for the Arts, Entertainment and Recreation, Accommodation and Food Services, and Health Care and Social Assistance sectors. It is beyond the scope of this paper to test this hypothesis, but we do believe that examining it in depth is a promising avenue of future research.

6.2. **Data Broken Down by NAICS3 Sectors.**

For a limited set of goods and services transacted online, we are able to categorize the conversions into NAICS 3-digit categories. We can do this for a few of the NAICS3 Retail Trade and Finance and Insurance subsectors. The results we obtain for Retail Trade are presented in Table 7. For goods that are fairly standard and have precise specifications, such as Motor Vehicles and Parts, Electronics and Appliances, Building Materials, and Health and Personal Care products, the magnitude of the US-Canada border effect is similar to the magnitude of the home bias. The international border, therefore, causes similar hindrances to trade as the home state or province border. This might be because there is little risk involved in ordering the wrong product. A Ford alternator, an iPhone, a certain type of paint or a Crest toothpaste are likely to be the same, whether purchased in the US or Canada. By comparison, Clothing has an estimated border effect that is much higher than the estimated home bias. Unlike the other goods mentioned, clothing is unlikely to have precise specifications. A medium-sized shirt purchased in the US vs. a medium purchased in Canada is likely to feel and fit differently. Finally, Food and Beverages also have a high border effect, most likely due to regulatory differences between the two countries.

Table 7. Results at the NAICS 3-digit level: Retail Trade

NAICS Sector	**Distance**				**Internal**				**CA↔US**			
OLS estimates for:	2008	2009	2010	2011	2008	2009	2010	2011	2008	2009	2010	2011
Motor Vehicle, Parts	-0.1	-0.2	-0.2	-0.3	4.9	7.1	6.7	8.1	4.2	5.3	5.6	6.6
Furniture	-0.2	-0.1	-0.1	-0.2	2.0	2.1	2.4	2.4	7.4	7.0	7.9	11.8
Electronics, Appliances	-0.1	*0.0*	*0.0*	*-0.0*	1.6	2.5	2.0	2.6	1.7	3.0	3.9	3.3
Building Materials	-0.1	-0.1	-0.2	-0.2	3.1	3.0	5.7	7.8	4.7	5.8	5.4	6.1
Food, Beverages	-0.1	-0.2	-0.2	-0.1	3.2	2.6	2.6	3.3	11.6	4.9	18.1	23.5
Health, Personal Care	-0.1	-0.1	-0.1	-0.1	2.2	2.3	2.8	5.2	2.7	2.3	3.2	4.6
Clothing	-0.1	-0.1	-0.1	-0.2	1.8	2.0	2.0	2.0	6.6	9.2	7.0	7.2
Sporting Goods, Hobby	*-0.0*	*-0.1*	*-0.0*	-0.1	2.1	2.0	2.4	2.8	4.0	4.8	4.5	3.7

We can also examine three subcomponents of the Finance and Insurance sector, namely Credit Intermediation, Securities and Investments, and Insurance Carriers. The estimated average border effect for Securities and Investments is much lower than that estimated for Credit Intermediation or Insurance Carriers. This indicates the fact that investing in mutual funds across the border is much easier than obtaining an international credit card or insurance policy. The border effect most likely reflects, once more, the difference in regulation between the US and Canada. Another result worth noting in Table 8 is that the border effect increases dramatically and consistently from 2008 to 2011 for both Credit Intermediation and Insurance Carriers, subsectors that were heavily affected by the economic crisis.

Table 8. Results at the NAICS 3-digit level: Finance and Insurance

NAICS Sector	**Distance**				**Internal**				**CA↔US**			
OLS estimates for:	2008	2009	2010	2011	2008	2009	2010	2011	2008	2009	2010	2011
Credit Intermediation	*-0.1*	-0.1	-0.2	*-0.1*	2.6	2.3	2.4	3.0	11.4	15.0	29.4	25.2
Securities, Investments	*0.0*	*-0.0*	*0.0*	*0.1*	1.8	1.4	2.1	2.4	3.9	3.1	3.6	5.7
Insurance Carriers	*-0.1*	*-0.1*	*-0.1*	*-0.1*	6.7	6.5	5.4	8.4	21.7	35.8	82.4	123.1

7. Conclusion

In this study, we use a proprietary data set from Google to examine the US-Canada border puzzle from a new perspective. Our primary finding is that the border effect has a large and negative effect on trade between US states and Canadian provinces, even in the online environment. This effect seems to be stronger for services that are heavily regulated and for goods that do not have standard measurements and specifications.

The emergence of online trade creates new opportunities and challenges for the economics of trade. Technology enables buyers and sellers in different countries to reduce their search and, in some cases, transportation costs. We believe there is a substantial number of interesting trade economics questions that can be studied with the increasing size, richness and prevalence of datsets relating to online trade.

8. Appendix

Table 9 reproduces the results presented in the original Helpman, Melitz and Rubinstein (2008) paper. Table 9, on the other hand, presents the results we obtain when we apply the methodology proposed by Helpman, Melitz and Rubinstein (2008) to Google data.

Table 9. Helpman, Melitz and Rubinstein (2008) Results

	Offline: 1986				
	OLS	Probit	NLS	Polynomial	Non Parametric
	(Benchmark)				(100 bins)
Distance	-1.167***	-0.213***	-0.813***	-0.847***	-0.789***
	(0.040)	(0.016)	(0.049)	(0.052)	(0.088)
Land border	0.627***	-0.087	0.871***	0.845***	0.863***
	(0.165)	(0.072)	(0.170)	(0.166)	(0.170)
Island	-0.553**	-0.173**	-0.203	-0.218	-0.197
	(0.269)	(0.078)	(0.290)	(0.258)	(0.258)
Landlock	-0.432**	-0.053	-0.347**	-0.362*	-0.353*
	(0.189)	(0.050)	(0.175)	(0.187)	(0.187)
Legal	0.535***	0.049***	0.431***	0.434***	0.418***
	(0.064)	(0.019)	(0.065)	(0.064)	(0.065)
Language	0.147*	0.101***	-0.030	-0.017	-0.036
	(0.075)	(0.021)	(0.087)	(0.077)	(0.083)
Colony	0.909***	-0.009	0.847***	0.848***	0.838***
	(0.158)	(0.130)	(0.257)	(0.148)	(0.153)
Currency	1.534***	0.216***	1.077***	1.150***	1.107***
	(0.334)	(0.038)	(0.360)	(0.333)	(0.346)
FTA	0.976***	0.343***	0.124	0.241	0.065
	(0.247)	(0.009)	(0.227)	(0.197)	(0.348)
Religion	0.281**	0.141***	0.120	0.139	0.100
	(0.120)	(0.034)	(0.136)	(0.120)	(0.128)
Regulation costs	-0.146	-0.108***			
	(0.100)	(0.036)			
Days & proc.	-0.216*	-0.061**			
	(0.124)	(0.031)			
δ (from $\hat{\bar{\omega}}^*_{ij}$)			0.840***		
			(0.043)		
$\hat{\bar{\eta}}^*_{ij}$			0.240**	0.882***	
			(0.099)	(0.209)	
$\hat{\bar{z}}^*_{ij}$				3.261***	
				(0.540)	
$\hat{\bar{z}}^{*2}_{ij}$				-0.712***	
				(0.170)	
$\hat{\bar{z}}^{*3}_{ij}$				0.060***	
				(0.017)	
Obs.	6,602	12,198	6,602	6,602	6,602
R^2	0.693	0.573		0.701	0.706

Notes: Exporter and importer fixed effects. Marginal effects at sample means and pseudo R^2 reported for Probit. Regulation costs are excluded variables in all second stage specifications. Bootstrapped standard errors for NLS; robust standard errors (clustering by country pair) elsewhere.

*** significant at 1%, ** significant at 5%, * significant at 10%

Table 10. Worldwide Gravity Results

	Online: 2011				
	OLS (Benchmark)	Probit	NLS	Polynomial	Non Parametric (100 bins)
Distance	-0.667*** (0.029)	-0.177*** (0.012)	-0.574*** (0.032)	-0.591*** (0.035)	-0.597*** (0.035)
Land border	0.787*** (0.125)	0.112* (0.059)	0.809*** (0.122)	0.817*** (0.125)	0.808*** (0.125)
Island	0.106 (0.114)	-0.122*** (0.039)	0.154 (0.110)	0.144 (0.112)	0.136 (0.113)
Landlock	0.209** (0.096)	0.074* (0.039)	0.161* (0.094)	0.172* (0.096)	0.167* (0.097)
Legal	0.051 (0.035)	0.021 (0.015)	0.033 (0.034)	0.038 (0.035)	0.041 (0.035)
Language	1.357*** (0.070)	0.300*** (0.017)	1.217*** (0.077)	1.234*** (0.081)	1.237*** (0.082)
Colony	0.512*** (0.189)	0.149* (0.078)	0.515*** (0.179)	0.532*** (0.183)	0.534*** (0.184)
Currency	0.084 (0.117)	0.068 (0.075)	0.031 (0.118)	0.049 (0.120)	0.045 (0.121)
FTA	0.786*** (0.074)	0.218*** (0.030)	0.641*** (0.073)	0.674*** (0.075)	0.678*** (0.075)
Religion	0.834*** (0.083)	0.124*** (0.033)	0.802*** (0.081)	0.796*** (0.083)	0.790*** (0.083)
Warehouse	0.054 (0.057)	-0.077*** (0.024)			
δ (from $\hat{\omega}^*_{ij}$)			0.239*** (0.069)		
$\hat{\bar{\eta}}^*_{ij}$			1.651*** (0.066)	1.701*** (0.134)	
$\hat{\bar{z}}^*_{ij}$				1.592*** (0.351)	
$\hat{\bar{z}}^{*2}_{ij}$				-0.271** (0.109)	
$\hat{\bar{z}}^{*3}_{ij}$				0.018 (0.011)	
Obs.	10,541	20,532	10,541	10,541	10,541
R^2	0.787	0.716		0.805	0.807

Notes: Exporter and importer fixed effects. Marginal effects at sample means and pseudo R^2 reported for Probit. Regulation costs are excluded variables in all second stage specifications. Bootstrapped standard errors for NLS; robust standard errors (clustering by country pair) elsewhere.
*** significant at 1%, ** significant at 5%, * significant at 10%

The only difference between our methodology and Helpman, Melitz and Rubinstein's (2008) is the fact that they use two regulation cost measures as their exclusion restrictions, while we use a variable measuring the number of procedures required to build a warehouse as ours.

The results differ in expected ways. Distance still matters in the virtual world, though less than in the offline world. This finding is in line with previous studies done using online trade data. Cultural affinities, such as a shared language or religion, have a significant and positive effect on online trade, while they have an insignificant effect on traditional trade. The finding regarding language, in particular, is not surprising: in the online environment, the seller and the buyer interact via a website that must be written in a language that they can both understand. Finally, a shared currency influences traditional trade significantly and positively, while it has no statistically significant effect on online trade. This finding is likely due to the fact that in the online environment, currency is easily converted by automatic means, which facilitates trades between countries that do not use the same currency.

For variable definitions and methodology, please see the original Helpman, Melitz and Rubinstein (2008) paper.

9. Works Cited

Anderson, James E., and Eric van Wincoop. “Gravity with Gravitas: A Solution to the Border Puzzle.” *American Economic Review* 93.1 (2003): 170-92.

Anderson, James E., and Yoto V. Yotov. “The Changing Incidence of Geography.” *American Economic Review* 100.5 (2010): 2157-86.

Anderson, James E., Catherine Milot and and Yoto V. Yotov. “How Much Does Gravity Deflect Services Trade?” *International Economic Review* (2012): forthcoming.

Bergstrand, Jeffrey H., Peter Egger, and Mario Larch. *Gravity Redux: Structural Estimation of Gravity Equations and Asymmetric Bilateral Trade Costs.* University of Notre Dame, 2007.

Blum, Bernardo, and Avi Goldfarb. “Does the Internet Defy the Law of Gravity?” *Journal of International Economics* 70.2 (2006): 384-405.

Brynjolfsson, Erik, Hu Yu, and Mohammad Rahman. “Battle of the Retail Channels: How Product Selection and Geography Drive Cross-channel Competition.” *Management Science* 55.11 (2009): 1755-1765.

Brown, W. Mark, and William P. Anderson. “Spatial Markets and the Potential for Economic Integration between Canadian and U.S. Regions.” *Papers in Regional Science* 81.1 (2002): 99-120.

Forman, Chris, Anindy Ghose, and Avi Goldfarb. "Competition between Local and Electronic Markets: How the Benefit of Buying Online Depends on Where You Live." *Management Science* 55.1 (2009): 47-57.

Feenstra, Robert C. "Border Effects and the Gravity Equation: Consistent Methods for Estimation." *Scottish Journal of Political Economy* 49.5 (2002): 491-506.

Goolsbee, Austin. "Competition in the Computer Industry: Online Versus Retail," *Journal of Industrial Economics*, 49.4 (2001): 487-499.

Head, Keith, and Thierry Mayer. *Illusory Border Effects: Distance Mismeasurement Inflates Estimates of Home Bias in Trade.* CEPII research center, 2002.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein. "Estimating Trade Flows: Trading Partners and Trading Volumes." *The Quarterly Journal of Economics* 123.2 (2008): 441-87.

Hortacsu, Ali, F. Asis Martinez-Jerez, and Jason Douglas. "The Geography of Trade in Online Transactions: Evidence from eBay and MercadoLibre." *American Economic Journal: Microeconomics* 1.1 (2009): 53-74.

Lendle, Andreas, Marcelo Olarreaga, Simon Schropp, and Pierre-Louis Vezina. *There Goes Gravity: How eBay Reduces Trade Costs.* Centre for Economic Policy Research Discussion Paper No. 9094, 2012.

McCallum, John. "National Borders Matter: Canada-U.S. Regional Trade Patterns." *American Economic Review* 85.3 (1995): 615-23.

Ramsey, J. B. "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis." *Journal of the Royal Statistical Society*, Series B.31 (1969): 350-71.

Roy, Francis. "Cross-border Shopping and the Loonie: Not What It Used to Be." *Canadian Economic Observer* (2007).

Silva, J. M. C. Santos, and Silvana Tenreyro. "The Log of Gravity." *The review of economics and statistics* 88.4 (2006): 641-58.

—. *Trading Partners and Trading Volumes: Implementing the Helpman-Melitz-Rubinstein Model Empirically.* Centre for Economic Performance, LSE, 2009.