Market Design, Borders, and Gravity in the Virtual World

Cosmina Liana Dorobanțu
Balliol College
University of Oxford



Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Economics
Trinity Term 2015

To my parents,
Carmen and Vasile Dorobanțu,
For their unwavering support

To my sister,
Ruxandra Dorobanțu,
For her steady companionship

To my grandmother,
Ana Joița,
For her endless love

To my American host parents,
Karin Jacobson and Bob Seeman,
For believing in me

Acknowledgements

I am grateful to Beata Javorcik for working with me on both my MPhil and my DPhil theses. She has provided numerous suggestions and comments throughout the years, and I am a much better researcher as a result of the time and effort she invested in me.

Steve Bond, Paul Klemperer, and Meg Meyer provided invaluable support during my time at Oxford. Steve convinced me to teach and my work with students was one of the most enriching experiences I have had within this university. Paul and Meg advised me at crucial points in my academic career and I am thankful for their thoughts, insights, and time.

Professionally, I wouldn't be where I am today without Kim Clausing and Bill Kipp. Over the past 15 years, they have been my strongest supporters and I cannot thank them enough for that. Kim introduced me to economic research and provided amazing guidance and support at every stage of my academic career. Bill, on the other hand, worked tirelessly to help me build my career at Google.

I am lucky to have wonderful friends. Ever since I met them, Kim Gupta, Vince Gueriguian, Ioana Hyde, Brandon Brockmyer, and Blake Goud have made my life infinitely more beautiful. Kim travelled the world with me, and with every trip we took, we laughed more, we shared more, we learned more. I am so thankful for having her in my life. Vince is one of the few people who can talk me out of taking things too seriously and I am grateful to him for that. Ioana and I started out by sharing a desk in high school. As we navigated into adulthood, we ended up sharing much more and I am thankful for her friendship and for being so close to me all these years. In my undergraduate thesis, I thanked Brandon for being in close competition with my parents over whose shoulder I cried on the most. Not much has changed since; he is an incredibly supportive friend

and I am grateful to him for that. Blake was my Econ study buddy at Reed; he helped me get through Jeff Parker's macro class and shared the cookies he received from his Mom with me at a time when my Mom was too far away to send me cookies of my own. I am grateful for all the times he has been by my side.

At Oxford, Sang Seok Lee, Nick Woolley, Ruth Cheadle, Felix Chow, Silja Baller, and Banu Demir were wonderful companions throughout my MPhil and DPhil journeys. I am thankful to each one of them. Every time I needed someone to talk to, Sang was there for me; it meant the world to me, on many occasions. On my birthdays, Nick used to show up at my doorstep with cake and candles. I will always treasure those memories. Ruth and Felix joined me for countless walks around Oxford, rain or shine. Silja and Banu were my study buddies and they kept me grounded during some of the most stressful times of my life.

A boisterous group of Oxford-based Romanians ensured that I had a home away from home throughout my studies. I am so grateful, to each and every one of them. Ever since she moved to Oxford, Cora Miron brought so much joy, music, and fun into my life. I treasure her friendship, company, and support, and I am thankful for the countless times she helped me gain perspective. Mihai Vidrighin is one of the kindest people I've ever met. Alexandra Tomescu provided constant encouragement and support throughout my DPhil, especially during the month prior to submission. Adina Brădeanu is, by far, the best neighbor I've ever had. Mariela Neagu and Daniel Bălănescu welcomed me into their home at any time of the day or night and shared their art and poetry with me. Raluca David rekindled my love for the mountains and for long, meaningful chats in front of the fire.

Mo Percival taught me more about myself than I ever thought I would be able to learn. I am thankful to her for bringing so much color and meaning into my life. I am also grateful to Aniela Sârbu, who accompanied me along this journey, sharing stories, thoughts, and feelings along the way.

Elise and Théodore, my niece and my nephew, were a constant reminder throughout my DPhil of what life is all about. We jumped on the trampoline, read stories about

dragons, ran around parks and forests, danced, sang, and laughed. I am thankful to them – and to Olivier Le Gouanvic – for being such a big part of my life and for their healthy outlook on the world.

Lizzie and Charlie Seeman, were, by far, my closest friends during my first year studying abroad. I am so grateful for their warmth, love, hugs, and the many laughs we shared.

They are not here to see me graduate, but they are still very much in my thoughts. I am grateful to Marin Joița, as well as Gherghina and Tănase Dorobanțu, for the beautiful memories I have of them, and for the love and care that they surrounded me with when I was young.

Last, but not least, I would like to thank Lea, Aldo, Lara, Lana, and Ralph for their protection, cuddles, and the sheer joy they brought to my life.

Market Design, Borders, and Gravity in the Virtual World

Cosmina Liana Dorobanțu
Balliol College
University of Oxford

Thesis submitted for the degree of Doctor of Philosophy in Economics
Trinity Term 2015

Thesis Abstract

This thesis consists of three separate papers which examine different aspects of the economics of online commerce.

The first paper studies a natural experiment in the release of a new ad targeting feature onto an online advertising platform. The experiment affects the specificity of advertising assets in certain geographic ad markets. The paper finds evidence that the additional specificity negatively affects auction participation in the treated areas, an effect that has not been anticipated by the incumbent theoretical literature. The paper also finds evidence that despite negatively affecting auction participation, the additional specificity leads to higher revenue growth for the online platform in the treated areas. The paper's results highlight the importance of considering entry and exit decisions in theoretical models of specificity choices by market designers.

The second paper uses a proprietary data set from Google to find that online trade between two US states or two Canadian provinces is 6.7 times higher than trade between a US state and a Canadian province. This finding is surprising given that in the online environment, information costs and business-to-business transactions involving intermediate inputs are largely absent. When disaggregating the data by sectors of economic activity, the study finds that the largest US-Canada border effects occur for services whose

consumption is tied to a particular location and goods that face large regulatory hurdles at the border.

The third paper analyzes geographical patterns of cross-country Internet transactions using proprietary data from Google. The paper finds the effect of distance on online trade to be around -0.53. The study also finds that cultural characteristics, such as shared languages or religions, have a large impact on e-commerce, while economic ties, such as a common currency, have an insignificant effect. The paper underlines the importance of accounting for selection into trade in worldwide gravity estimations and identifies two exclusion restrictions that can be used when examining online trade flows.

Table of Contents

Acknowledgments **iii**

Thesis Abstract **vii**

Table of Contents **ix**

Chapter 1 – Competition and Specificity in Market Design: Evidence from Geotargeted Advertising **1**

1 Introduction 1
2 Relevant Literature 6
2.1 Auction and Market Design: Information Disclosure and Market Thickness 6
2.2 Internet Advertising 9
2.3 Internet Privacy 11
2.4 Asset Specificity 12
3 Setting and Institutional Details 12
3.1 Choices that Advertisers Make 12
3.1.1 Ad Content 12
3.1.2 Ad Targeting Criteria 13
3.1.3 Ad Costs 14
3.2 The Position of the Ads on the Page 15
4 The Natural Experiment 17
5 Data 19
5.1 Panel Index Variables 20
5.1.1 Location 20
5.1.2 Time 21
5.1.3 Industry 21

5.2 Outcome Variables 22
5.2.1 Auction Participation 22
5.2.2 Platform Revenue 24
5.2.3 Advertiser Outcomes 26
5.3 Explanatory Variables 27
5.3.1 Zip Code Targetability 27
5.3.2 Additional Controls 33
6 Specifications and Identification 34
6.1 Differences-in-Differences 35
6.2 Instrumental Variable Approach Combined with a Fuzzy Regression Discontinuity Design 36
7 Results 40
7.1 Auction Participation 40
7.2 Platform Revenue 42
7.3 Advertiser Outcomes 45
8 Conclusion 47
References 50
Appendix A: Industries 53
Appendix B: Summary Statistics and Detailed Data Sources for U.S. Census Figures 54

Chapter 2 – The US-Canada Border Effect in Online Commerce 57

1 Introduction 57
2 Relevant Literature 61
2.1 The Beginnings of a Puzzle: Measuring the Size of the US-Canada Border Effect 61
2.2 Intranational Border Effects 62
2.3 Explanations for the Existence of the Border Effect Offline 63
2.4 The Border Effect in the Online World 66
3 Empirical Approach 67
3.1 Empirical Specification 67
3.2 Estimation Method 69
4 Data 71
4.1 Google Data 71
4.1.1 Data Source 71
4.1.2 Type of Information Recorded 73
4.1.3 Coverage 75

4.1.4 Comparability to Border Effect Estimates Obtained Using Government Data Sources 76
4.1.5 Additional Google Data 77
4.2 Distance Measure 77
5 Aggregate Results 78
5.1 Main Specification 78
5.1.1 Discussion of Distance Effects 79
5.1.2 Discussion of Border Effects 81
5.1.3 Discussion of Contiguity Effects 81
5.2 Limiting the Data Set to Conversions Where No Product is Shipped 82
5.3 The Border Effect in Clicks versus Conversions 84
6 Sector Level Analysis 87
6.1 Results at the NAICS2 Level 87
6.2 Retail Trade Data Broken Down by NAICS3 Sectors 90
7 Conclusion 93
References 95
Appendix A: NAICS2: Full Set of Results 99
Appendix B: NAICS3 - Retail Trade: Full Set of Results 100

Chapter 3 – Worldwide Gravity in Online Commerce 101

1 Introduction 101
2 Relevant Literature 104
2.1 Gravity Equations 104
2.1.1 Distance 105
2.1.2 Other Geographical Factors 107
2.1.3 Cultural Factors 108
2.1.4 Economic Factors 108
2.2 Internet Trade and Gravity Equations 109
2.3 Overview of Helpman et al. (2008) 111
3 Data 114
3.1 Google Data 114
3.1.1 Data Source 114
3.1.2 Coverage 114
3.1.3 Comparability to Estimates Obtained by Helpman et al. (2008) 115
3.2 Explanatory Variables 116
3.2.1 Distance 116
3.2.2 Other Geographic Characteristics 118

3.2.3 Cultural Characteristics 119
3.2.4 Economic Characteristics 120
3.2.5 Exclusion Restriction 120
4 Empirical Approach 122
5 Results 124
5.1 Main Results 124
5.1.1 Distance 126
5.1.2 Other Geographical Characteristics 127
5.1.3 Language 127
5.1.4 Other Cultural Characteristics 128
5.1.5 Economic Characteristics 128
5.1.6 PPML Results 130
5.2 Alternative Exclusion Restriction 132
6 Conclusion 134
References 135
Appendix: Helpman et al. (2008)'s Results (Reproduced) 138

CHAPTER 1

Competition and Specificity in Market Design: Evidence from Geotargeted Advertising

Bo Cowgill and Cosmina Dorobanțu*

Abstract

This paper studies a natural experiment in the release of a new ad targeting feature onto an online advertising platform. The experiment affects the specificity of advertising assets in certain geographic ad markets. The paper finds evidence that the additional specificity negatively affects auction participation in the treated areas, an effect that has not been anticipated by the incumbent theoretical literature. The paper also finds evidence that despite negatively affecting auction participation, the additional specificity leads to higher revenue growth for the online platform in the treated areas. By showing that for the online platform in the study, the additional specificity was profitable despite concerns about market thinness, partitioned demand, and bargaining power, the paper's results uncover a new angle on the trade-off between competition and specificity. Furthermore, the paper's results highlight the importance of considering entry and exit decisions in theoretical models of specificity choices by market designers.

1 Introduction

Market designers often face choices about the specificity of tradable assets. For example, one design choice is to partition markets into separate, highly-specific asset classes. This design gives buyers a mechanism to pay higher amounts for goods or services that are well-targeted for their needs. However, this design also fragments competition between buyers into separate markets. Competition between buyers is useful to give bargaining power to the seller and place upward pressure on prices.

*Authors are listed in alphabetical order. The authors thank the online platform for sharing the data. We also thank Mark Armstrong, Alexandre de Cornière, Beata Javorcik, Ian Jewitt, Paul Klemperer, Jonathan Levin, Randall Lewis, Preston McAfee, John Morgan, Dennis Novy, Ferdinand Rauch, and seminar participants at Oxford for helpful comments. Dorobanțu acknowledges financial support from the Economic and Social Research Council.

By contrast, an alternative design pools separate asset types into a smaller number of more general, less specific markets. This design denies buyers the ability to pay more for better targeted assets. However, the lack of specificity enhances market participation. Is the tradeoff worthwhile? When is higher asset specificity desirable in a marketplace?[1]

This question is particularly salient in the Internet industry, where new technology firms have designed marketplaces to facilitate a variety of new transactions. As technology companies' capacity to collect data grows, designers at these firms can create increasingly specific markets where buyers can condition bids on lots of data.

In online advertising, ad platforms have a choice over how much ad targeting they enable. Highly specific targeting is valuable to buyers (advertisers). However, if advertisers can see all the available information about users, many of them might choose to not advertise to certain audiences at all, instead focusing on a smaller number of users in their target market. Because each firm has a slightly different target market, the result may be fragmented competition between advertisers and less pricing power for the ad platform.[2] By contrast, broader or more general targeting offers less value to the advertisers, but pools demand for the platform's ad inventory and creates price competition between advertisers.

A large literature in the economics of auction and market design has emphasized the importance of market thickness (Roth (2008)) and extra bidders (Bulow and Klemperer (1996)) for auction revenue. Despite these results, online advertising platforms continue to release more and more targeting features – which would seem to decrease the number of bidders per auction and reduce thickness. A related literature (Tadelis et al. (2015)) has stressed the value of information in auctions to improve match quality and increase the auctioneer's revenues. Through the lens of this literature, better targeting is a form of offering more information to buyers about the ad inventory for sale.

How does increased targeting affect an online platform's revenue? Researchers like

[1] A related question is – when is niche marketing desirable to mass marketing? For example, should a social networking site accept only users who meet certain characteristics or all users who are interested in joining?

[2] This issue is particularly relevant for auction-based ad platforms (Edleman et al. (2007) and Varian (2007)), where competition between advertisers plays a direct role in ad click prices and platform revenue.

Hummel and McAfee (2015) and Fu et al. (2012) provide a partial theoretical answer, providing conditions under which additional ad targeting can increase platform revenue. However, this nascent literature lacks empirical validation. In addition, these papers ignore the effect of better targeting on auction participation – a critical issue in the context of online advertising auctions.

We study this phenomena empirically using a natural experiment in the online ad space. In our setting, a large online advertising platform released a new feature permitting advertisers to target ever-smaller geographical areas within the United States. Prior to this release, only city-level targeting was available. After the release, the platform's advertisers were able to target individual ZIP codes.[3]

For privacy and accuracy reasons, the platform in our study did not enable targeting for all US ZIP codes. In particular, the online platform decided not to release finer geographical targeting in areas where its engineers were unable to detect 1,000 or more Internet cookies.[4] This policy was meant to ensure that the online behavior of individual users cannot be tracked by advertisers.

Given that some ZIP codes in our study were treated, while some remained untreated, we use a differences-in-differences approach to uncover the effect that the introduction of the new targeting feature had on the outcomes of interest. As a robustness check, we utilize the 1,000 Internet cookie cutoff rule for treatment in order to construct an instrumental variable approach combined with a fuzzy regression discontinuity design.

Our data set covers over 1 billion ad auctions and it enables us to examine how increased ad targeting affects auction participation and platform revenue. Furthermore, it enables us to study whether the platform's advertisers benefit from the release of the new targeting feature. We have four main findings.

First, we find that the introduction of finer geographical targeting had a statistically significant impact on auction participation. In particular, we find that in the treatment regions, the release of the finer targeting feature had a negative effect on the number of

[3] A system of postal codes used in the United States since 1963.

[4] An Internet cookie is a small text file sent from a website and stored on a user's hard drive.

ads displayed on the online platform.

Second, we find that although it had a negative impact on auction participation, the release of the new targeting feature had a positive effect on the online platform's revenue in the areas where treatment occurred. Our estimates suggest that a unit increase in treatment strength resulted in a revenue increase of 2.36 standard deviations relative to its underlying trend in the areas where the platform introduced ZIP code targeting.

Next, we next turn to the mechanism of this revenue increase. Revenue for the online platform in our study consists of two parts, the number of clicks accrued by the ads and the cost-per-click paid by the advertisers.[5] We find that the revenue growth was due to a higher volume of clicks; the average cost-per-click on the online platform did not increase relative to its underlying trend in the treatment areas following the introduction of ZIP code targeting.

Combined, what do these three findings suggest? In the treatment areas, there was a negative impact on the number of ads displayed on the online platform following the introduction of the new targeting feature, and a positive impact on the volume of clicks accrued by the ads. The release of the finer geographical targeting feature, in other words, had the important effect of increasing the relevancy of the ads running on the platform. From a revenue perspective, the online platform in our study benefitted from the introduction of ZIP code targeting because these more relevant ads generated more clicks.

Our fourth and final finding concerns advertiser outcomes. We find that enhanced targeting benefited not only the online platform, but also the advertisers that used the platform to promote their products. In particular, we find that in the treatment areas, the introduction of ZIP code targeting had a positive effect on the number of conversions.[6] Furthermore, we find that on average, the cost-per-conversion did not increase relative to its underlying trend in the treatment areas following the release of the new targeting feature. Combined, these two results show that ZIP code targeting was beneficial to

[5] Advertisers in this market only pay the platform when a user clicks on one of their ads.

[6] A conversion is an action that a user who clicks on an ad undertakes on the advertiser's website, such as purchasing, placing an item in a shopping cart, downloading or other forms of online interactions that are valuable to businesses.

advertisers: in the regions where treatment occurred, ZIP code targeting had a positive effect on the volume, but not the price, of the conversions.

Taken together, our findings uncover a new angle on the tradeoff between competition and specificity. We show that for the online platform in our study, enhanced ad targeting (ZIP code targeting) was profitable despite the concerns about market thinness, partitioned demand, and bargaining power.

We also show that the introduction of the new targeting feature had a statistically significant impact on auction participation. In the existing theoretical models of targeting and auction design, the number of participants in the marketplace is considered to be exogenous. As such, our findings suggest a new avenue for theoretical research.

From a policy making perspective, our results have implications for how regulators view the economics of privacy and data gathering. Collecting users' data in order to release finer targeting features is profitable for online platforms; it has a positive impact on the platforms' revenues and the relevancy of the ads running on the platforms. Enhanced targeting based on users' characteristics is also profitable for advertisers, whose advertisements generate a larger volume of sales.

Although our results show that enhanced targeting is beneficial for both online platforms and their advertisers, Internet users' welfare might be negatively affected by having their private browsing histories analyzed and used for commercial purposes. Regulation aimed at limiting the use of Internet users' data for ad targeting features should strike a balance between the negative effects the policy measures might have on the online advertising industry and the benefits Internet users will derive from having their privacy protected online.

The remainder of this paper proceeds as follows. Section 2 reviews the related literature. Section 3 outlines the setting and institutional details. Section 4 describes the natural experiment and Section 5 presents our data. Section 6 proposes the empirical specifications and Section 7 explores the results. We conclude discussing the implications of our findings and future avenues of research in Section 8.

2 Relevant Literature

This project aims to contribute to several parts of the literature about auction and market design, Internet advertising, Internet privacy, as well as asset specificity.

2.1 Auction and Market Design: Information Disclosure and Market Thickness

Our work relates to the effect of information disclosure on auction outcomes – a topic studied within the context of auction design. Milgrom and Weber (1982) are the first to address information revelation in auctions in a seminal work that identifies the "linkage" principle. Assuming that the bidders' valuations are affiliated (one bidder's high valuation makes other bidders' high valuations more likely), Milgrom and Weber (1982) show that in a second price auction, the auctioneer's revenue increases when she commits herself to a policy of full disclosure. The intuition behind this result is simple: with access to more information about the object auctioned, bidders' valuations are more closely aligned. This intensifies the competition between the bidders and encourages them to submit higher bids, which translate into increased revenues for the auctioneer.

Milgrom and Weber (1982)'s assumption that bidders' valuations are affiliated holds for certain auctions. For example, this assumption is plausible in the context of auctions for mineral rights, where each bidder's valuation of the auctioned object is a function of the amount of natural resources that can be extracted. Ganuza (2004), however, argues that there are other auctions, such as Internet auctions similar to the one we study, where bidders' valuations need not be correlated. In these settings, bidders' valuations can be heterogeneous, which means that in response to the auctioneer revealing additional information about the auctioned object, some bidders' valuations will increase, while other bidders' valuations will decrease.

In a setting where bidders have heterogeneous valuations, Ganuza (2004) identifies two competing effects. First, he shows that information revelation increases the efficiency of the auction by improving the match between bidders' preferences and the

attributes of the auctioned object. This has a positive effect on the winning bidder's willingness to pay for the auctioned object and on the auctioneer's revenue. Second, he shows that information revelation also increases the winning bidder's informational rents, which has a negative effect on the auctioneer's revenue. Ganuza (2004) proves that as the number of participating bidders increases, the first effect dominates the second. In other words, the auctioneer has an incentive to withhold information about the auctioned object when the number of bidders is low and to release information when the number of bidders is high.

Board (2009) generalizes Ganuza (2004)'s results and shows that in second price auctions – regardless of whether bidders' valuations are affiliated or heterogeneous – revealing information decreases the auctioneer's revenue when there are only two bidders and increases the auctioneer's revenue when there are four or more bidders. Board (2009) argues that this is due to the "allocation" effect, which causes the order of the bidders' valuations to change as a result of the information disclosed.

With two bidders, it is straightforward to see how Board (2009)'s "allocation" effect works. Assume that an online advertising platform does not reveal any information about the users visiting its website. Two advertisers, A and B, are competing in a second price auction for advertising space on the platform. Advertiser A values local users at $4 and non-local users at $2, while advertiser B values local users at $2 and non-local users at $4. Assuming that there is an equal probability of a user being local or non-local, each advertiser will bid $3 and the online platform's expected revenue from the auction will be $3. Now assume that the online platform will reveal to the advertisers the users' geographic location before the advertisers place their bids. If the user is local, advertiser A will bid $4 and advertiser B will bid $2. In a second price auction, the online platform's expected revenue will be $2, which is $1 less than the revenue the platform expects when it does not reveal information about its users.

In most auction models, information disclosure is considered to be exogenous. The works of Milgrom and Weber (1982), Ganuza (2004), and Board (2009) are revolutionary precisely because they relax this assumption and treat information revelation as an auctioneer's endogenous choice. However, despite their significant theoretical contribu-

tions, the works of Milgrom and Weber (1982), Ganuza (2004), and Board (2009) do have some limitations. In particular, they assume that the number of bidders participating in the auction is fixed. This assumption has the important implication that the information revealed by the auctioneer does not affect, in any way, the number of bidders participating in the auction. In other words, there is no auction entry or exit as a result of the information disclosed. Furthermore, they assume that the auctioneer has a single object for sale and conducts only one auction. This second assumption rules out the possibility that there are multiple objects for sale and that bidders can use the information revealed to decide what item (if any) to bid on.

In a recent empirical paper, Tadelis et al. (2015) call into question the plausibility of these two assumptions in the context of US wholesale car auctions. Using a large scale field experiment, they show that disclosing information about the quality of the vehicles auctioned raises the auctioneer's revenue, a result that is consistent with Milgrom and Weber (1982)'s predictions. However, Tadelis et al. (2015) also find that the increase in the auctioneer's revenue holds across all vehicle quality levels. In other words, the information revealed positively influences the auctioneer's revenue for both high and low quality cars. In the context of Milgrom and Weber (1982)'s theoretical model, this finding is surprising: bad news about the quality of an auctioned object should lower the bidders' willingness to pay.

Tadelis et al. (2015) cannot directly measure auction participation, but they hypothesize that their finding that information disclosure raises the seller's revenue regardless of whether the information is good news or bad news is due to a better match between the bidders and the auctioned object. In a setting with multiple objects for sale, bidders use the information disclosed by the auctioneer to choose which auction to participate in; they will enter auctions where they value the auctioned object highly relative to other bidders – and exit auctions where they do not. Tadelis et al. (2015) argue that this endogenous sorting of bidders into auctions as a result of information disclosure leads to a better match between the bidders and the auctioned items, which intensifies competition in each auction and results in higher bids and higher revenues for the seller.

Our work is similar to Tadelis et al. (2015) in that it underlines, in the context of mul-

tiple object auctions, the importance of allowing for an endogenous sorting of bidders into auctions as a result of information disclosure. We are able to measure auction participation in our data and to show that information disclosure has a statistically significant effect on auction participation. In particular, we find that information disclosure has a negative effect on auction participation. By directly measuring and identifying the effect of information revelation on auction participation, our work brings an important contribution to the literature studying the effect of information disclosure on auction outcomes.

Our work also relates to the literature studying the benefits of market thickness. In a seminal paper, Bulow and Klemperer (1996) emphasize the importance of market thickness by examining whether it is more profitable to sell a company via a public auction with no reserve price or an optimally structured negotiation with one less bidder. They find that the public auction is preferable, as long as it attracts at least one extra bidder compared to the negotiation process. A decade later, Roth (2008) describes the history of regional kidney exchanges in the United States in order to show that market thickness, defined as a large number of potential participants, is a key condition for marketplaces to function well.

In our paper, although we find that information disclosure has a negative effect on auction participation, we also find that the auctioneer's revenue increases as a result of information revelation. By showing that thinner markets need not always result in lower revenues for the auctioneer, our work also contributes to the literature studying the benefits of market thickness.

2.2 Internet Advertising

Issues surrounding information disclosure and thickness arise concretely in the case of Internet advertising markets. In these markets, online platforms like the one we study auction off advertising space. The platforms hold a considerable amount of information about their users, such as their geographical locations and search histories, and are often faced with the question of how much information they should disclose to potential ad

bidders. A small branch of the literature has sought to address this setting directly.[7]

In an informal essay, Levin and Milgrom (2010) discuss the tradeoffs that online platforms face when offering finer ad targeting. The authors point out that by disclosing information which allows bidders to target their ads to relevant users, online platforms ensure a good match between ad bidders and ad viewers. Levin and Milgrom (2010) hypothesize that the quality of the match encourages entry into the auction for advertising space, increases the competition between bidders, and results in higher revenues for the online platforms. However, the authors also note that revealing too much information about the users may create thin markets. If ad bidders can target very narrow sets of users, there will not be enough bidders participating in each auction. This can create monetization problems for the online platform.

Hummel and McAfee (2015) and Fu et al. (2012) propose theoretical models addressing the question of whether releasing enhanced targeting options in a second price online ad auction increases the platform's revenue. Using a single object auction model with a fixed number of symmetric bidders, Hummel and McAfee (2015) obtain results similar to Board (2009). They find that finer targeting options decrease the online platform's revenue when there is a relatively small number of players, increase the online platform's revenue when there is a relatively large number of players, and have an ambiguous effect for an intermediate number of players. Fu et al. (2012) also consider a single item auction with an exogenous number of bidders and find that enhanced targeting generally results in higher expected revenues for the online platform.

We contribute to the literature that studies whether releasing enhanced targeting options in a second price online ad auction increases platform revenues by being the first to address this question empirically. We also contribute to the literature by calling into

[7] Although it is not directly relevant to our work, we would like to point the reader to a related branch of the literature which addresses other interesting questions regarding the economics of targeting in online advertising. Bhawalkar et al. (2014) study how much advertisers bidding in an online auction for ad views might be willing to pay a data source for targeting data. Abraham et al. (2013) show that in the presence of information asymmetries (i.e. when some ad bidders have access to targeting data, while others do not), the revenue of an online platform employing a second price auction to allocate its advertising space is negatively affected. De Corniere and Nijs (2014) examine how the information revealed by an online platform about its users affects the equilibrium prices of the goods sold by the firms that place ads on the platform. Finally, Athey and Gans (2010) and Bergemann and Bonatti (2011) investigate how the offline market for advertising is affected by the appearance of online advertising, which allows advertisers to target their ads to users.

question the plausibility of some of the underlying assumptions used by Hummel and McAfee (2015) and Fu et al. (2012). In particular, we show that enhanced targeting options do affect auction participation – a finding that contradicts the theoretical models' assumption that better targeting options have no effect on the number of ads participating in the auction.

2.3 Internet Privacy

A key question addressed in this paper – the benefits of enhanced targeting for online platforms and advertisers – affects the economics of privacy. A widespread concern among privacy activists is that online platforms, as well as advertisers, face irresistible business incentives to invade consumers' privacy.

Little research has been done to date to help regulators understand whether they have to intervene to protect users' privacy online – and what impact their interventions might have on the online advertising industry. One exception is Johnson (2013), who builds a structural model to measure how three privacy-related regulatory options affect revenues for online platforms and advertiser surplus. He finds that a complete ban on ad targeting would decrease platforms' revenues and advertiser surplus by more than half.[8]

Our finding that enhanced targeting options are beneficial to online platforms and their advertisers contributes to the literature on two fronts. First, it confirms that online platforms and the advertisers promoting their products on these platforms do not have a financial incentive to protect Internet users' data. In doing so, it strengthens the case for policy interventions that protect online users' privacy. Second, it indicates that potential privacy-related measures may negatively impact the revenues of online advertising platforms and their advertising clients. In doing so, it encourages regulators to design policies that strike a balance between the negative effects on the online advertising industry and the benefits that Internet users derive from having their privacy protected online.

[8] A related literature shows that in the presence of targeting constraints, ads are less effective. See, for example, Goldfarb and Tucker (2011) and Bailey (2012).

2.4 Asset Specificity

Finally, this paper also contributes to the literature by attempting to integrate the notion of "asset specificity" (Williamson (1975)) into market design. The notion of asset specificity has been mostly used in the organization literature, where the specificity of assets plays a role in vertical integration and "boundaries of the firm" decisions. In this literature, the specificity of assets is exogenous. However in many emerging market design settings, the specificity of assets is a strategic, endogenous choice of a market designer. We provide a concrete example of this in the Internet advertising market, where online platforms choose how much specificity (targeting) to enable in the marketplace.

3 Setting and Institutional Details

Our empirical setting is a keyword search advertising platform. In this type of Internet advertising, users enter keywords into a search engine and are shown a labeled combination of organic results and paid advertisements. In this section, we introduce the main choices that advertisers face when placing an ad and we describe the ways in which the online platform uses the information provided by the advertisers to determine how it displays the ads.

3.1 Choices that Advertisers Make

Advertisers who want their ads to appear in the paid advertisements section of the online platform's search results page must set up an advertising account. When setting up an account, they face several choices regarding the content of the ad, the ad's targeting criteria, and the costs accrued by the ad. We discuss each one of these choices in turn.

3.1.1 Ad Content

Advertisers choose the text of their ad (limited to an ad title and two short lines of descriptive text) and the URL of the website that they wish to advertise. Suppose a mort-

gage broker based in New York City wants to advertise his website.[9] To do so, he writes the following ad:

Expert mortgage advice
Get impartial mortgage advice from
New York City's top rated broker!
`www.nyctopmortgagebroker.com`

Internet users that come across the ad will see the ad text above and if they choose to click on it, they will be taken to the mortgage broker's website.

3.1.2 Ad Targeting Criteria

Advertisers also choose the keywords (or search terms) that trigger their ads. By choosing the right keywords, advertisers ensure that their ads are only seen by users who are actively searching for the advertised product or service. Our mortgage broker, for example, might choose keywords such as 'business loans,' 'mortgage loans,' 'first home loans' or 'mortgage calculator' to trigger his ad. By choosing these keywords, he ensures that his ad will appear only to users who conduct an online search for one of these terms.

In addition to keywords, advertisers can specify other user characteristics as targeting criteria for their ads. For example, on some online advertising platforms, advertisers can opt for their ads to be visible only to female users, or only to college graduates. For the advertising platform in our study, geographic targeting is the only other additional way of targeting.

Note that many advertising platforms – including the one in our study – do not have direct information about the geographical location of their users. However, they are generally able to ascertain information about their users' location via the users' IP (Internet Protocol) address and other cues. Often an IP address can be mapped back to an Internet Service Provider (ISP) whose geographic coverage is known. The targeting software may also know information about how a user's request was routed through intermediary servers before arriving at the platform (including the name or other characteristics of

[9] The example given here is fictional. At the time of writing, the website `www.nyctopmortgagebroker.com` does not exist.

the wireless or ethernet network). This information may also yield some clues about the user's geographic location. Lastly, in some cases, platforms are able to utilize the GPS system on users' devices to establish their location.

Geographic targeting is attractive to advertisers, as it allows them to tailor their marketing strategy based on the users' location. Being based in New York City, our mortgage broker might want his ads to appear only to users based in New York City, who can easily reach his office for face-to-face consultations. If this is the case, he can choose to have his ads visible only within the New York City limits and thus avoid incurring advertising costs for users who live outside of his service area. Similarly, a large device manufacturer might value advertising more in certain areas than others. The latest tech gadgets are usually in high demand in areas like Silicon Valley, so the device manufacturer might want to spend more for ads seen by users within Silicon Valley and less for users in less technologically savvy locations. For both the mortgage broker and the device manufacturer, geographical targeting allows the advertisers to alter their marketing strategy depending on the geography of the users.[10]

3.1.3 Ad Costs

Advertisers in this setting pay per ad click. When choosing a particular keyword to trigger an ad, each advertiser enters a maximum cost-per-click (CPC). This represents the highest dollar amount that the advertiser is willing to pay for a click accrued by the ad. While there is no upper limit on the amount they can specify, advertisers are required to enter a maximum CPC that is at least one cent ($0.01).

The online platform in our study discloses to advertisers information regarding the average cost-per-click for each keyword they are considering, as well as information regarding the average number of monthly searches for that particular keyword. The av-

[10] Although our paper studies a natural experiment in geographic targeting, similar economics apply to other forms of ad targeting and other changes to asset specificity in markets. For example, suppose that a platform allowed advertisers to target their ads based on users' hobbies (in addition to keyword targeting). Revealing this knowledge may lead to many advertisers withdrawing bids from auctions where users' hobbies are irrelevant. However, the advertisers' valuations for users with relevant hobbies may go up. These dynamics (and others) mirror the economic mechanisms at work for the online platform in this paper as it introduces more geographical targeting options.

erage cost-per-click and the average search volume can vary considerably from keyword to keyword. For example, we set up an advertising campaign for the fictitious mortgage broker based in New York City using four keywords: 'business loans,' 'mortgage loans,' 'first home loans,' and 'mortgage calculator.' For an ad targeted to users within New York City, the online platform informed us that for the keyword 'business loans,' the average CPC is $53.69 and this keyword receives, on average, 1,000 monthly searches. For the keyword 'mortgage loans,' the average CPC is $27.56 and the average monthly search volume is 201. For 'first home loans,' the average CPC is $10.53 and the average monthly search volume is 10. Finally, for 'mortgage calculator,' the average CPC is $0.78 and the average monthly search volume is 90,500.

3.2 The Position of the Ads on the Page

The online platform uses the information provided by the advertisers when setting up their ad campaigns to determine the order in which it displays the ads on the page and the cost of each click accrued by the ads. We describe this process below.

When users conduct a search on the online platform, they reach a search results page, where there are shown a list of organic results and a labelled list of ads. The online platform displays a maximum of eight ads per search result page. The allocations of each advertiser to one of the eight ad positions on the page are determined through a generalized second price (GSP) position auction, which is widely used within the online advertising industry.

One important feature of this auction setting is that bids are quality-adjusted. The bids in the auction are the maximum cost-per-click amounts that the advertisers specify for each keyword. The quality adjustment comes in the form of a quality score that the online platform calculates for each keyword. The quality score is on a scale of 1 to 10, with 1 being the lowest and 10 the highest. The online platform calculates this score based on several factors, such as the relevance of the ad and the quality of the advertiser's website.[11] Advertisers do not know their keywords' exact quality scores,

[11] The platform does not reveal to the advertisers any information regarding the way it weighs each factor, nor does it publicize all the factors that it takes into account when calculating the quality score.

but they can see, for each keyword, whether their scores are 'low,' 'medium' or 'high.'

To measure an ad's relevance, the online platform has algorithms in place that analyze factors such as whether the language in the ad relates to the keyword chosen. For example, if the mortgage broker in our example uses the ad text in Section 3.1.1, but picks an irrelevant keyword, such as 'restaurants,' the online platform will consider the broker's ad to be irrelevant for this particular keyword. To measure the quality of an advertiser's website, the online platform uses algorithms that examine factors such as whether the page is easy to navigate, whether the content of the page is relevant to the ad, whether the website takes a long time to load, or whether it abuses users' data.

The following example illustrates how a generalized second price position auction works. Suppose there are four eligible advertisers in the auction for a given keyword. Advertiser 1 has a maximum CPC of \$4 and a quality score of 2, advertiser 2 has a maximum CPC of \$3 and a quality score of 5, advertiser 3 has a maximum CPC of \$2 and a quality score of 10, and advertiser 4 has a maximum CPC of \$1 and a quality score of 9. What will be their positions on the page and how much will each advertiser pay per click?

The total auction score of each advertiser is his bid times his quality score.[12] For advertiser 1, the total auction score is 8 (\$4 $\times$ 2), for advertiser 2, the total auction score is 15 (\$3 $\times$ 5), for advertiser 3, it is 20 (\$2 $\times$ 10), while for advertiser 4, it is 9 (\$1 $\times$ 9).

Advertiser 1 has the lowest auction score, so he will be the last one ranked on the page and he will pay \$4 per click. With an auction score of 9, advertiser 4 will rank right above advertiser 1, and he will pay just enough to obtain a total auction score above the auction score of advertiser 1. In other words, he will pay \$0.89 per click (\$0.89 $\times$ 9 = 8.01, which is greater than 8). With an auction score of 15, advertiser 2 will rank above advertiser 4, and she will pay \$1.81 per click (\$1.81 $\times$ 5 = 9.05, which is greater than 9). Finally, with an auction score of 20, advertiser 3 will take the top position on the page and she will pay \$1.51 per click (\$1.51 $\times$ 10 = 15.1, which is greater than 15).

[12] In this example, we weigh bids and quality scores equally, though this might not be the case. Because of its proprietary nature, we do not have access to information as to how the online platform weighs bids and quality scores.

Theoretical analyses of the auction format described above are available from Edleman et al. (2007) and Varian (2007). For our purposes, a key piece of information to retain is that within this auction format, advertisers can reduce their cost-per-click by improving the relevance of their ads.

4 The Natural Experiment

In March 2012, a technological breakthrough enabled the online platform in our study to introduce ZIP code targeting. This is a form of geographical ad targeting of small geographical units known as "ZIP codes." There are close to 30K standard ZIP codes in the United States.[13] The average standard US ZIP code covers a land area of 87.6 miles^2 (227 km^2) and is home to 10,600 people.

Prior to the introduction of ZIP code targeting, the most granular form of targeting available on the online platform we study was city targeting. The platform also offered other forms of wider geographical targeting: county targeting (there are 3,007 counties in the US), congressional district targeting (there are 435 congressional districts in the US, which cover the electoral constituencies that elect the members of the House of Representatives), designated market area (DMA) targeting (there are 210 DMA regions within the US, where the population has similar TV station, radio, newspaper, and often Internet content coverage), and state targeting (there are 50 states in the US).

In March 2012, the online platform in our study released ZIP code targeting to roughly one third of US' standard ZIP codes. For the remaining ZIP codes, geographical targeting remained unavailable. In September 2013, an additional 1% of US' standard ZIP codes became targetable thanks to technological improvements that the online platform made to the algorithm identifying users' geographical locations. The rest of the ZIP codes remain un-targetable.

[13]The US also has roughly 12.5K non-standard ZIP codes. These ZIP codes are assigned to overseas military bases, facilities that house post office boxes, and government agencies, companies or buildings that receive high volumes of mail, such as the CIA, Wal-Mart or the Empire State Building. Our analysis focuses only on the standard ZIP codes, as a household's online behavior within a standard ZIP code is likely to differ significantly from the online behavior of army personnel, post office employees or CIA operatives.

Through various panel data techniques described in Section 6, we use geographic areas with little ZIP code targeting enabled as "control groups" to estimate the effect of introducing finer geographical targeting. Areas with more targetable ZIP codes are the "treatment group."

An important econometric detail in our study is that the platform did not estimate which markets – neither geographic markets, industry markets, nor the combination – would yield the greatest profit from ZIP code targeting in order to introduce the feature only in those areas. Had the platform been strategic, it would have introduced ZIP code targeting selectively into markets where it would have the highest impact. For example, local services is a category of keywords that might be particularly strategic to introduce ZIP code targeting. Alternatively, the company might have introduced targeting into regions with high income, so that the ZIP code targeting could be used to direct advertising efforts to buyers with high purchasing power. However, rather than introducing ZIP code targeting into strategic industries and geographies, the platform offered the feature according to a threshold strategy – creating a regression discontinuity. In every geography where it offered ZIP targeting, all keywords were available to be targeted in all categories.

The platform's selection criteria into ZIP code targeting were a mixture of privacy and technological related reasons. On the privacy side, the platform decided not to release ZIP code targeting in areas where there was any risk of advertisers being able to track the online behavior of individual users. On the technological side, as discussed in Section 3, user locations must be inferred and cannot be perfectly known. The accuracy of geolocation tools is often below 100%. Companies in this industry typically establish internal quality requirements, and do not release geographical targeting features to advertisers in regions where the accuracy requirements are not met. The platform in our study chose to release ZIP code targeting only to ZIP codes that met the privacy and accuracy requirements set by the firm's engineers.

Notably for us, on the privacy side, the platform's engineers set the following cutoff rule: they did not launch ZIP code targeting in ZIP codes where they were unable to detect 1,000 or more Internet cookies. Below this threshold, no ZIP code got targeting.

Above this threshold, the ZIP codes that met the engineers' accuracy requirements got targeting.

In some of our empirical specifications, we use this privacy related cutoff rule to instrument for treatment status. We note that the existence of the 1,000 cookie threshold was not public knowledge, ruling out any possibilities for there being endogenous sorting of ZIP codes into treatment.

5 Data

The data from the online platform for this study were originally recorded on a per-query basis. For each search performed by a user, the platform recorded some information about the user and some information about the advertisers participating in the auction for ad space.

For the user, the online platform recorded the words entered into the search engine (often referred to as search terms or keywords in the advertising industry), as well as some characteristics of the search.[14] These characteristics include the estimated geographic location of the user, in as much granularity as the platform was able to infer, and a set of binary variables describing whether the user made the query from a computer, a tablet, or a mobile device.

For the advertisers who participated in the auction for the particular word combination searched by the user, the online platform recorded the following variables: the number of ad views (often referred to as impressions in the advertising industry), the number of ad clicks, and the price the advertisers paid for each ad click (as established by the "GSP" position auction). Furthermore, the online platform recorded in the underlying dataset for advertisers some data on purchases made by users after clicking on an ad.

[14] For privacy reasons, we were not shown the exact queries submitted by the users.

5.1 Panel Index Variables

For the econometrics in this study, the per-query data have been aggregated into sets of two-period geographic panels. The panels are indexed by location $\times$ industry $\times$ time to allow for a richer set of controls.

5.1.1 Location

We chose not to run our analysis at the ZIP code level. There are two reasons for this. The first is the quality of the data for the untreated ZIP codes. In order to obtain data at the ZIP code level, we must be able to accurately map IP addresses to each standard ZIP code in the US. However, one of the key reasons why some ZIP codes in our study did not get ZIP code targeting is the fact that the platform's engineers were unable to properly map IP addresses to these locations. Analysis at the ZIP code level would mean that for the untreated ZIP codes (unlike the treated ones), we might either misattribute observations or miss them out entirely due to the poor mapping.

The second reason deals with the stable unit treatment value assumption (SUTVA) violations between treated and untreated ZIP codes. This assumption requires treatment units and control units to not be able to affect each other. Had we used ZIP codes as a unit of analysis, there may have been interactions between treatment and control locations. For example, a "control" ZIP code that was adjacent to a "treated" ZIP code may be affected by the nearby treated unit (advertisers may substitute spending away from the control unit into the treated unit). The new targeting option may cause advertising revenue to shift out of un-targetable regions and into adjacent, targetable ones. If we were performing our analysis at the ZIP code level, this shift would appear to be "increased" revenue, even if overall revenue stayed constant and was simply distributed differently after the targeting. In order to avoid this, our basic unit of geography must be larger than a ZIP code.

The geographical unit we choose for data aggregation purposes is a "designated market area." As noted in Section 4, DMAs are geographies within the United States where

the population has similar TV station, radio, newspaper, and often Internet content coverage. Substitution across DMA boundaries is unlikely, so our unit of geographical aggregation satisfies SUTVA. Furthermore, the quality of the IP mapping at the DMA level is high for the entirety of the United States, eliminating concerns over quality differences between treated and untreated regions.

In no case is the geographic granularity of units in our panel specifications more fine (or less fine) than the DMA level. We refer to a DMA as having a higher level of "treatment" if its component ZIP codes can be targeted, and as relatively "untreated" if its component ZIP codes cannot be targeted.[15]

5.1.2 Time

On the time dimension, data are aggregated at the monthly level. For our analyses, we use data from February 2011, February 2012, and February 2013. The month in the middle, February 2012, was the month right before ZIP code targeting was launched. The first month, February 2011, was exactly a year before, while the last month, February 2013, is precisely a year later.

We chose a period of 12 months between the dates we study for two reasons. First, the effects of better targeting need time to evolve. Advertisers' choice to utilize ZIP code targeting is endogenous. It requires them to know about the existence of the new targeting feature and to shift their pre-existing strategy to utilize it. A year allows for this shift to take place. Second, online sales are seasonal. By studying the same calendar month, our estimations are not affected by seasonality movements.

5.1.3 Industry

We also aggregate the data by industry, in addition to location and time. "Industry" is a classification associated with a query and is assigned through semantic text analysis. These industries follow the same structure as NAICS sectors . The online platform

[15]Section 5.3 describes in detail how we construct the variable measuring treatment strength at the DMA level.

has 25 top level industry classifications, which are similar to 2-digit NAICS sectors. In our study, we use the top level classification to split our location $\times$ time panel into 25 industries. A full list of these industries is available in Appendix A.

5.2 Outcome Variables

We are interested in understanding how the introduction of the new targeting feature affected auction participation and the online platform's revenue. Furthermore, we are also interested in understanding whether the platform's advertisers benefited from the release of the new targeting feature. We will discuss, in turn, the outcome variables that we have at our disposal to uncover these effects.

5.2.1 Auction Participation

The first question we would like to address is whether the introduction of finer geographical targeting had an effect on auction participation. In light of the existing literature, answering this question is important for several reasons.

First, as we discussed in Section 2, there is a significant theoretical literature that addresses the more general question of information provision in auctions, as well as a smaller branch of the literature that tackles the more specific question of the effect of releasing enhanced targeting options in the context of second price online ad auctions. A key assumption, common to all existing theoretical models, is that auction participation is exogenous and cannot be affected by the information disclosed. By asking whether the introduction of ZIP code targeting has an effect on auction participation, we are the first to test the empirical validity of this assumption.

Second, a number of previous researchers have underlined the importance of market thickness and additional bidders for auction revenue. In light of this literature, the online platform's release of a finer targeting feature is risky: the introduction of ZIP code targeting may have the undesirable effect of reducing market thickness and creating monetization problems for the platform. By asking whether the release of the enhanced

targeting feature affects auction participation, we seek to understand whether the online platform in our study does, indeed, risk creating thinner markets.

The outcome variable that allows us to gain insight into auction participation is the number of ad impressions.[16] This represents the total number of ads that appeared on the online platform in response to searches conducted by users. If the introduction of ZIP code targeting resulted in more ads competing for the platform's advertising positions, we expect to find a positive effect of finer targeting on ad impressions. However, if the release of the new ZIP code targeting feature resulted in less ads competing for the platform's advertising positions, we expect to find a negative effect of enhanced targeting on ad impressions.

Note that a priori, it is ambiguous whether the number of ad impressions should be positively or negatively affected by the introduction of ZIP code targeting. Suppose a user living in San Francisco's Cole Valley neighborhood searches online for a pizza delivery service. Prior to the introduction of ZIP code targeting, three ads got triggered by the user's search for 'pizza delivery,' all of them targeting the entire city of San Francisco. One ad was for a pizza place that delivered to Cole Valley, while the other two were for pizza places that delivered to other neighborhoods in San Francisco. After the introduction of ZIP code targeting, however, only one ad gets triggered by the user's search for 'pizza delivery:' the one belonging to the pizza restaurant that services Cole Valley. The other two advertisers turn off their ads for neighborhoods where they cannot deliver. In this example, the number of impressions decreases from 3 to 1. The market, in other words, is becoming thinner.

[16] Our data set does not contain information about the number of advertiser accounts. However, even if we did have access to data regarding the number of accounts, in our setting, the number of ad impressions is a much more precise measure of auction participation than the number of advertiser accounts. As we have seen in Section 3, the online platform in our study auctions off advertising space. Each time a user performs a search, the platform runs an auction to determine the allocation of ads on the search results page. What matters in this setting is the number of participating ads, rather than the number of advertising accounts. A simple example will illustrate this point. Assume that a marketing company has an advertising account on the online platform in our study. Within this account, the marketing company manages the advertising campaigns of two different businesses, each selling mobile phone plans. When a user submits a search for 'mobile phone plans,' he sees two ads: both of them placed by the marketing company. If we measure auction participation by counting the number of ad impressions, we correctly measure it as two. However, if we measure auction participation by counting the number of participating advertiser accounts, we would incorrectly measure it as one, as there is only one account participating in the auction: the one belonging to the marketing company.

Now assume a different scenario: the same user living in San Francisco's Cole Valley neighborhood searches online for a pizza delivery service. Prior to the introduction of ZIP code targeting, the user's search triggered three ads: one for Domino's, one for Papa John's, and one for Pizza Hut. These are big restaurant chains, with large marketing budgets, that deliver to all of San Francisco's neighborhoods. A small pizza restaurant, that only services Cole Valley, did not have an ad presence on the online platform prior to the introduction of ZIP code targeting as it did not have the marketing budget to compete with the established players for clicks originating from anywhere in San Francisco. After the introduction of the ZIP code targeting feature, however, the small pizza restaurant in Cole Valley joins the ad platform and competes with the established players for clicks within its service area. As a result of the small restaurant entering the ad auction, the Cole Valley user searching for 'pizza delivery' sees four ads. In this example, the number of impressions increases from 3 to 4. The market, in other words, is becoming thicker.

Our two examples show that there might be some exit and some entry into the auction in response to the new targeting feature. In our data, a negative effect of ZIP code targeting on impressions would indicate that the exit effect dominates the entry effect.

5.2.2 Platform Revenue

The second question we would like to address is whether the online platform's revenue increased or decreased as a result of the ZIP code targeting feature. By answering this question, we hope to contribute to the existing literature on several fronts.

First, as noted in Section 2, uncovering the direction of the revenue effect will provide the first empirical answer to the question of whether enhanced targeting options increase an online platform's revenues in a second price auction. Thus far, this question has been addressed in the literature only from a theoretical perspective.

Second, by examining how the release of the new targeting feature affects auction participation and the platform's revenue, we also contribute to the literature studying the importance of market thickness for auction revenue. In particular, we ask how the auctioneer's revenue is affected when introducing a feature that risks creating thinner mar-

kets.

Finally, we contribute to the literature studying the economics of Internet privacy. In particular, by understanding the revenue effects of finer targeting options for online platforms, we shed light not only on the platform's financial incentives to release data about its users, but also on the impact on the profitability of online platforms of potential privacy-related measures.

The first outcome variable that allows us to gain insight into auction revenue is the online platform's total revenue from the advertising auction. Since advertisers pay the online platform only when users click on their ads, the platform's total revenue from its advertising auction is simply the sum over the dollar amounts paid by advertisers for each click accrued by their ads.

Two additional outcome variables allow us to decompose the revenue effect into its underlying components: the total number of clicks and the average realized cost-per-click (CPC). The total number of clicks represents the total number of times that users clicked on the ads displayed alongside the online platform's search results. If the number of clicks increases following the introduction of ZIP code targeting, this might be due to an increase in the number of ads running on the platform, to users finding the ads more relevant and clicking on them more, or a combination of the two.

If the average realized CPC decreases following the introduction of ZIP code targeting, this might signal one of three factors.[17] First, it might signal that advertisements on the online platform are becoming more relevant. As noted in Section 3.2, an important feature of the online platform's auction setting is that bids are quality-adjusted. If the quality of the ads increases following the introduction of the new targeting feature, the realized CPCs will be lower. Second, it might also signal that competition on the online platform diminished. In a generalized second price auction setting, thinner markets translate into lower realized CPCs. Finally, a decrease in the realized CPC following the introduction of ZIP code targeting might also signal that participating advertisers are, on average, submitting lower bids. This scenario is plausible if, for example, the new

[17] As a reminder to the reader, the realized CPC (the price the advertisers pay for each click) is sometimes, but not always, equal to the maximum CPC (the advertiser's bid). This is explained in Section 3.2.

ZIP code targeting feature attracts a large number of small, local businesses, that cannot afford to submit high bids.

5.2.3 Advertiser Outcomes

The final question we would like to address is whether finer targeting benefitted the advertisers that promoted their products on the online platform. Most of the existing literature focuses solely on the platform's perspective when studying the effects of enhanced targeting. The advertisers' perspective, however, is particularly relevant from a policy perspective. Do the advertisers have a financial incentive to obtain data about Internet users? If so, this would indicate that there is scope for policy makers to intervene in order to ensure that Internet users' privacy is protected. Also, do advertisers benefit or lose from potential regulation aimed at protecting Internet users' data? If they lose, this would indicate the need for policy makers to find a balance between the negative effects of the intervention and the benefits users derive from having their privacy protected online.

Two outcome variables allow us to study the advertisers' perspective. The first one is the total number of conversions. In order to provide its advertisers with useful intelligence about their ads, the online platform offers free "conversion tracking" software. This software tracks users anonymously from clicking on an ad to purchasing, placing the item in a shopping cart, downloading or other forms of online "conversions" that are valuable to businesses. If the release of the new ZIP code targeting feature has a positive effect on the number of conversions, this would be good news for the advertisers, as it would imply that a higher number of users ended up purchasing or showing interest in the items promoted by the online platform's advertisers.

The second outcome variable that allows us to study the advertisers' perspective is the cost-per-conversion. This is the average price that the online platform's advertisers pay for a conversion. We obtain this variable by dividing the total amount of money that the advertisers spend to promote their products on the online platform by the total number of conversions. If the introduction of ZIP code targeting has a negative effect on the

average cost-per-conversion, this would be good news for the advertisers, as it would imply that, on average, they are paying less in order to persuade the online platform's users to purchase their products.

Table 1 summarizes all the outcome variables introduced above.

Table 1: Outcome Variables

Variable	Description
Auction Participation	
Impressions	Total number of ads displayed on the platform in response to searches conducted by users
Platform Revenue	
Revenue	Online platform's total ad revenue
Clicks	Total number of clicks accrued by the ads running on the platform
Cost-per-click	Average price paid by the advertisers for a click
Advertiser Outcomes	
Conversions	Total number of times that users who clicked on ads completed purchases, downloads, or other forms of online interactions that are valuable to businesses
Cost-per-conversion	Average price paid by the advertisers for a conversion

5.3 Explanatory Variables

5.3.1 ZIP Code Targetability

ZIP code targetability is the variable whose change we are using for identification. This variable was zero during the pre-period for all designated market areas. For the post period, the value for each DMA lies between zero and one. This is because in many DMAs, only a faction of the component ZIP codes were made targetable. To construct the [0,1] variable, we measured what fraction of the DMA's queries were ZIP-targetable immediately after the launch. This represents the "level" of treatment for each DMA at the moment it was applied.[18]

[18] An alternative metric might have examined the fraction of targetable queries during the post-period, rather than immediately after launch. We did not use this metric for the following reason: suppose that ZIP

Figure 1 shows that treatment strength across DMAs is highly varied, alleviating any concerns that treatment strength at the DMA level is concentrated around a particular value.

Figure 1: Frequency of Treatment Strength at the DMA Level

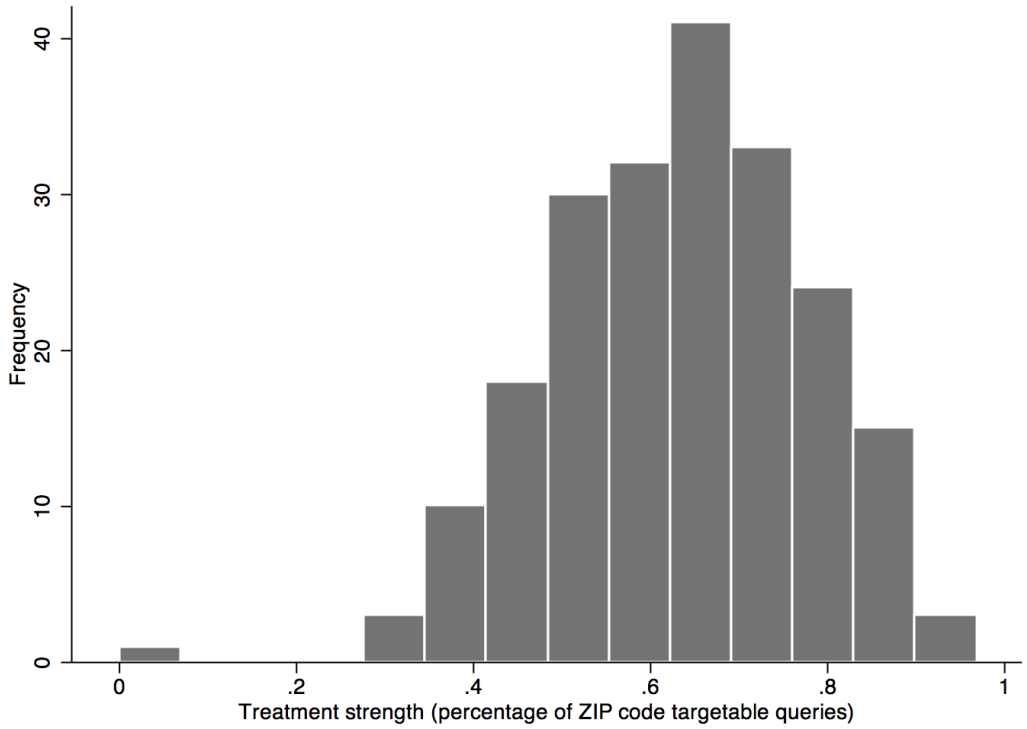
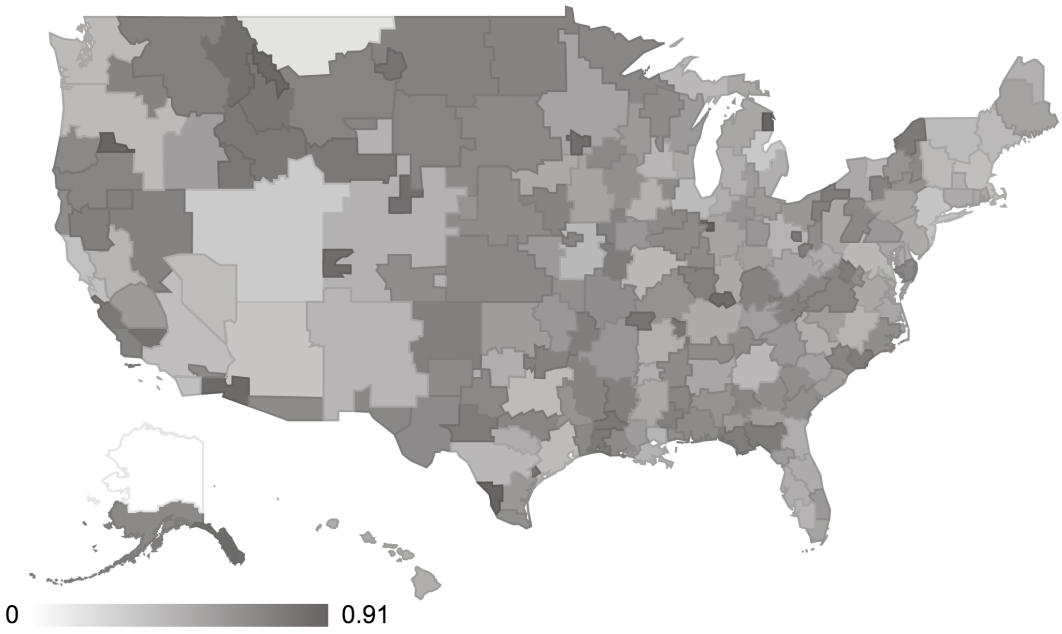


Figure 2 shows a map of all DMAs in the United States and their respective treatment strengths. ZIP code targetability, the variable we use to measure treatment strength, takes higher values for the darker ares on the map, and lower values for the lighter areas. Figure 2 is reassuring in that it shows that treatment strength is evenly distributed across the United States. For example, we do not see high concentrations of treatment predominantly along the East or West coasts or low concentrations of treatment predominantly in the middle states.

code targeting encouraged users to search more (or less) in treatment areas. If we encoded treatment status as the number of targeted queries in the post-period, then this would confuse the cause (lots of targetable queries) with the effects (people entering queries that can be ZIP targeted).

Figure 2: Geographical Distribution of Treatment Strength at the DMA Level



Besides examining the geographical distribution of treatment strength, we are also interested in knowing whether treatment strength is correlated with business presence or social and economic characteristics at the DMA level. In order to determine this, we use publicly available data sources to compile a database at the DMA level of several variables of interest, such as the number of commercial establishments per capita, median age, educational attainment, yearly income, and the unemployment rate.

Since designated market areas are not a commonly used geographical aggregation unit, none of the publicly available data sources within the United States report data at the DMA level. However, we circumvent this problem by collecting data at the ZIP code level and then using our ZIP code to DMA mapping to generate aggregate DMA figures. For the variables of interest in our study, ZIP code level data are available from the United States Census Bureau's County Business Patterns and American Community Survey. The data we collect are for 2011, the year prior to treatment.[19] We report summary statistics and detailed data sources for each variable of interest in Appendix B.

We examine whether treatment strength is correlated with business presence or social

[19]We extracted the data using American FactFinder, `http://factfinder2.census.gov`, on 26 December 2015.

and economic characteristics by regressing each variable we collect from the US Census Bureau on ZIP code targetability and a constant term. Tables 2, 3, and 4 report the results we obtain. In column (1) of each table, we present the results we obtain by estimating a simple linear regression model. In column (2) of each table, we present the results we obtain by estimating a median regression model.[20]

In Table 2, we investigate the relationship between treatment strength at the DMA level and business presence. In particular, we regress the number of commercial establishments per capita, a variable intended to capture the intensity of business activity at the DMA level, on ZIP code targetability, x_i. The regression results indicate that the coefficient on x_i is not significant at any conventional levels. This result is reassuring, as it confirms that treatment strength is not correlated with business presence.

Table 2: Treatment Strength and Business Presence

	Estimation method:	OLS	Quantile regression
Dependent variable:		(1)	(2)
Number of commercial establishments per capita	x_i	1.521	-3.999
		(2.510)	(2.419)
	constant	22.601***	25.108***
		(1.415)	(1.417)
	Observations	210	210
	R^2	0.003	0.011

The data for the dependent variable are for 2011. x_i is our treatment strength variable. It is bounded between $[0, 1]$ and it represents the percentage of "treated" queries that can be ZIP code targeted in DMA i at time of launch. Both the dependent and the independent variables are in levels. In column (2), we report quantile regression estimates for $q = 0.5$ (the median). We estimate the standard errors using the bootstrap method with 100 resamples.
*** Significant at 1%, ** significant at 5%, * significant at 10%

In Table 3, we examine the relationship between treatment strength and several social characteristics such as age, educational attainment or the prevalence of family households. We find that the coefficient of x_i, ZIP code targetability, is not significant at the 5% level for any of the variables we investigate. These results are equally reassuring, as they confirm that treatment strength is not correlated with social characteristics.

[20]For asymmetric distributions, a conditional quantile estimation can often pinpoint the central tendency of the distribution more accurately than the usual conditional mean estimation.

Table 3: Treatment Strength and Social Characteristics

	Estimation method:	OLS	Quantile regression
Dependent variable:		(1)	(2)
Median age	x_i	0.069	1.724
		(1.745)	(2.168)
	constant	37.857***	36.961***
		(1.079)	(1.354)
	Observations	210	210
	R^2	0.000	0.003
Educational attainment: high school or higher (% of people over 25 years)	x_i	-0.060*	-0.057
		(0.031)	(0.037)
	constant	0.890***	0.897***
		(0.016)	(0.022)
	Observations	210	210
	R^2	0.029	0.010
Family households (% of total households)	x_i	0.017	0.008
		(0.020)	(0.016)
	constant	0.656***	0.660***
		(0.012)	(0.009)
	Observations	210	210
	R^2	0.006	0.002

The data for the dependent variable are for 2011. x_i is our treatment strength variable. It is bounded between $[0, 1]$ and it represents the percentage of "treated" queries that can be ZIP code targeted in DMA i at time of launch. Both the dependent and the independent variables are in levels. In column (2), we report quantile regression estimates for $q = 0.5$ (the median). We estimate the standard errors using the bootstrap method with 100 resamples. *** Significant at 1%, ** significant at 5%, * significant at 10%

Finally, in Table 4, we investigate the relationship between treatment strength and income. The results we report in the first row of Table 4 indicate that there is a negative and statistically significant relationship between ZIP code targetability and yearly income per capita. The magnitude of the coefficient we estimate for x_i indicates that a 1% increase in ZIP code targetability is associated with a $112 decrease in yearly per capita income. This finding is surprising: if the online platform in our study was strategic in choosing the geographical areas for the release of the new ad targeting feature, we would expect the platform to choose higher income neighborhoods, with greater purchasing power, rather than lower income regions.

Table 4: Treatment Strength and Economic Characteristics

	Estimation method:	OLS	Quantile regression
Dependent variable:		(1)	(2)
Yearly income per capita (thousands 2011 USD)	x_i	-11.206*** (1.949)	-12.279*** (1.874)
	constant	31.695*** (1.287)	32.003*** (1.211)
	Observations	210	210
	R^2	0.180	0.124
Unemployed (% of people over 16 years)	x_i	-0.012 (0.011)	-0.012 (0.012)
	constant	0.092*** (0.007)	0.091*** (0.007)
	Observations	210	210
	R^2	0.007	0.005
Housing units with no vehicles available (% of occupied housing units)	x_i	-0.034* (0.018)	-0.011 (0.011)
	constant	0.092*** (0.011)	0.074*** (0.007)
	Observations	210	210
	R^2	0.052	0.005
Owner-occupied housing units (% of occupied housing units)	x_i	0.041 (0.027)	0.022 (0.032)
	constant	0.658*** (0.017)	0.676*** (0.018)
	Observations	210	210
	R^2	0.017	0.004
Housing units with less than one occupant per room (% of occupied housing units)	x_i	-0.001 (0.014)	0.004 (0.007)
	constant	0.975*** (0.008)	0.978*** (0.005)
	Observations	210	210
	R^2	0.000	0.001

The data for the dependent variable are for 2011. x_i is our treatment strength variable. It is bounded between $[0, 1]$ and it represents the percentage of "treated" queries that can be ZIP code targeted in DMA i at time of launch. Both the dependent and the independent variables are in levels. In column (2), we report quantile regression estimates for $q = 0.5$ (the median). We estimate the standard errors using the bootstrap method with 100 resamples.
*** Significant at 1%, ** significant at 5%, * significant at 10%

In the remaining rows of Table 4, we further probe this result. In particular, we are concerned that the yearly income per capita measure we are using might be measured with error. This is due to the fact that there is a significant amount of research indicating

that income figures in household survey data are measured with error.[21] For example, Pedace and Bates (2000), Bollinger (1998), and Hurst et al. (2014) convincingly show that survey respondents grossly misreport their income in household surveys.

If people misreport their income in household surveys, are there other variables that we can look at to assess their financial wellbeing? We obtain four such variables from the U.S. Census: the unemployment rate, the percentage of housing units that do not own any vehicle, the percentage of owner-occupied housing units (as opposed to rented housing units), and the percentage of housing units that have less than one occupant per room. We find that ZIP code targetability does not have a statistically significant effect on any of these variables. These results are encouraging, in that they suggest that the relationship between treatment strength and financial wellbeing might not be as strong as implied by the results we obtain using yearly income data.

Overall, the results presented in Tables 2, 3, and 4 alleviate concerns that treatment strength is correlated with business presence or the population's social and economic characteristics. The results also support the platform's position that the selection criteria into ZIP code targeting were a mixture of privacy and technological related reasons, rather than strategic motivations.

5.3.2 Additional Controls

In our regressions, we also include a few additional controls. First, we control for the device used by the people searching on the online platform: $percent_computer_{ijt}$ measures the percentage of queries originating from personal computers, while $percent_tablet_{ijt}$ measures the percentage of queries originating from tablets. The excluded category is $percent_mobile_{ijt}$, the percentage of queries originating from mobile phones.

Second, we control for the query mix on the platform by including the percentage of public queries, $percent_public_queries_{ijt}$. The online platform in our study considers a public query any search term that meets a global frequency threshold. This threshold requires the search term to have been queried by at least 20 unique users or 20 unique IP

[21] Bound et al. (2001) provide an excellent review of this literature.

addresses within a 90 day window. Once a search term passes the threshold, it remains labeled in the platform's databases as a public query permanently.

Finally, we include in our regression time, as well as (DMA $\times$ industry) fixed effects.

6 Specifications and Identification

We want to identify the effect of changes in the "treatment" status of DMAs on the various outcomes identified in Section 5.2. In other words, we are interested in estimating γ in the equation below:

$$\begin{aligned} \triangle y_{ijt} = \gamma x_{ijt} + \beta_1 percent_computer_{ijt} + \beta_2 percent_tablet_{ijt} \\ + \beta_3 percent_public_queries_{ijt} + \delta_t + \zeta_{ij} + \epsilon_{ijt} \end{aligned} \tag{1}$$

where t = February 2012 and February 2013, and:

$\triangle y_{ijt}$ = Change in the outcome variable at the (DMA i $\times$ industry j) level between periods t and $t-1$: a full list of outcome variables is available in Table 1.
x_{ijt} = Variable of interest: percentage of "treated" queries that can be ZIP code targeted in DMA i and industry j at time of launch.
$percent_computer_{ijt}$ = Control: percentage of queries originating from personal computers.
$percent_tablet_{ijt}$ = Control: percentage of queries originating from tablets.
$percent_public_queries_{ijt}$ = Control: percentage of public queries.
δ_t = Controls: period fixed effects. This is essentially a dummy variable taking the value of 0 prior to the launch, and 1 following the launch.
ζ_{ij} = Controls: (DMA $\times$ industry) fixed effects.
ϵ_{ijt} = Error term: standard errors clustered at DMA level.

On the left-hand side, we choose to look at changes in outcomes, rather than outcomes, for the variables of interest. In other words, instead of estimating our regressions using the data that we have for each outcome variable for February 2012 and February 2013, we estimate our regressions on the differences for each outcome variable between the months of (February 2012 - February 2011) and (February 2013 - February 2012). By

transforming our outcome variables in this way, we control for trends that are common to (DMA $\times$ industry) combinations. We also include in our regressions (DMA $\times$ industry) fixed effects, which allow us to ensure that our results are not affected by any underlying heterogeneity across (DMA $\times$ industry) pairs that is fixed over time.

We normalize the left-hand side variables in our study by subtracting the mean and dividing by the standard deviation.[22] We estimate Equation 1 by differences-in-differences, as well as by employing an instrumental variable (IV) approach combined with a fuzzy regression discontinuity (RD) design. We discuss each method below.

6.1 Differences-in-Differences

Since treatment was applied orthogonally to any of the outcome variables in our study, a difference-in-difference estimator is an appropriate way to estimate Equation 1. However, there is one additional step we take in order to ensure that this is, indeed, the case.

We calculate the growth rates between February 2011 and February 2012 (the year before the introduction of the ZIP code targeting feature), for revenue, clicks, impressions, conversions, and queries. We then estimate a regression with x_{ij} (the percentage of "treated" queries that can be ZIP code targeted in DMA i and industry j at the time of launch) as the dependent variable and the growth rates we obtain as independent variables.[23] If the platform launched ZIP code targeting only to high growth areas, we would expect the estimation to return positive and significant coefficients. However, if the platform launched ZIP code targeting orthogonally to the growth of the variables in our regression, we would expect the estimation to return insignificant results.

Table 5 summarizes the results we obtain. All coefficients are insignificant, with the exception of the coefficient of the variable measuring growth in the number of clicks. This coefficient, however, is only significant at the 5% level. In terms of magnitude, the estimate is very close to zero and it has a negative sign, which is the opposite of what

[22] The outcome variables in our study contain highly sensitive information. By normalizing the variables, we ensure that the confidentiality of the online platform's data is protected.

[23] Note that we do not include the cost-per-click and the cost-per-conversion in our regression as they are ratios of the variables whose growth rates we do include.

we would expect if the platform was being strategic in its selection of ZIP codes for the release of the new targeting feature. Given this, we believe that the results in Table 5, combined with the analysis presented in Section 5.3.1, provide convincing evidence for the appropriateness of a difference-in-differences estimator.

Table 5: Treatment Selection Analysis

Revenue Growth$_{ij}$	0.0008
	(0.0005)
Clicks Growth$_{ij}$	-0.0069**
	(0.0032)
Impressions Growth$_{ij}$	0.0042
	(0.0027)
Conversions Growth$_{ij}$	-0.0005
	(0.0004)
Query Growth$_{ij}$	0.0020
	(0.0018)
Industry fixed effects (ζ_j)	Yes
DMA fixed effects (α_i)	Yes
Observations	5,123
R^2	0.978

Dependent variable: treatment status (percentage of "treated" queries that can be ZIP code targeted in DMA i and industry j at time of launch). Each variable measures growth prior to treatment, from Feb 2011 to Feb 2012. Standard errors clustered at DMA level.
*** Significant at 1%, ** significant at 5%, * significant at 10%

6.2 Instrumental Variable Approach Combined with a Fuzzy Regression Discontinuity Design

Although we believe that our use of a differences-in-differences estimator is justified, as a robustness check, we ensure that γ, our parameter of interest in Equation 1, is properly identified by employing a instrumental variable (IV) approach combined with a fuzzy regression discontinuity (RD) design.

As noted in Section 4, due to privacy reasons, ZIP codes where the platform's engineers were not able to detect 1,000 cookies or more did not get ZIP code targeting. We use this cutoff rule to instrument for treatment status. In particular, we estimate the

following first stage regression:

$$x_{ijt} = \theta z_{ijt} + \beta_1 percent_computer_{ijt} + \beta_2 percent_tablet_{ijt} + \beta_3 percent_public_query_{ijt} + \beta_4 percent_1000+_queries_{ijt} + \delta_t + \zeta_{ij} + \nu_{ijt} \quad (2)$$

where:

x_{ijt} = Dependent variable: percentage of "treated" queries that can be ZIP code targeted in DMA i and industry j at time of launch.

z_{ijt} = Instrument: takes the value of 0 pre-treatment and equals the percentage of queries (targetable and un-targetable) in DMA i and industry j coming from ZIP codes with more than 1,000 cookies post-treatment.

$percent_computer_{ijt}$ = Control: percentage of queries originating from personal computers.

$percent_tablet_{ijt}$ = Control: percentage of queries originating from tablets.

$percent_public_queries_{ijt}$ = Control: percentage of public queries.

$percent_1000+_queries_{ijt}$ = Control: percentage of queries (targetable and un-targetable) in DMA i and industry j coming from ZIP codes with more than 1,000 cookies.[24]

δ_t = Controls: period fixed effects. This is essentially a dummy variable taking the value of 0 prior to the launch, and 1 following the launch.

ζ_{ij} = Controls: (DMA $\times$ industry) fixed effects.

ν_{ijt} = Error term: standard errors clustered at DMA level.

Our econometric approach is similar to that of Angrist et al. (1996). In a seminal work, the researchers use the cuttoff rule for assignment to treatment as an instrument for treatment status and show that in the second stage regression, the estimated coefficient of the treatment status variable measures the local average treatment effect. This in the casual effect of receiving treatment for the subset of the population that is assigned to treatment and receives the treatment. Angrist et al. (1996) work with binary treatments and instruments, but Angrist and Imbens (1995) extend their analysis to multi-valued treatments and instruments.

Our instrument, z_{ijt}, is strong by design. The platform's engineers released ZIP code targeting only to ZIP codes above the 1,000 threshold. This cutoff rule necessarily in-

[24]Since our instrument is the interaction of two variables, $percent_1000+_queries_{ijt}$ and δ_t, we must also include the uninteracted terms in both our first and second stage regressions.

duces a correlation between the number of queries from ZIP codes with cookie populations above 1,000 and the number of treated queries.[25] This ensures that θ in Equation 2 is different from zero.

Also by design, our instrument is uncorrelated with ϵ_{ijt}, the error term in Equation 1. The 1,000 cookie threshold was exogenously determined by the privacy related cutoff rule. It is orthogonal to the online platform's revenues and to other outcomes of interest to us. This ensures that any effect of the cutoff rule on y_{ijt}, the outcome of interest in Equation 1, comes solely through its effect on determining the treatment status, x_{ijt}.

We use $\widehat{x_{ijt}}$, the predicted values we obtain from the first stage regression, to replace the actual values of x_{ijt} in Equation 1. This allows us to estimate γ, the casual effect of ZIP code targetability on the outcomes of interest.

To construct our instrument, we obtain Internet cookie data collected by the online platform in our study.[26] Ideally, we would like to have access to the data that the platform's engineers used for the treatment cutoff rule. Unfortunately, those data are no longer available. The platform's privacy rules require it to delete all Internet cookie data older than two years. We are, however, able to obtain Internet cookie data from September 2, 2013. Although this is not an exact measure of the Internet cookie population at the time the treatment was applied, it is a good approximation. The United States is a mature market in terms of Internet usage. As such, the growth rates of variables such as Internet cookies, which are closely tied to the number of Internet users, do not experience large increases from year to year.[27]

In Figure 3, we show the probability of a ZIP code being treated as a function of the number of Internet cookies detected in that ZIP code. There is a clear discontinuity around 1,000 Internet cookies. To the left of this threshold, the probability of receiving treatment is below 0.15, while to the right of the threshold, the probability rises to 0.54.

Figure 3 also makes it clear that the discontinuity at 1,000 Internet cookies is fuzzy.

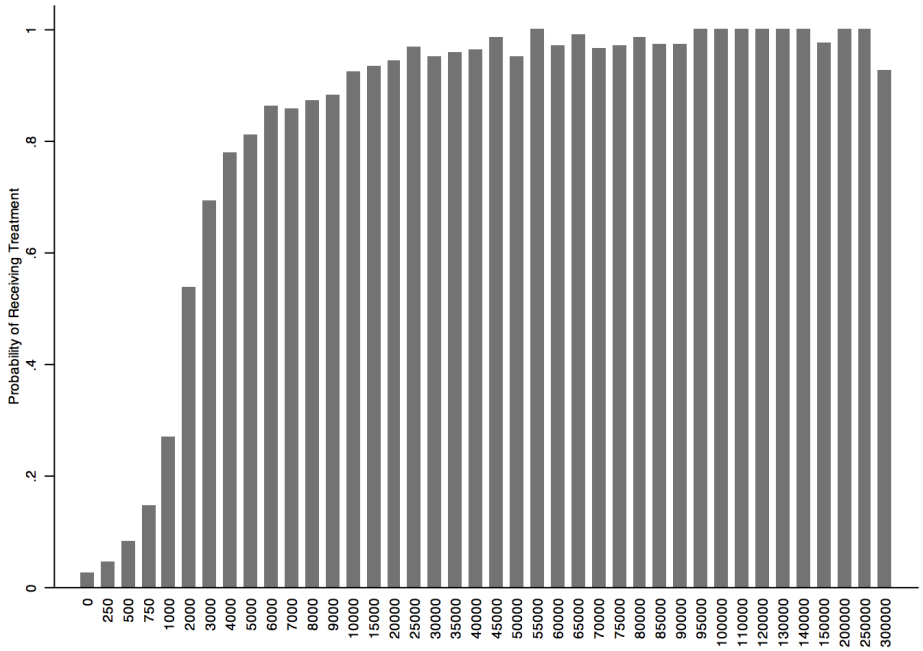
[25] Note that this is not a perfect correlation: not all ZIP codes above the 1,000 threshold received treatment. Only those ZIP codes that met the platform's accuracy requirements got ZIP code targeting.

[26] Internet cookie data are not available from public sources.

[27] We do experiment with various thresholds higher than 1,000 cookies in order to account for possible growth in the number of Internet cookies. Our results, however, remain unchanged.

The probability of receiving treatment below the 1,000 cookie threshold is not exactly zero due to our use of Internet cookie data from September 2013 to approximate the cookie population at the time of treatment. At the same time, the probability of receiving treatment above the 1,000 cookie threshold is not exactly one, either. This is because some ZIP codes that met the 1,000 cookie threshold remained un-targetable as they did not meet the platform's accuracy requirements.[28]

Figure 3: Probability of a ZIP Code Being Treated by Number of Internet Cookies



We estimate Equation 2, our first stage regression, and we obtain an F-statistic of 5,769.[29] This is well above the conventional required threshold. The estimated coefficient for z_{ijt}, our instrument, is 0.988, while the standard error is 0.003. With these results, we are satisfied that both our instrument and our first stage equation are very strong.

[28] 2,821 out of the total 11,539 ZIP codes that met the 1,000 Internet cookie threshold remained untreated.

[29] We report here the Kleibergen-Paap, rather than the Cragg-Donald, F-statistic as we are using heteroscedasticity robust standard errors.

7 Results

In this section, we present the results we obtain for each outcome variable and estimation method, discussing auction participation, platform's revenue, and advertiser outcomes following the introduction of the ZIP code targeting feature.

For each outcome variable, we report the coefficient we obtain for x_{ijt}, our variable of interest, by estimating the regressions using a differences-in-differences (DiD) approach and using an instrumental variable approach combined with a fuzzy regression discontinuity design (IVFRD). The DiD coefficients can be interpreted as the average effect on the dependent variable of a unit increase in treatment strength (in our case, ZIP code targetability) for the treated areas (in our case, the areas where the platform released ZIP code targeting). The IVFRD coefficients can be interpreted as the average effect on the outcome of interest of a unit increase in ZIP code targetability for areas whose treatment status was influenced by the instrument (in our case, the 1,000 Internet cookie cutoff rule).

7.1 Auction Participation

Table 6 details out the results we obtain when examining the change in the total number of ad impressions – the outcome variable that allows us to examine the effect of enhanced targeting on auction participation.

When estimated using the differences-in-differences approach, the coefficient we obtain for x_{ijt}, the variable measuring treatment strength, is negative and significant at the 5% level.[30] The magnitude of the estimated coefficient implies that a unit increase in treatment strength resulted, on average, in a decrease of 1.92 standard deviations in the number of ad impressions relative to their underlying trend within the areas where the online platform introduced ZIP code targeting.

When estimated using the instrumental variable approach combined with a fuzzy regression discontinuity design, the coefficient we obtain for x_{ijt} is negative and significant

[30] With a p-value of 0.017, the coefficient barely misses the significance threshold at the 1% level.

at the 1% level. The magnitude of the estimated coefficient implies that a unit increase in treatment strength resulted, on average, in a decrease of 1.82 standard deviations in the number of ad impressions relative to their underlying trend within the areas where treatment status was influenced by the instrument.

Table 6: Auction Participation Results

Dependent variable:	Change in the number of impressions (normalized)	
Estimation method:	DiD	IVFRD
	(1)	(2)
x_{ijt}	-1.916**	-1.821***
	(0.797)	(0.543)
Query origin controls ($percent_computer_{ijt}$, $percent_tablet_{ijt}$)	Yes	Yes
Query mix controls ($percent_public_query_{ijt}$)	Yes	Yes
Uninteracted IV control ($percent_1000+_queries_{ijt}$)	No	Yes
Period FEs (δ_t)	Yes	Yes
DMA $\times$ industry FEs (ζ_{ij})	Yes	Yes
Observations	10,500	10,500
R^2	0.419	

The dependent variable is the change in the total number of ad impressions between the months of (February 2013 - February 2012) and (February 2012 - February 2011). The independent variables are in levels, based on data from February 2013 and February 2012. ZIP code targeting was launched in March 2012. Standard errors are clustered at the DMA level.
*** Significant at 1%, ** significant at 5%, * significant at 10%

The results we present in Table 6 show that the introduction of ZIP code targeting had a negative and statistically significant effect on auction participation. With this finding, we contribute to the literature on several fronts.

First, similar to Tadelis et al. (2015), we call into question the theoretical assumption that information disclosure has no effect on auction participation. However, unlike Tadelis et al. (2015), we are able to measure auction participation in our data and to show, empirically, that in an auction with multiple objects for sale, the information disclosed by the auctioneer has a statistically significant effect on auction participation. By identi-

fying this effect in the data, we underline the importance of allowing for an endogenous sorting of bidders into auctions when modeling the effect of information disclosure on auction outcomes.

Second, we confirm Levin and Milgrom (2010)'s intuition that by offering finer targeting options, online platforms risk creating thinner markets. Indeed, in the setting that we study, the introduction of ZIP code targeting had a negative effect on auction participation. But – did the thinner markets have a negative effect on the online platform's revenue? This is the question we turn to next.

7.2 Platform Revenue

Table 7 details out the results we obtain for the outcome variables that allow us to shed light on the effect of enhanced targeting on the online platform's revenues. Columns (1) and (2) report the coefficients we obtain when examining the change in the platform's total revenue. Whether estimated using the differences-in-differences approach or the instrumental variable approach combined with a fuzzy regression discontinuity design, the coefficients we obtain for x_{ijt}, the variable measuring treatment strength, are positive and significant at the 1% level. This means that DMAs with higher treatment intensity saw greater revenue increases following the introduction of ZIP code targeting. In terms of magnitudes, the estimated coefficient for x_{ijt} is 2.36 standard deviations using the differences-in-differences approach and 2.60 standard deviations when using the instrumental variable approach combined with a fuzzy regression discontinuity design.

The revenue results we report in columns (1) and (2) of Table 7 are important for several reasons. First, combined with the results reported in Table 6, our findings suggest that although the introduction of ZIP code targeting had a negative effect on auction participation, the release of the finer targeting feature had a positive effect on the online platform's revenues. By showing that thinner markets need not always result in lower revenues for the auctioneer, our work contributes to the literature studying the benefits of market thickness in an auction context.

Second, our results are the first empirical confirmation of the theoretical predictions

advanced by papers like Hummel and McAfee (2015) and Fu et al. (2012), which ask, from a theoretical perspective, whether enhanced targeting increases an online platform's revenues in a second price auction.

Finally, our findings indicate that online platforms do not have a financial incentive to protect Internet users' privacy. In light of this, policy measures designed to limit the amount of targeting offered by online platforms may be needed in order to ensure that Internet users' data are protected.

Table 7: Platform Revenue Results

Dependent variable:	Change in platform's revenue (normalized)		Change in the number of clicks (normalized)		Change in the cost-per-click (normalized)	
Estimation method:	DiD	IVFRD	DiD	IVFRD	DiD	IVFRD
	(1)	(2)	(3)	(4)	(5)	(6)
x_{ijt}	2.362***	2.599***	3.557***	3.708***	-0.239	-0.442***
	(0.836)	(0.597)	(1.178)	(0.823)	(0.197)	(0.144)
Query origin controls (percent_computer$_{ijt}$, percent_tablet$_{ijt}$)	Yes	Yes	Yes	Yes	Yes	Yes
Query mix controls (percent_public_query$_{ijt}$)	Yes	Yes	Yes	Yes	Yes	Yes
Uninteracted IV control (percent_1000+_queries$_{ijt}$)	No	Yes	No	Yes	No	Yes
Period FEs (δ_t)	Yes	Yes	Yes	Yes	Yes	Yes
DMA $\times$ industry FEs (ζ_{ij})	Yes	Yes	Yes	Yes	Yes	Yes
Observations	10,500	10,500	10,500	10,500	10,500	10,500
R^2	0.513		0.330		0.566	

The dependent variables are the change in revenue (columns (1) and (2)), clicks (columns (3) and (4)), and cost-per-click (columns (5) and (6)) between the months of (February 2013 - February 2012) and (February 2012 - February 2011). The independent variables are in levels, based on data from February 2013 and February 2012. ZIP code targeting was launched in March 2012. Standard errors are clustered at the DMA level.
*** Significant at 1%, ** significant at 5%, * significant at 10%

In columns (3), (4), (5), and (6) of Table 7, we decompose the online platform's revenue into its two components: the number of clicks and the cost-per-click.[31] The coefficient on the variable measuring treatment strength is positive and significant at the 1% level in columns (3) and (4), where the outcome variable is the change in the number

[31] Because we normalize each dependent variable to protect the confidentiality of the online platform's data, the coefficients for the change in the number of clicks and the change in the cost-per-click cannot be added in the usual way to obtain the coefficients for total revenue.

of clicks. The magnitude of the coefficient estimated using the differences-in-differences approach implies that a unit increase in treatment strength resulted, on average, in an increase of 3.56 standard deviations in the number of ad clicks relative to their underlying trend within the areas where the online platform introduced ZIP code targeting. The instrumental variable approach combined with a fuzzy regression discontinuity design produces an estimate of a similar magnitude.

In Section 7.1, we saw that the introduction of ZIP code targeting had a negative effect on the number of ads running on the platform. The results reported in columns (3) and (4) of Table 7, however, show that despite its negative impact on the number of ads appearing on the platform, the release of the enhanced targeting feature had a positive effect on the number of clicks accrued by the ads. The ads, in other words, became more relevant.

The intuition behind this result is simple: with better access to information about the Internet users conducting searches on the online platform, advertisers were able to choose more precisely what auctions to participate in: they exited irrelevant ad auctions and participated only in the auctions where the users seeing the ads were likely to show interest in the advertised product or service. This sorting of advertisers into auctions following the release of ZIP code targeting resulted in a better match between the ads displayed on the platform and the users seeing them.

In columns (5) and (6) of Table 7 we report the results we obtain when our outcome variable is the change in the realized cost-per-click. Depending on the estimation method used, we find that the coefficient for x_{ijt} is either insignificant or negative.[32]

As noted in Section 5.2.2, in the context of the auction that we study, potential declines in the realized CPCs might arise due to several factors. First, the realized CPCs on the platform might decline as a result of ads becoming more relevant. Second, the realized CPCs might decline due to diminishing competition on the online platform. Finally, the

[32]Differences between the coefficients estimated using the two methodologies may arise due to the fact that the estimators recover different effects. While the differences-in-differences estimator recovers the average effect on the dependent variable of a unit increase in treatment strength for the treated areas, the instrumental variable approach combined with a fuzzy regression discontinuity design recovers the average effect on the dependent variable of a unit increase in treatment strength for the areas whose treatment status was influenced by the instrument. Imbens and Angrist (1994) discuss this in detail.

realized CPCs might also decline due to advertisers submitting lower bids.

In light of our previous results – namely that the introduction of ZIP code targeting increased the relevancy of the ads running on the online platform and decreased auction participation – the first two factors identified in the paragraph above are likely to be the key drivers behind a potential cost-per-click decline following the release of the finer targeting feature. Although we cannot test this empirically, we note that from a theoretical perspective, it is unlikely for advertisers to submit lower bids following the introduction of ZIP code targeting: most auction models predict that a better match between the auction participants and the auctioned items intensifies competition between the bidders and results in higher bids.[33]

While we are not able to identify the underlying dynamics behind the cost-per-click results reported in Table 7, one thing is for certain: in the treated areas, there were no increases in the average realized cost-per-click relative to its underlying trend following the introduction of ZIP code targeting. This result has an important implication: in the treated areas, the platform's revenues following the release of the finer targeting feature saw higher gains solely due to increases in the volume of the clicks accrued by ads – and not due to increases in the cost-per-click. Existing theoretical works focus only on the effect of enhanced targeting on auction prices, ignoring its effect on the volume of items sold. Our findings suggest that this is a potentially problematic oversight: for the online platform in our study, when it comes to revenue, the volume effect dominates.

7.3 Advertiser Outcomes

In Table 8, we report the results we obtain for the change in the total number of conversions and the change in the cost-per-conversion – the outcome variables that allow us to examine whether the introduction of the finer targeting feature benefitted the advertisers promoting their products on the online platform.

[33] The reason we cannot test this empirically is that we do not have access to maximum cost-per-click data. The differences that necessarily exist in a generalized second price position auction between the realized cost-per-click (the price that an advertiser pays for a click) and the maximum cost-per-click (the advertiser's actual bid in the auction) mean that we cannot use the results we report in columns (5) and (6) of Table 7 to ascertain whether the average bid on the platform increased or decreased following the introduction of ZIP code targeting.

Columns (1) and (2) report the coefficients we obtain when examining the change in the total number of conversions. Regardless of the estimation method used, the coefficients we obtain for x_{ijt}, the variable measuring treatment strength, are positive and significant at the 1% level. These results imply that DMAs with higher treatment intensity experienced greater growth in the number of conversions following the introduction of ZIP code targeting. The magnitude of the estimated coefficient using the difference-in-difference approach suggests that a unit increase in treatment strength led, on average, in an increase of 3.54 standard deviations in the number of conversions relative to their underlying trend within the ares where the platform introduced ZIP code targeting. The instrumental variable approach combined with a fuzzy regression discontinuity design produces an estimate of 3.63 standard deviations.

Table 8: Advertiser Outcomes

Dependent variable:	Change in conversions (normalized)		Change in the cost-per-conversion (normalized)	
Estimation method:	DiD	IVFRD	DiD	IVFRD
	(1)	(2)	(3)	(4)
x_{ijt}	3.543***	3.625***	-0.213	-0.419**
	(1.213)	(0.842)	(0.258)	(0.172)
Query origin controls (percent_computer$_{ijt}$, percent_tablet$_{ijt}$)	Yes	Yes	Yes	Yes
Query mix controls (percent_public_query$_{ijt}$)	Yes	Yes	Yes	Yes
Uninteracted IV control (percent_1000+_queries$_{ijt}$)	No	Yes	No	Yes
Period FEs (δ_t)	Yes	Yes	Yes	Yes
DMA $\times$ industry FEs (ζ_{ij})	Yes	Yes	Yes	Yes
Observations	10,500	10,500	10,500	10,500
R^2	0.221		0.726	

The dependent variables are the change in total conversions (columns (1) and (2)) and cost-per-conversion (columns (3) and (4)) between the months of (February 2013 - February 2012) and (February 2012 - February 2011). The independent variables are in levels, based on data from February 2013 and February 2012. ZIP code targeting was launched in March 2012. Standard errors are clustered at the DMA level.
*** Significant at 1%, ** significant at 5%, * significant at 10%

Columns (3) and (4) in Table 8 report the coefficients we obtain when examining the change in the cost-per-conversion. When employing the difference-in-difference ap-

proach, we obtain an insignificant coefficient, while when employing the instrumental variable approach, we obtain a negative coefficient that does not meet the significance threshold at the 1% level.

The results in Table 8 suggest that advertisers benefitted from the introduction of ZIP code targeting. The release of the finer targeting feature had a positive impact on the number of conversions. This suggests that following treatment, a higher number of users completed purchases, downloads, or other forms of valuable online interactions on the advertisers' websites. At the same time, the introduction of ZIP code targeting did not lead to increases in the average cost-per-conversion. Combined, these two results indicate that enhanced targeting was beneficial to advertisers: in regions with higher treatment strength, ZIP code targeting had a positive effect on the volume, but not on the price, of online conversions.

Our final finding, regarding the benefits of enhanced targeting to advertisers, contributes to the literature designed to help policy makers understand whether they should intervene to protect Internet users' privacy and how potential policy measures might affect advertisers. In particular, our results indicate that like online platforms, advertisers have business incentives to gather and use Internet users' data. This finding strengthens the case for policy interventions aimed at protecting online users' privacy. Furthermore, our results indicate that potential privacy-related measures might negatively impact the profitability of the advertisers who promote their products on online platforms. This finding encourages regulators to design policies that mitigate the negative effects of regulation on the online advertising industry.

8 Conclusion

Our results provide empirical evidence for the key finding of stylized models of ad targeting, namely that better targeting leads to increases in revenue. We show that for the online platform in our study, enhanced ad targeting was profitable despite concerns about market thinness and partitioned demand.

We also show that the introduction of ZIP code targeting had a statistically significant effect on auction participation. The theoretical literature on targeting and auction design ignores the effect of information revelation on auction participation. With this paper, we underline the importance of allowing for endogenous entry and exit into the advertising space when modeling the effects of enhanced targeting.

From a policy perspective, our findings suggest that both online platforms and their advertisers have business incentives to gather – and use – even more information about people searching online. For Internet users who value their privacy, the prospect of having more of their data collected and used for advertising purposes is undoubtedly worrisome. Regulation can limit the online industry's ability to make use of Internet users' data for commercial purposes, but policy makers will need to carefully design it so that it strikes a balance between its negative effects on the online platforms and their advertisers and its benefits to Internet users.

Our findings raise several questions for future research. Our paper aims to inform a more generic market design question about the tradeoff between specificity and competition between buyers. Little theoretical or empirical work deals directly with this tradeoff in market design. Although we find results in an empirical setting, this setting also has some limitations. We have estimates for one type of market (online ads) and one type of intervention (ZIP code targeting).

The natural experiment we study led to higher revenue growth in the treated areas. However, other changes to specificity in this market (and others) may not. At some point, thinning markets through targeting may deleteriously impact seller revenue. How can market design practitioners and regulators know the tipping point where a proposed targeting feature starts to harm revenue? Future research could inform this question through a combination of new theory and structural modeling.

In this paper, we focus on aggregate data and dynamics. Future researchers might be interested in understanding how the nature of improved specificity impacts outcomes. Some types of specificity may help sellers differentiate themselves horizontally. For example, in some cases, ZIP code targeting could help local businesses, such as restaurants,

advertise to local customers. Other types of specificity can also permit vertical differentiation. For example, ZIP code targeting could be useful for businesses who want to bid differently for wealthy neighborhoods vs poor neighborhoods. A specificity increase (such as ZIP code targeting) may have different impacts on revenue, auction participation, and advertiser outcomes depending on whether vertical or horizontal segmentation are in demand from advertisers. We hope to see these – and other – research questions explored as better data become available.

References

Abraham, Ittai, Susan Athey, Moshe Babaioff, and Michael Grubb, "Peaches, lemons, and cookies: Designing auction markets with dispersed information," in "Proceedings of the fourteenth ACM conference on Electronic commerce" ACM 2013, pp. 7–8.

Angrist, Joshua D and Guido W Imbens, "Two-stage least squares estimation of average causal effects in models with variable treatment intensity," *Journal of the American statistical Association*, 1995, *90* (430), 431–442.

_ , _ , and Donald B Rubin, "Identification of causal effects using instrumental variables," *Journal of the American statistical Association*, 1996, *91* (434), 444–455.

Athey, Susan and Joshua S Gans, "The impact of targeting technology on advertising markets and media competition," *American Economic Review*, 2010, *100* (2), 608–613.

Bailey, Michael Charles, "Essays in applied economics: Evidence from natural experiments." PhD dissertation, Stanford University 2012.

Bergemann, Dirk and Alessandro Bonatti, "Targeting in advertising markets: Implications for offline versus online media," *RAND Journal of Economics*, 2011, *42* (3), 417–443.

Bhawalkar, Kshipra, Patrick Hummel, and Sergei Vassilvitskii, "Value of targeting," in "Algorithmic Game Theory," Springer, 2014, pp. 194–205.

Board, Simon, "Revealing information in auctions: The allocation effect," *Economic Theory*, 2009, *38* (1), 125–135.

Bollinger, Christopher R, "Measurement error in the current population survey: A nonparametric look," *Journal of Labor Economics*, 1998, *16* (3), 576–594.

Bound, John, Charles Brown, and Nancy Mathiowetz, "Measurement error in survey data," *Handbook of Econometrics*, 2001, *5*, 3705–3843.

Bulow, Jeremy and Paul Klemperer, "Auctions versus negotiations," *American Economic Review*, 1996, *86* (1), 180–194.

Corniere, Alexandre De and Romain De Nijs, "Online advertising and privacy," *Available at SSRN 2191124*, 2014.

Edleman, Benjamin, Michael Ostrovsky, and Michael Schwarz, "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," *American Economic Review*, 2007, *97* (1), 242–259.

Fu, Hu, Patrick Jordan, Mohammad Mahdian, Uri Nadav, Inbal Talgam-Cohen, and Sergei Vassilvitskii, "Ad auctions with data," in "Algorithmic Game Theory," Springer, 2012, pp. 168–179.

Ganuza, Juan-José, "Ignorance promotes competition: An auction model with endogenous private valuations," *Rand Journal of Economics*, 2004, *35* (3), 583–598.

Goldfarb, Avi and Catherine E Tucker, "Privacy regulation and online advertising," *Management Science*, 2011, *57* (1), 57–71.

Hummel, Patrick and R Preston McAfee, "When does improved targeting increase revenue?," in "Proceedings of the 24th International Conference on World Wide Web" International World Wide Web Conferences Steering Committee 2015, pp. 462–472.

Hurst, Erik, Geng Li, and Benjamin Pugsley, "Are household surveys like tax forms? Evidence from income underreporting of the self-employed," *Review of Economics and Statistics*, 2014, *96* (1), 19–33.

Imbens, Guido W and Joshua D Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Johnson, Garrett, "The impact of privacy policy on the auction market for online display advertising," *Available at SSRN 2333193*, 2013.

Levin, Jonathan and Paul Milgrom, "Online advertising: Heterogeneity and conflation in market design," *American Economic Review: Papers and Proceedings*, 2010, *100* (2), 603–607.

Milgrom, Paul R and Robert J Weber, "A theory of auctions and competitive bidding," *Econometrica*, 1982, *50* (5), 1089–1122.

Pedace, Roberto and Nancy Bates, "Using administrative records to assess earnings reporting error in the survey of income and program participation," *Journal of Economic and Social Measurement*, 2000, *26* (3, 4), 173–192.

Roth, Alvin E, "What have we learned from market design?," *Economic Journal*, 2008, *118* (527), 285–310.

Tadelis, Steven, Florian Zettelmeyer et al., "Information disclosure as a matching mechanism: Theory and evidence from a field experiment," *American Economic Review*, 2015, *105* (2), 886–905.

Varian, Hal R, "Position auctions," *International Journal of Industrial Organization*, 2007, *25* (6), 1163–1178.

Williamson, Oliver E, "Markets and hierarchies," *New York*, 1975, pp. 26–30.

Appendix

A Industries

Adult

Apparel

Arts and Entertainment

Beauty and Personal Care

Business and Industrial

Computers and Consumer Electronics

Dining and Nightlife

Family and Community

Finance

Food and Groceries

Health

Hobbies and Leisure

Home and Garden

Internet and Telecom

Jobs and Education

Law and Government

News, Media, and Publications

Occasions and Gifts

Other

Real Estate

Retailers and General Merchandise

Sensitive Subjects

Sports and Fitness

Travel and Tourism

Vehicles

B Summary Statistics and Detailed Data Sources for U.S. Census Figures

Table 9: Summary Statistics at the DMA Level

Variable	Number of observations	Mean	Median	Standard deviation	Min	Max
Number of commercial establishments (per capita)	210	23.526	22.816	4.372	14.395	42.213
Median age (years)	210	37.890	38.000	3.048	28.700	48.800
Educational attainment: high school or higher (% of people over 25 years)	210	0.853	0.863	0.052	0.606	0.947
Family households (% of total households)	210	0.667	0.665	0.032	0.570	0.835
Yearly income (per capita, thousands of 2011 USD)	210	24.875	24.315	3.931	13.887	41.104
Unemployed (% of people over 16 years)	210	0.084	0.084	0.021	0.029	0.175
Housing units with no vehicles available (% of occupied housing units)	210	0.071	0.068	0.023	0.036	0.286
Owner-occupied housing units (% of occupied housing units)	210	0.684	0.689	0.048	0.542	0.825
Housing units with less than one occupant per room (% of occupied housing units)	210	0.975	0.981	0.020	0.838	0.999

Table 10: Detailed Data Sources

Variable	U.S. Census Bureau data publication	Table number
Number of commercial establishments	County Business Patterns	CB1100CZ11
Median age (years)	American Community Survey	S0101
Educational attainment: high school or higher (% of people over 25 years)	American Community Survey	DP02
Family households (% of total households)	American Community Survey	DP02
Yearly income (per capita, thousands of 2011 USD)	American Community Survey	DP03
Unemployed (% of people over 16 years)	American Community Survey	DP03
Housing units with no vehicles available (% of occupied housing units)	American Community Survey	DP04
Owner-occupied housing units (% of occupied housing units)	American Community Survey	DP04
Housing units with less than one occupant per room (% of occupied housing units)	American Community Survey	DP04

CHAPTER 2

The US-Canada Border Effect in Online Commerce

Cosmina Dorobanțu and Bo Cowgill*

Abstract

This is the first study to examine the US-Canada border effect in the virtual world. Using a proprietary data set from Google, the paper finds that the border between the United States and Canada has a large and statistically significant impact on online trade. At the aggregate level, the estimates for the border effect indicate that intranational trade on Google is 6.7 times higher than international trade. This finding is surprising given that in the online environment, information costs and business-to-business transactions involving intermediate inputs are largely absent. When disaggregating the data to the sector level, the study finds that large estimated US-Canada border effects for services whose consumption is tied to a particular location and goods that face large regulatory and bureaucratic hurdles at the border.

1 Introduction

Online commerce is a large and growing form of trade. 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 in 2012, 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 online trade are rare. This paper contributes

*The authors thank Google for sharing the data used in this paper. We also thank Avinash Dixit, Keith Head, Beata Javorcik, Bertin Martens, Marc Melitz, Dennis Novy, Ferdinand Rauch, Andrés Rodríguez-Clare, Tony Venables, Pierre-Louis Vézina, Noam Yuchtman, and seminar participants at Oxford, the European Trade Study Group, and the European Commission for helpful comments. Dorobanțu acknowledges financial support from the Economic and Social Research Council.

[1]The figures come from the U.S. Department of Commerce's *Quarterly Retail E-commerce Sales Report (Q1 2012)*.

[2]The figures come from Statistics Canada and eMarketer's 2012 *Canada Retail E-commerce Forecast*.

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 light on this well-known puzzle in international trade.

Our study comes at a time in which the future of national barriers is being publicly questioned by leaders in business, technology, and government. In a March 2012 speech in Germany, Google Chairman Eric Schmidt claimed before an audience including German Chancellor Angela Merkel and Brazilian President Dilma Rousseff that "the Web will dissolve national borders."[3] Similarly, in his 2005 bestselling book, *The World is Flat*, New York Times columnist Thomas Friedman heralded the arrival of a new, borderless economy made possible by advances in computing and telecommunications. Comparable ideas have been expressed by academics in the fields of business and international relations (Ōmae (1995)) and even by leaders of the US Central Intelligence Agency, National Intelligence Council, and advisors to the Director of National Intelligence.[4] Because of our use of Internet data to study the effect of national borders on Internet trade, our paper is particularly well-positioned to contribute to this debate.

The relationship between national borders and commerce is familiar to international trade economists. By estimating a simple gravity equation, McCallum (1995) showed that trade between two Canadian provinces is more than 20 times larger than trade between Canadian provinces and US states. Obstfeld and Rogoff (2001) included this result on their list of six major puzzles in international macroeconomics. Subsequent theoretical and empirical research shows that the border effect is slightly lower than that reported by McCallum (1995), but the estimated effect remains surprisingly high.

Over the years, researchers proposed several explanations to justify the high border effects estimated using traditional, offline trade data. These explanations point to the existence of trade barriers, information costs, multistage production, or data aggregation biases. The e-commerce transactions reflected in our data evade several of these explanations. Furthermore, the richness of the data allows us to investigate other untested

[3] "Schmidt: The Web will dissolve national barriers," `http://news.cnet.com/8301-30685_3-57390831-264/schmidt-the-web-will-dissolve-national-barriers/`. Accessed November 11, 2013.

[4] "The End of the Nation-State?", `http://www.nytimes.com/2013/10/13/opinion/sunday/the-end-of-the-nation-state.html`. Accessed November 17, 2013.

hypotheses, such as the role of consumer preferences.

Hortaçsu et al. (2009) argue that in the online environment, several of the traditional explanations outlined above do not apply. Information costs, for example, are much lower online. Unlike in traditional trade models, geographically distant buyers and sellers find each other easily and costlessly on the Internet. Exporters and importers do not need to incur high search costs or rely on the existence of migrant or business networks to discover and transact with each other. The cost-free online platform where they meet, be it eBay or Google, is the only network they need to find each other and to interact.

Multistage production is another unlikely candidate for justifying online border effects. Transactions conducted over the Internet are predominantly business-to-consumer (B2C) or consumer-to-consumer (C2C) exchanges in final goods. Business-to-business (B2B) transactions in intermediate goods hardly exist in the online world. A manufacturer, like Boeing, that needs to purchase intermediate inputs for its production process, like fuselage parts, will not buy them from an online store that it finds by searching on Google. Manufacturing companies usually have contracts with their suppliers that are negotiated and signed offline, and that cover large quantities of the intermediate input needed for production. The products and services that are sold online, by comparison, usually target final consumers and they have relatively low dollar values. By being an unlikely medium of exchange for intermediate inputs or B2B transactions, Hortaçsu et al. (2009) argue that the geographic agglomeration patterns chosen to optimize production processes in the offline world will not appear online.

If information costs or multistage production do not influence online trade patterns, we should expect to find a low border effect in the virtual world. However, in this paper, we show that the border between US and Canada has a large and negative impact on trade flows. Our estimates for the average border effect indicate that intranational online trade is 6.7 times higher than international trade.

What are the key drivers behind the online border effect? The first part of our analysis examines the possibility that one of the factors driving the online border effect is the difficulty of shipping goods across the US-Canada border. Our results, however, suggest

that the difficulty of shipping physical goods is not one of the drivers behind the online border effect.

The second part of our analysis examines the possibility that the US-Canada border effect is due, in part, to consumers' reluctance to transact with a foreign seller. We use a unique feature of our data, namely that a domestic consumer can transact with a foreign business only if the business makes a conscious decision to advertise its products aboard. If domestic consumers are not reluctant to trade with foreign sellers, once a foreign seller places an online ad in the consumer's country – and the consumer clicks on the ad – we would expect to see no difference in the data in the way in which ad clicks convert into purchases for consumers and sellers that are located in the same country versus the way in which ad clicks convert for consumers and sellers that are located on different sides of the US-Canada border. However, our results show the opposite: ad clicks have a lower probability to convert into sales when the transacting parties are not located within the same country. This indicates that consumers' reluctance to trade with foreign sellers might be one of the factors giving rise to the US-Canada border effect.

In the third part of our analysis, we classify our data by NAICS2 sectors and confirm previous researchers' finding that border effects vary greatly across sectors of economic activity. The US-Canada border affects the sector with the highest estimated border effect 7.4 times more than it affects the sector with the lowest estimated effect. This large difference points to the importance of using disaggregated data when studying border effects in a gravity setting.

Our results at the NAICS2 level show that in the virtual world, the US-Canada border effect is large for services that need to be consumed in a particular location. Our finding mirrors Hortaçsu et al. (2009)'s, who also find that within the United States, the home bias is highest for goods such as opera tickets, which are tied to a specific location.

We devote the fourth and final part of our analysis to examining Retail trade, a sector-level classification that encompasses cross-border transactions in final goods. We find that all else being equal, intranational Retail trade online is 6.2 times higher than international Retail trade. The magnitude of this effect clearly shows that even in the ab-

sence of B2B transactions in intermediate inputs, the border between the US and Canada severely hinders trade. Our results indicate that for certain products, like motor vehicles, regulatory and bureaucratic burdens faced by end consumers may contribute to the high US-Canada border effect that we observe in the data.

Section 2 reviews the literature on US-Canada border effects. Section 3 introduces the empirical approach and discusses the estimation method. Section 4 presents the data. Section 5 contains the results we obtain on aggregated data. Section 6 discusses the results we obtain when we disaggregate our data by sectors of economic activity and Section 7 concludes.

2 Relevant Literature

2.1 The Beginnings of a Puzzle: Measuring the Size of the US-Canada Border Effect

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 provinces or a province and a state 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 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.[5]

Anderson and van Wincoop (2003) is a major breakthrough. They derive the gravity equation from theory, basing their model on differentiated goods and constant elasticity of substitution preferences. The model shows that bilateral trade between two regions

[5] Although we do not cover it here, there is a different, but related strand of the literature which examines the US-Canada border effect through the lens of price differences on either side of the border. See, for example, Gorodnichenko and Tesar (2009), Gopinath et al. (2011), Boivin et al. (2012), and Broda and Weinstein (2008).

depends not only on the regions' GDPs and the distance between them, but also on their implicit price indexes. The authors show that 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 consistent 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 can include fixed effects for both the importing and the exporting regions. Using this methodology, Feenstra (2002) estimates an average border effect using traditional trade figures from 1993 of 4.7.[6] As 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.

2.2 Intranational Border Effects

Wolf (2000) showed that border effects affect intranational trade, as well. The researcher looked at shipments within the United States and found that crossing a state border diminished trade by almost as much as crossing an international border. Depending on the methodology he used, Wolf (2000)'s estimates show that trade within a US state is 3.3 to 4.4 times larger than trade with other American states.

[6]Researchers have found that similarly large international border effects exist for trade within the European Union. See, for example, Chen (2004) and Evans (2003).

Several researchers, starting with Hillberry and Hummels (2003), questioned the size of Wolf (2000)'s internal border estimates. However, the inescapable conclusion is that there is a home bias in trade that operates within, not just across, countries' borders. In a recent paper, Coughlin and Novy (2013) not only confirmed this conclusion, but also showed that within the United States, the intranational border effect is larger than the international border effect.

2.3 Explanations for the Existence of the Border Effect Offline

The empirical papers that uncovered sizable international, as well as intranational, border effects spurred a vast literature attempting to explain the existence of these border effects. We will summarize the most important explanations put forward below.

Trade barriers

One of the earliest explanations proposed for the existence of the international border effect relates to tariff and non-tariff barriers to trade, such as regulatory differences or standards requirements. However, the researchers that investigate this explanation (see, for example, Head and Mayer (2000), Chen (2004), Evans (2003)) find limited evidence to support it. Even the authors that find trade barriers to impact the border effect significantly underline the fact that the estimated impact is too small to explain the entirety of the border effect. In more recent work, Yi (2010) calculates that the trade weighted cost of shipping a good between Canada and the United States, relative to the average cost of shipping a good internally, is 14.8%.[7]

Information costs

Another explanation put forward for the existence of border effects relates to product specific information costs, namely the costs required to obtain information about the existence and quality of foreign products. Using Rauch (1999)'s classification, Chen (2004)

[7] In his trade cost calculation, Yi (2010) includes tariffs, non-tariff trade barriers, transport costs and wholesale distribution costs.

shows that the international border effect is larger in industries with a high degree of product differentiation, which face bigger information costs, compared to industries with homogenous goods. Around the same time, Evans (2003) also finds evidence in support of the information costs explanation for the existence of large international border effects.

In a related strand of the literature, several researchers find evidence that the existence of migrant and business networks reduce the intranational border effect. Networks encourage trade through two main channels: they reduce information costs and they contribute to the diffusion of preferences. Combes et al. (2005) study trade between French regions and find that combined, the existence of business and migrant networks reduce the intranational border effect by 53%. Millimet and Osang (2007) show that within the US, migrant networks dampen the intranational border effect, while Garmendia et al. (2012) reach similar conclusions by examining trade patterns within Spain.

Multistage production

Other researchers argue that the border effect arises due to goods being produced in multiple stages. Yi (2010) proposes a Ricardian model of trade similar to Eaton and Kortum (2002) to explain the international border effect between the United States and Canada. However, while Eaton and Kortum (2002) have single stage production, Yi (2010)'s model has multistage, multi-region production. In a world where different stages of the same good are produced in different regions , Yi (2010) shows that over two fifths of the border effect between US and Canada can be explained. By comparison, in a world with single stage production, only one sixth of the border effect can be explained. The differences between these two results is due to the fact that in a multistage model, a given trade cost gives rise to a larger border effect than in a standard model. In Yi (2010)'s model, goods end up crossing the border multiple times. As they do so, they incur the trade costs multiple times.

Rossi-Hansberg (2005) and Hillberry and Hummels (2008) also turn to multistage production processes in order to explain the border effect. However, unlike in Yi (2010)'s

model, the border effect in these two papers arises as a result of the spatial clustering of economic activity. Hillberry and Hummels (2008) show that shipments within the United States are highly localized. In particular, they find that shipments within a ZIP code are three times higher than shipments outside the ZIP code.[8]

Hillberry and Hummels (2008) turn to specialized intermediate inputs and clusters of up- and down-stream producers to explain their empirical finding. Suppose that region i produces an intermediate good x that can only be used as an input for the production of final good z. If region j does not produce z, it has no use for x, so trade between i and j will not occur. Furthermore, to avoid spatial frictions, the producers of specialized intermediate goods and the producers of the final goods that require them will locate next to each other. In other words, x and z producers will locate within the same region.

In Hillberry and Hummels (2008)'s world, intranational border effects are an artifact of geographic aggregation. The up- and down-stream producers' decision to locate next to each other means that shipments are very high over short distances, but they drop off significantly thereafter. On data sets that are aggregated at large geographical units, the authors argue, the intranational border effect captures up- and down- stream producers' preferences to trade with each other over very short distances.

Data aggregation

Finally, some researchers suggest that aggregating data over sectors of economic activity leads to biased estimates of the border effect. Hillberry (2002) shows that across industries, the US-Canada border effect varies from .03 for fur goods to 263.7 for gum and wood chemicals. The author argues that without disaggregating the data, researchers overestimate the border effect.

More recently, Anderson and Yotov (2010) focus on the manufacturing sectors and show that the aggregate US-Canada border effect is lower than the average border effect

[8]Hillberry and Hummels (2008) note that for the US data that they have, 88% of the intranational border effect comes from the number of shipments, rather than the average value of each shipment. In other words, the extensive, rather than the intensive, margin drives the border effect.

estimated with commodity level data.[9] Anderson et al. (2014) underline the need to estimate disaggregated gravity models for services, as well. They show that the coefficients on the US-Canada border effect dummies vary significantly across nine service sectors.

2.4 The Border Effect in the Online World

Only three previous papers report estimates for border effects using online trade data: Hortaçsu et al. (2009), Einav et al. (2014), and Lendle et al. (2015). Since border effects are not central to their research topics, two of these papers, Einav et al. (2014) and Lendle et al. (2015), simply provide estimates for the size of the border effect, with no further discussion. Only one of the papers, Hortaçsu et al. (2009), attempts to explain the factors which might give rise to the border effect in the online world. Notably, no other paper prior to ours studies the US-Canada border effect using online trade data.

Table 1 summarizes the intranational and international online border effects estimated by previous researchers.[10] Using eBay data, Hortaçsu et al. (2009) and Einav et al. (2014) find the same intranational border effect: all else being equal, trade within a US state is 1.75 times larger than trade outside the state. Hortaçsu et al. (2009) and Lendle et al. (2015) are the only researchers to estimate international border effects using online trade data. Their estimates, however, are widely divergent.

Table 1: Online Border Effects Obtained by Previous Researchers

Study	Intranat'l Border	Internat'l Border	Country Coverage	Estimator Used	Dependent Variable
Hortaçsu et al. (2009)	1.75	-	US	OLS	Number of sales (logs)
	2.75	431.82	9 South American states	OLS	Value of sales (logs)
Einav et al. (2014)	1.75	-	US	PPML	Number of sales
Lendle et al. (2015)		4.55	61 countries	OLS	Value of sales (logs)

[9] In related work, Chen and Novy (2011) document the fact that trade integration varies considerably across manufacturing industries within the European Union.

[10] In keeping with the rest of this literature review, the table reports the exponents of the estimated coefficients. The values reported in the table should be interpreted as the number of times that home trade exceeds out of state or out of country trade.

Hortaçsu et al. (2009) argue that in the online world, several of the explanations proposed by researchers studying offline trade flows do not apply. In particular, search costs are inexistent online: geographically distant buyers and sellers can easily find each other in the virtual world. Also, by being an unlikely medium of exchange for intermediate inputs or business-to-business transactions, Hortaçsu et al. (2009) argue that the physical proximity patterns chosen to optimize production processes identified in the offline world by Hillberry and Hummels (2008) will not appear online.

Hortaçsu et al. (2009)'s findings indicate that within the United States, the intranational border effect is not explained by the cost of shipping the products. The factors that drive the border effect are the existence of local products such as opera tickets, which must be consumed in a given location, and trust. On eBay, the estimated intranational border effect is higher for sellers that have low ratings than for sellers that have good ratings.

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} = Trade 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 trade 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,[11] we obtain the following operational econometric model:

$$\begin{aligned}\ln x_{ij} = {} & \beta_0 + \beta_1 \ln d_{ij} + \\ & + \beta_2\, contig\ US\text{-}CA_{ij} + \beta_3\, contig\ CA_{ij} + \beta_4\, contig\ US_{ij} + \beta_5\, internal_{ij} + \\ & + \beta_6\, border\ US \leftrightarrow CA_{ij} + \lambda_i + \chi_j + \varepsilon_{ij}\end{aligned} \tag{4}$$

where $contig\ US\text{-}CA_{ij}$, $contig\ CA_{ij}$, $contig\ US_{ij}$, $internal_{ij}$, and $border\ US \leftrightarrow CA_{ij}$ capture the observable intranational and international border barriers. We define the dummy variables as follows:

$contig\ US\text{-}CA_{ij}$ = 1 if a Canadian province and a US state share a land border, 0 otherwise.

$contig\ CA_{ij}$ = 1 if two Canadian provinces share a land border, 0 otherwise.

$contig\ US_{ij}$ = 1 if two US states share a land border, 0 otherwise.

$internal_{ij}$ = 1 for a Canadian province or a US state's trade with itself, 0 otherwise.

$border\ US \leftrightarrow CA_{ij}$ = 1 for exports between states and provinces and 0 otherwise.

[11] The fixed effects pick up other state- or province-specific unobservables. For example, Quebec is home to 6.3 out of the total of 7.3 million native francophones in Canada. Among all the states and provinces in this study, it is the only region where English is not an official language and where the population is predominantly French speaking. Being the only region with these linguistic characteristics, the fixed effects pick up this individual variation.

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

Note that in equation 4, $\beta_1 = (1-\sigma) \times \rho$. In other words, the coefficient in front of the bilateral distance variable is the product of the elasticity of substitution, $(1-\sigma)$, and the elasticity of total trade costs with respect to distance (ρ). Similarly, the coefficients in front of the observable intranational and international trade barriers variables (β_2, β_3, β_4, β_5, and β_6) are also the product of the elasticity of substitution, $(1-\sigma)$, and the elasticity of total trade costs with respect to each trade barrier variable. This is due to the fact that in equation 2, $\ln t_{ij}$ is preceded by $(1-\sigma)$.

Most studies assume that contiguity has a symmetric effect on trade. In our specification, however, we allow for asymmetric contiguity effects. 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 contiguity to affect trade asymmetrically.

3.2 Estimation Method

The prevalence of zero trade flows in the offline world gave rise to a substantial literature advocating against the use of linear estimators, which ignore the zeros in the data, in gravity exercises. Over the past decade, researchers such as Helpman et al. (2008) and Silva and Tenreyro (2006) proposed alternative estimation methods that account for the missing flows.

A particularity of our data set, however, is that we have very few missing trade flows. When working with aggregated data in the US-Canada online setting, we encounter no zero trade flows: all states and provinces export to all locations. When we disaggregate our data by sector of economic activity, we do encounter some zero trade flows, but even so, we find that in every industry we consider, at least 45 states and provinces export to all locations.

The relatively low incidence of zero trade flows in our data affects our choice of estimation methods. We follow Head and Mayer (2014)'s recommendation and choose an appropriate estimator for our US-Canada border effect analysis from the OLS, the Poisson Pseudo Maximum Likelihood (PPML),[12] and the Gamma Pseudo Maximum likelihood (GPML) estimators.[13] Even in the absence of zero trade flows, the PML estimators have a potential advantage over standard OLS: they do a better job at handling heteroskedasticity.

We perform the RESET test proposed by Silva and Tenreyro (2006) to ascertain whether the OLS or the PML estimators do a better job at addressing potential heteroskedasticity problems in our data.[14] We find that the PPML estimator has the best performance out of the three estimators considered, so we decide to use the PPML estimator for our analysis. For comparison purposes, for our main specification, we also provide the results we obtain by using OLS.

[12] Suggested by Silva and Tenreyro (2006).

[13] Note that we cannot employ a methodology similar to that suggested by Helpman et al. (2008), as it relies on a first stage regression that models the probability of two locations trading. With a data set where there are many zero trade flows, estimating a first stage Probit regression is a reasonable proposition. With our data set, however, estimating a first stage Probit regression is not feasible: in the absence of zero trade flows, the probability of exporting is always 1.

[14] The RESET test is a special case of White's test for heteroskedasticity. The test is conducted by examining the significance of an additional regressor, $x'\hat{\beta}$, where $\hat{\beta}$ is the vector of estimated parameters.

4 Data

4.1 Google Data

4.1.1 Data Source

The data we use 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, placing the item in a shopping cart, downloading or other forms of online "conversions" that are valuable to businesses. It allows Google's clients to measure the return on their investment in advertising in terms of dollars and transaction counts. The transaction data we use in this study come from the records generated by the conversion tracking software.

The following example illustrates how the conversion tracking software works. Suppose a bicycle producer wishes to advertise her bikes online, so she signs up for a Google AdWords account. Thanks to her new account, when users search on Google – or on millions of other websites that partner with Google – for bicycles, an ad for her business appears. Every time a user clicks on her ad, the bicycle producer pays Google for that click.

The question our hypothetical bicycle producer faces is whether the clicks she receives on her ad convert into sales for her. Google's conversion tracking software provides the answer to this question. This software consists of a piece of code that the bicycle producer places on her "thank you for your order" page – the part of her site that users see only after completing a purchase. Once placed, the code records for the advertiser the number of times the users who click on ads reach the "thank you for your order" page. In order words, the conversion tracking code records the number of sales our bicycle producer makes thanks to her ads.

Since the conversion tracking software tracks user behavior on millions of websites, including Google, we are confident that the data set we assemble from its records is the

most comprehensive one used to date to examine trade related questions in the digital world. However, like any other data set, the 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.

First, 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. Second, 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 recorded in the data set, as it happened outside of Google's advertising program. Third, the code generates a conversion count if a user clicks on an 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.

The final potential limitation comes from the way in which we determine location data. 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, the level of sophistication needed to set up a VPN network makes us believe that the users that mask their IP address represent a small enough percentage of the total Internet users to make our location identification on the buyers' side reliable. On the seller side, we have access to two address fields. The first one is a general mailing address that is self-reported by the sellers. This is the address field that we choose to use. The second one is the address where the credit card used to pay for the ads is registered. A data check reveals that

the general mailing address reported by the sellers very rarely differs from the billing address.

4.1.2 Type of Information Recorded

The United States and Canada are some of Google's most advanced markets, so we are able to extract from Google's databases additional information regarding the conversions recorded by the conversion tracking software. In particular, we are able to obtain information regarding the types of conversions recorded and we are also able to get insight into the sectors of economic activity to which the conversions in our data set correspond.

Conversion types

Online conversions can constitute a variety of advertiser-specified events. Some advertisers choose to measure conversions that do not immediately lead to money changing hands. For example, they may track mailing list signups or page views. Although they may not lead immediately to purchases, we believe all conversions tracked have economic value to the advertisers. If they did not, advertisers would not spend money to promote their sites on Google's platform in order to generate these conversions.

When setting up conversion tracking for their accounts, Google's advertisers can select the types of conversions they are tracking. Selecting a type is optional and the selection must fall into one of the categories listed in Table 2 below. We use the information contained in our database regarding conversion types for the analysis presented in Section 5.2.

Table 2: Conversion Types

Conversion type	Conversion is due to:
Purchase	A purchase, sale, or "order placed" event
Sign-up	A sign-up user action
Lead	A lead-generating action
Page view	A user visiting a page
Login	A user logging into a 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 referral of new customers

Sector of economic activity

Google has automatic mechanisms in place that classify the ads running on its online platforms into verticals. These verticals are similar to NAICS sectors and they follow the same structure. On Google, there are 27 top level vertical 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 Google 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's Chief Economist, Hal Varian, in Choi and Varian (2012). A few examples of how the mapping is done are provided in the Table 3 below.

Where Choi and Varian (2012) did not provide a NAICS/vertical mapping, we assigned an encoding. We manage to successfully assign a 2-digit NAICS classification to over 99% of the conversions in our data set. Only 0.2% of our data remain unclassified. Out of the classified conversions, 1.5% of them fall within the Agriculture, Mining, Construction, and Manufacturing sectors, leaving the vast majority of the classified con-

versions, over 98%, reflected within one of the service sectors.[15]

Table 3: Google Vertical to NAICS Sector Mapping

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
...	...	...	...

The fact that Manufacturing is so underrepresented in our data set warrants further explanation. Google does have a vertical for Industries, the company's equivalent of what in the economics world is commonly described as Manufacturing. Less than 0.2% of the conversions in our data set are classified under Industries. Google also has verticals such as 'Computers and electronics.' We can classify this vertical as NAICS 334, Computer and electronic product manufacturing, or as NAICS 443, Electronics and appliance stores. The first classification would fall under Manufacturing, the second one would fall under Services. We follow Choi and Varian (2012) and classify 'Computers and electronics' within Services, under Retail trade. The reason they suggest this classification and we choose to follow it, is because retailers, rather than manufacturers, are Google's usual advertising clients.

We note that over 30% of the conversions reflected in our data set fall within Retail trade. Due to the large number of conversions in this category, we are able to classify Retail trade conversions at the NAICS3 level.

4.1.3 Coverage

Positive conversion counts are reported for advertisers and sellers in all 50 states, the District of Columbia (DC), the 10 Canadian provinces, and the 3 Canadian territories. In line with previous offline studies, for our main specification, we exclude from our

[15]The only service sector that we cannot map any conversions to is "Management of companies and enterprises."

analysis the 3 Canadian territories located in the north of the country: Yukon, Nunavut, and the Northwest Territories. The territories are very scarcely populated, with just 0.30% of Canada's population residing in these areas. For the US, we focus our analysis on the 48 continental states and DC. The data set covers all conversion counts recorded over a period of four years, from 2008 to 2011. Similar to other researchers who examine online trade flows (see, for example, Lendle et al. (2015)), we average trade flows over this four year period.

4.1.4 Comparability to Border Effect Estimates Obtained Using Government Data Sources

Traditionally, studies that examine the US-Canada border effect (for example, McCallum (1995)) or intranational border effects (for example, Wolf (2000)) use government data. The estimates that these studies report are not directly comparable to ours. As noted in Section 4.1.2, over 98% of the conversions in our data set are reflected within one of the service sectors, with more than 30% of all conversions reflected in the two-digit NAICS services sector Retail trade.

Unfortunately, government data sources do not collect trade flow figures between US states and Canadian provinces for Services. For trade flows between US states and Canadian provinces, Statistics Canada and the US Census Bureau provide figures only for the two-digit NAICS sectors 11: Agriculture, forestry, fishing and hunting, 21: Mining, quarrying, and oil and gas extraction, 22: Utilities, 23: Construction, and 31-33: Manufacturing. No data are available for total trade in Services or for any of the two-digit NAICS service sectors.

Data for most service sectors are also unavailable at the US state to US state level. Despite the relevance of intrastate trade to the US Constitutional system,[16] the US government collects limited information on trade flows between US states. The Commodity Flow Survey (CFS), a joint effort of the US Census Bureau and the US Department of Transportation, is the primary source of data for trade flows between US states. The CFS

[16]Interstate commerce is mentioned in Article 1, Section 8 of the US Constitution, which has been the basis for many high-profile Supreme Court cases about the limits of government power to regulate the economy.

mostly covers the movement of commodities and business-to-business transactions such as mining, manufacturing, and wholesale distribution. The US Department of Transportation specifically notes that "most retail and services industries are excluded from the [Commodity Flow] Survey."[17]

Given the discrepancy between our data and traditional data sources' coverage of the service sectors, our estimates should be seen as complementing – rather than providing a direct comparison to – the estimates reported by researchers who examined the US-Canada border effect using traditional, offline data figures.

4.1.5 Additional Google Data

Besides conversion numbers, we are also able to obtain some data from Google regarding the number of ad clicks. The figures we obtain represent the number of times users on Google or Google's partner sites clicked on online ads placed by Canadian and American businesses. Positive click counts are reported for each US state and Canadian province. The click data cover a period of three years, from 2008 to 2010.[18] Click data are available only at the aggregate level; unlike conversion counts, we are unable to break down the figures by conversion type or sector of economic activity. Even so, we are able to use the aggregate click counts in Section 5.3 to understand whether consumers' reluctance to trade with foreign sellers may be one of the drivers behind the US-Canada border effect.

4.2 Distance Measure

Besides the trade flow figures obtained from Google, 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 $d_{ii} =$

[17] `http://www.rita.dot.gov/bts/sites/rita.dot.gov.bts/files/publications/commodity_flow_survey/html/methodology_2012.html`. Accessed on January 15, 2016.

[18] Conversion data cover a period of four years, from 2008 to 2011.

$0.67\sqrt{\text{internal area}_i/\pi}$. 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 Aggregate Results

5.1 Main Specification

Table 4 provides the results we obtain by using conversion count data aggregated across conversion types and sectors of economic activity. In column (1), we report the results we obtain when controlling only for distance, same state trade, and the US-Canada border. Column (2) is our preferred specification, as it also controls for contiguity effects. Column (3) alleviates concerns related to us excluding from our main analyses the non-continental US states and the 3 Northern Canadian territories. If we estimate our regressions with all 50 US states, DC, the 10 Canadian provinces, and the 3 Canadian territories, our results remain virtually unchanged. Finally, column (4) reports the estimates we obtain for our preferred specification by using OLS, rather than PPML.

All regressions have importer and exporter fixed effects. We include same state and same province conversions in all regressions in order to be able to estimate the home market effect. Including same state trade has an additional advantage. Hillberry and Hummels (2008) argue that trade is extremely localized and high national border effects might be due to researchers failing to control for this. In the absence of finer distance measures, as is the case here, one way to ensure that extremely localized trade does not drive the high national border effect is by including same state trade and same state dummy variables. For each estimated coefficient, we report robust standard errors, clustered by region pair.

Table 4: Results at the Aggregate Level for the Main Specification

	PPML			OLS
	48 US states, DC, 10 CA provinces		50 US states, DC, 10 CA provinces, 3 CA territories	48 US states, DC, 10 CA provinces
	(1)	(2)	(3)	(4)
$\ln d_{ij}$	-0.019	0.037	0.032	-0.152***
	(0.016)	(0.024)	(0.023)	(0.022)
contig US-CA$_{ij}$		0.223**	0.218**	0.169**
		(0.106)	(0.105)	(0.085)
contig CA$_{ij}$		0.028	0.021	-0.242
		(0.194)	(0.194)	(0.306)
contig US$_{ij}$		0.219***	0.212***	0.086**
		(0.042)	(0.041)	(0.037)
internal$_{ij}$	0.555***	0.706***	0.696***	1.261***
	(0.087)	(0.095)	(0.094)	(0.159)
border US $\leftrightarrow$ *CA*$_{ij}$	-1.912***	-1.903***	-1.903***	-1.858***
	(0.067)	(0.080)	(0.080)	(0.097)
Implied US$\leftrightarrow$CA Border Effect	6.76	6.71	6.71	6.41
Observations	3,481	3,481	4,096	3,481
RESET p-values	0.134	0.136	0.113	0.000

Importer and exporter fixed effects included. The dependent variable is the level of conversion counts aggregated across all conversion types and sectors of economic activity for the PPML specification and the log of conversion counts similarly aggregated for OLS. The data cover four years, 2008 to 2011. Robust standard errors (clustering by region pair). In columns (2) and (3), the coefficients for *border US* $\leftrightarrow$ *CA*$_{ij}$, and their standard errors, look identical due to rounding. In column (2), the estimated coefficient is -1.902716, with a standard error of 0.0800522. In column (3), the estimated coefficient is -1.903024, with a standard error of 0.0797963.
*** Significant at 1%, ** significant at 5%, * significant at 10%

5.1.1 Discussion of Distance Effects

There are several reasons why distance might matter less online. In particular, the information frictions that hinder trade in the offline world are mostly absent online. In the virtual world, search costs are close to zero: buyers and sellers find each other with a simple internet search, the cost of which is low and uncorrelated with the distance between the transacting parties. Communication costs online are also close to zero and independent of distance: buyers and sellers can easily communicate via email or web-based forms. Finally, contracting costs are also likely to be lower online, with transacting parties generally able to rate each other's behavior after a sale.

Previous researchers that examined online trade flows in a North American context found that, indeed, distance plays a smaller role online. Authors like Einav et al. (2014) and Hortaçsu et al. (2009) used data on the number of eBay transactions conducted within the United States to estimate distance coefficients of -0.10, and -0.07 respectively.[19] These coefficients are lower than those traditionally found in offline data, but they are significantly different from zero.

By contrast, our findings indicate that in the online world, distance is dead. The results reported for the PPML estimations in Table 4 indicate that distance does not have a statistically significant effect on the number of online conversions. The differences between our estimates and those reported by Einav et al. (2014) and Hortaçsu et al. (2009) might be due to the fact that our data sets cover different sets of transactions. The previous studies use data from eBay, which is a website dedicated to e-commerce transactions. Unlike eBay, Google is a generic, all-purpose search engine for a wide variety of online goods and services. Additionally, Google partners with millions of sites offering goods and services that cover a vast spectrum of economic activity online. The transactions reflected in our data set include diverse services such as legal advice, web hosting, banking, real estate services, and health care in addition to the more obvious categories of online commerce such as antiques, clothing, books, and electronics.

The differences we notice between our estimates and those reported by Einav et al. (2014) and Hortaçsu et al. (2009), therefore, might be due to the fact that the average service or good traded on Google and its partner sites is easier to provide at a distance than the average product sold on eBay. For example, while many eBay sellers must physically ship the product sold to the buyer, some service providers on Google or Google's partner sites might be able to transact electronically with their clients, without needing any physical shipments.

Note that in column (4), where we report the estimates we obtain by using the OLS, rather than the PPML estimator, the coefficient on distance is negative and statistically significant. The difference between the results we obtain by using PPML and the results

[19] Like us, Einav et al. (2014) use the PPML estimator. Hortaçsu et al. (2009), on the other hand, estimate their regressions using OLS. Both papers include in their estimations same-state trade.

we obtain by using OLS are likely due to heteroskedasticity. With a p-value of zero, the RESET test we perform for the OLS estimation clearly rejects the null hypothesis of homoskedastic errors. In the presence of heteroskedastic errors, Silva and Tenreyro (2006) conclusively show that a log linear model cannot be expected to produce consistent estimates of the parameters of interest. By contrast, the PPML specifications in Table 4 all pass the RESET test.

5.1.2 Discussion of Border Effects

The border between the United States and Canada continues to impact trade flows, even in the digital age. The coefficient of the US↔CA dummy variable in column (2) of Table 4 is large, negative, and statistically significant. Taking the exponent of this coefficient gives an average border effect of 6.7, which indicates that in a US-Canada setting, online trade between two US states or two Canadian provinces is 6.7 times higher than trade between a US state and a Canadian province. In the virtual world, where information costs and business-to-business transactions in intermediate inputs are largely absent, this is a surprising finding.

If distance has an insignificant effect on trade, the home bias influences online trade positively. The coefficient of the $internal_{ij}$ dummy variable in column (2) of Table 4 is economically and statistically significant. This result is in line with previous work. Einav et al. (2014) report a coefficient of 0.56 for their same state dummy. Hortaçsu et al. (2009) also report a coefficient of 0.56 when studying the number of eBay sales within the United States. At 0.71, our estimated coefficient is slightly higher. It implies that all else being equal, trade within a state or a province is 2.03 times larger than intranational trade.

5.1.3 Discussion of Contiguity Effects

Contiguity does not appear to have a large influence on trade in the virtual world. The coefficients of the dummy variables measuring the effect of sharing a land border between a US state and a Canadian province, as well as between two US states, are posi-

tive, but their magnitudes are not very large. They imply that all else being equal, trade between two contiguous US states or a contiguous US state and Canadian province is approximately 25% higher than trade between non-contiguous locations. The coefficient of the dummy variable measuring the effect of sharing a land border between two Canadian provinces is not significant at any conventional level. Our results are similar to the ones reported by Anderson and Yotov (2010), who find that contiguity has a positive and significant effect on trade between US states and Canadian provinces. Like us, they also find that sharing a land border makes no difference for trade between Canadian provinces.

5.2 Limiting the Data Set to Conversions Where No Product Is Shipped

Section 5.1 makes is clear that a large and significant US-Canada border effect exists in the online world. In this section, we use the data we have on conversion types to gather the first insight into the factors that give rise to this border effect. Yi (2010) calculates that the trade weighted cost of shipping a good between Canada and the United States, relative to shipping a good internally, is 14.5%. While we cannot directly calculate the cost of shipping goods for our data, we can use the data we have on conversion types in order to understand whether the difficulty of shipping goods across the US-Canada border is the factor that drives the border effect.

As we noted in Section 4.1.2, Google records 10 different types of conversions. Out of these 10 types, only 3 of them describe a direct sale, where the product being sold might change hands, physically, between the buyer and the seller. These 3 types are *Purchase* (a purchase, sale or "order placed" event), *Order charged* (a purchase or order that was successfully charged for), and *Shopping cart post* (an item put into a shopping cart). The other 7 conversion types (*Page view*, *Install*, *Download*, *Login*, *Signup*, *Referral*, *Lead*) involve an online event only, where no physical good is being shipped from the buyer to the seller.

In column (2) of Table 5, we report the results we obtain by limiting our data set to the 7 conversion types that refer to online events only. In column (1) of Table 5, we replicate

from Table 4 the results we obtain when using data aggregated over all 10 conversion types. If the difficulty of shipping a good across the US-Canada border is one of the factors which drive the international border effect in our estimations, we expect to find a smaller coefficient on the US↔CA border dummy for the subset of our data that excludes the 3 types of conversion which might describe goods being physically shipped.

We find that for the subset of the data that describes online events only, the US-Canada border effect appears to be even higher than the border effect we find when we consider all conversion types. The coefficient for the US↔CA border dummy in column (2) of Table 5 is 2.071, larger than the one we report in column (1) of the same table for all conversion types. However, a t-test reveals that the difference between these two coefficients is not statistically significant, so we cannot reject the null hypothesis that they are the same.

Table 5: Results for Conversions Where No Product is Shipped

	All conversions	Conversions where no product is shipped
	(1)	(2)
$\ln d_{ij}$	0.037	0.033
	(0.024)	(0.026)
contig US-CA$_{ij}$	0.223**	0.173**
	(0.106)	(0.085)
contig CA$_{ij}$	0.028	-0.009
	(0.194)	(0.158)
contig US$_{ij}$	0.219***	0.229***
	(0.042)	(0.046)
internal$_{ij}$	0.706***	0.717***
	(0.095)	(0.107)
border US $\leftrightarrow$ *CA*$_{ij}$	-1.903***	-2.071***
	(0.080)	(0.063)
Implied US↔CA Border Effect	6.71	7.93
Observations	3,481	3,481

Importer and exporter fixed effects included. The dependent variable is the levels of conversion counts aggregated over all sectors of economic activity and over all conversion types (column 1) and aggregated over all sectors of economic activity and the 7 conversion types that refer purely to online events (column 2). The data cover four years, 2008 to 2011. Estimations use PPML. Robust standard errors (clustering by region pair).
*** Significant at 1%, ** significant at 5%, * significant at 10%

The comparison between the two estimated US-Canada border effects in Table 5 is telling: when no product is being physically shipped, we find that the border effect is just as large as when we examine all conversion types. This result indicates that in the online world, the difficulty of shipping goods across the US-Canada border does not explain the large international border effect that we find in the data. Our finding is similar to Hortaçsu et al. (2009), whose results show that within the United States, shipping costs cannot explain the border effect.

5.3 The Border Effect in Clicks versus Conversions

In this section, we use the data that we have on ad clicks to try to understand whether consumers' reluctance to trade with foreign sellers is one of the factors behind the US-Canada border effect. In column (1) of Table 6, we run our preferred specification with the number of clicks as the dependent variable.

A priori, we expect to find the US-Canada border to have a dampening effect on ad clicks. For a cross-border click to have happened, a domestic business must have made a conscious decision to target the foreign consumer's geographical area and to pay for ads in that area.[20] A profit-maximizing enterprise would make this decision only after overcoming all the financial, logistic, and regulatory hurdles associated with exporting and ensuring that it can – and wants to – conduct business abroad. In this sense, we can think of the US-Canada border effect for clicks as capturing the difficulties of a business expanding abroad.

Furthermore, for a click to have happened, the foreign consumer must have also been interested in – and actively searching for – the product offered by the domestic business. Otherwise, the consumer would not have googled the product and clicked on the ad placed by the domestic business. In this sense, we can also think of the US-Canada border effect for clicks as capturing the difficulties of domestic businesses and foreign consumers finding each other.

[20] A user cannot see an ad unless the advertiser that placed the ad specifically targets the user's geographic location.

Our results show that the US-Canada border does, indeed, have a dampening effect on clicks. In column (1) of Table 6, the estimated coefficient for the US↔CA border dummy is negative and statistically significant. Taking the exponent of this coefficient gives a border effect of 4.06, which indicates that the number of clicks that users and businesses located within the same country generate are 4.06 times higher than the number of clicks generated by users and businesses located on opposite sides of the US-Canada border.

Table 6: Results for Clicks versus Conversions

	Clicks	Clicks minus conversions	Conversions	Conversion to click ratio
	(1)	(2)	(3)	(4)
$\ln d_{ij}$	-0.001	-0.006	0.035	-0.002
	(0.013)	(0.012)	(0.023)	(0.002)
$contig\ US\text{-}CA_{ij}$	-0.009	-0.024	0.079	0.005
	(0.068)	(0.071)	(0.111)	(0.006)
$contig\ CA_{ij}$	0.002	0.018	-0.044	0.011
	(0.215)	(0.219)	(0.232)	(0.030)
$contig\ US_{ij}$	0.120***	0.107***	0.231***	0.003
	(0.023)	(0.022)	(0.043)	(0.003)
$internal_{ij}$	0.503***	0.484***	0.686***	0.014
	(0.057)	(0.055)	(0.099)	(0.012)
$border\ US \leftrightarrow CA_{ij}$	-1.400***	-1.347***	-1.805***	-0.023***
	(0.069)	(0.071)	(0.091)	(0.006)
Implied US↔CA Border Effect	4.06	3.85	6.08	n/a†
Observations	3,481	3,481	3,481	3,481

Importer and exporter fixed effects included. The dependent variable is the number of clicks in column (1), the difference between the number of clicks and the number of conversions in column (2), the number of conversions in column (3), and the ratio of conversions to clicks in column (4). All dependent variables are aggregated over all sectors of economic activity and over all conversion types and they cover three years, 2008 to 2010. Estimations use PPML in columns (1) to (3) and fractional Probit in column (4). The coefficients reported in column (4) represent average marginal effects. Robust standard errors (clustering by region pair).
† Not available, as the estimation is performed using a fractional Probit.
*** Significant at 1%, ** significant at 5%, * significant at 10%

In columns (2) and (3) of Table 6, we decompose the total number of clicks into clicks that do not convert and clicks that do convert for the advertisers. In column (2), the dependent variable is the difference between clicks and conversions. This variable measures the number of times that users clicked on ads but did not purchase anything. In column (3), the dependent variable is conversions. This variable measures the number of

times that users clicked on ads and reached the advertisers' "thank you for your order" pages.

The results reported in columns (2) and (3) show a discrepancy between the coefficient estimated for the US↔CA border dummy when using the difference between clicks and conversions as the dependent variable and the coefficient estimated when using conversions on the left hand side. In particular, the US-Canada border seems to have a more severe dampening effect on conversions than on clicks that do not lead to conversions.

The difference between these two estimates is likely due to consumers' reluctance to transact with foreign sellers. Assume we have a California-based bicycle seller that places ads in Ontario. Now assume that following an ad click, no transaction takes place between the seller and the user. This can happen if the seller is unwilling to sell to an Ontario-based user or if the buyer is unwilling to buy from a California-based seller. The latter is much more likely. By making the ad visible within Ontario and by paying for the ad click, the bicycle seller signals his willingness and ability to trade with the Ontario-based user. The user, however, upon clicking on the ad, might decide that he is unwilling to trade with a California-based seller. This can happen, for example, if the user learns upon visiting the seller's website that the seller offers limited warranty coverage to Canadian customers or if the seller is unable to offer next-day deliveries to international addresses.

The example above shows that following an ad click, consumer, rather than producer, preferences are likely to dictate whether a cross-border sale happens. The business placing the ad is willing to sell abroad, the consumer finding the ad is interested in the product, but upon visiting the business' site, the consumer decides – for whatever reason – to not purchase from a foreign-based seller.

In column (4), we further investigate the possibility that consumers might be reluctant to trade with foreign sellers. In this column, the dependent variable is the ratio of conversions to clicks. Since this variable lives on the [0,1] interval, we estimate our regression using a fractional Probit. The coefficients we report in Table 6 represent average marginal effects.

If consumers are reluctant to trade with foreign sellers, we would expect to see a negative and significant US-Canada border effect in column (4). This is, indeed, the case. The estimated coefficient for the US↔CA border dummy is -0.023, and it is significant at the 1% level. This indicates that all else being equal, an ad click has a lower probability to convert into a sale when the transacting parties are not located within the same country. Interestingly, we do not notice the same bias for same-state transactions. The coefficient on the same-state dummy variable is insignificant. Our results show that the only border that has a negative impact on the conversion to click ratio is the international border. This is the first indication we have that consumers' reluctance to trade with foreign sellers might be one of the factors behind the US-Canada border effects observed in the data.

6 Sector Level Analysis

We underlined in the literature review the fact that a number of researchers suggest that aggregating data over sectors of economic activity leads to biased estimates for the border effect. To address these concerns, in this section we present the results we obtain when we disaggregate our data by sectors of economic activity.

6.1 Results at the NAICS2 Level

The first column of Table 7 presents the coefficient for distance obtained by applying our main specification to each NAICS2 sector reflected in our data set. The second column reports the exponent of the coefficient of the same-state dummy variable, which gives the home bias effect. The third column contains the exponent of the coefficient of the US-Canada border dummy, which gives the average border effect. We do not report standard errors in order to make the table easier to read, but we do gray out and mark with an ^{o} the coefficients that are not statistically significant at the 1% level.[21]

[21]The complete set of results, which reflect the values of all coefficients included in our regressions and their corresponding standard errors, is presented in Appendix A.

Table 7: Results at the NAICS2 Level

NAICS2 Sector:	$\ln d_{ij}$ (1)	$internal_{ij}$ (exponent) (2)	$border\ US \leftrightarrow CA_{ij}$ (exponent) (3)
Administrative, support, and waste management	0.056^{o}	2.01	26.63
Public administration	-0.117	1.74^{o}	21.48
Real estate and rental and leasing	-0.221	1.48^{o}	19.83
Health care and social assistance	0.102^{o}	2.47	13.21
Accommodation and food services	-0.119^{o}	2.74	12.88
Transportation and warehousing	-0.039^{o}	1.70	10.86
Finance and insurance	0.023^{o}	1.85	9.78
Education services	0.057^{o}	3.17	8.92
Retail trade	0.008^{o}	1.67	6.20
Information and cultural industries	0.059^{o}	2.02	5.04
Arts, entertainment, and recreation	0.064^{o}	6.22	4.90
Wholesale trade	-0.070	1.37	4.59
Other services	0.007^{o}	1.97	4.18
Professional, scientific, and technical services	-0.102	1.31	4.05
Utilities	-0.124^{o}	4.21	3.59

Importer and exporter fixed effects included. The dependent variable is the levels of conversion counts aggregated over all conversion types. Estimations performed using PPML. The data cover four years, 2008 to 2011. Robust standard errors (clustering by region pair). Sample size is 3,481 for all regressions.
^{o} not significant at the 1% level.

We order the NAICS2 sectors by the size of the estimated US-Canada border effect. The double line differentiates between the sectors for which the border effect is higher, and the sectors for which it is lower, than the 6.71 estimate we obtain when we use aggregate data.

The wide differences across sectors between border coefficients provide further justification for previous researchers' suggestion to use disaggregated data. Aggregation can bias gravity estimates, and it hides important sector-level dynamics. In our data set, the lowest coefficient that we find for the US↔CA dummy variable is -1.278, while the highest is -3.282. This large difference points to the importance of using disaggregated

data when studying border effects in a gravity setting.

Before we discuss in detail the coefficients reported in Table 7, we would like to remind the reader as to what these estimated coefficients represent. As noted in Section 3.1, the coefficients we are estimating are the product of the elasticity of substitution, $(1-\sigma)$, and the elasticity of total trade costs with respect to each trade barrier variable. Prior research provides substantial evidence that σ, the elasticity of substitution, varies greatly across industries.[22] Since we cannot assume that σ is constant across industries, we cannot attribute the differences between the coefficients reported in Table 7 solely to differences in trade frictions across sectors. In other words, if we estimate a smaller coefficient on distance or the border dummies, for one sector, we cannot conclude that trade frictions are necessarily smaller for that sector. The estimated coefficient might be smaller simply because the bundle of products in that sector has a lower elasticity of substitution than the bundle of products in the other sectors.

Although we cannot disentangle the effect of the elasticity of substitution from the effect of the elasticity of the total trade costs with respect to each trade barrier, the ranking of the sectors in Table 7 may nevertheless provide some evidence that the subset of services which can only be consumed locally encounters large border effects. [23]

In our data, the sector with the largest estimated coefficient for the US↔CA dummy variable is NAICS 56: Administrative, support, and waste management services. This sector covers establishments which provide services such as waste disposal, office cleaning and administration, or security and surveillance. Since these services can only be consumed locally, in the place where they are performed, we would expect them to encounter high intranational as well as international border effects.

The second sector ranked in Table 7 is Public administration. Services under this sector travel fairly well within a country, as evidenced by the insignificant home bias effect. However, we would expect them to face large international border barriers. This is because Public administration services primarily target a country's own citizens.

[22] See, for example, Broda and Weinstein (2006), Chen and Novy (2011) or Imbs and Mejean (2015).
[23] This is a hypothesis proposed by Hortaçsu et al. (2009), as well, to explain high levels of same city sales.

The sectors ranked third and fourth in Table 7 are Real estate and rental and leasing, and Health care and social assistance. The services provided under both of these classifications tend to be heavily localized. Real estate services, in particular, are physically bound to a particular location. In our regressions, this sector is one of only four industries for which the coefficient on distance is statistically different from zero.

The fact that all four service sectors ranked at the top of Table 7 describe bundles of services that can only be consumed locally make it less likely that the large coefficients estimated for the US↔CA dummy variable for each of these sectors are merely the result of these industries having large elasticities of substitution. The local or national character of the services described by these four classifications is likely to be one of the drivers behind the large border effects for these industries – and the large estimates we obtain for the US↔CA dummy variable coefficients.

In Table 7, we are also interested in the estimated border effect for the Retail trade sector. As noted in the data section, almost a third of all conversions reflected in our data set fall within Retail trade. In the online world, this sector primarily covers business-to-consumer or consumer-to-consumer transactions in final goods. The estimates reported in Table 7 show that the US-Canada border substantially hinders Retail trade. All else being equal, intranational trade for this sector is 6.2 times higher than international trade. We devote our next section to understanding the factors that drive this high border effect.

6.2 Retail Trade Data Broken Down by NAICS3 Sectors

We are able to categorize the conversions that fall within the Retail trade sector at the NAICS 3-digit level. Although this further disaggregation does not provide a complete explanation for the existence of the border effect, it is helpful in identifying potential factors that give rise to it, and in understanding whether appropriate policy measures can be designed to lower it. Table 8 presents the results we obtain by applying our main specification to each NAICS3 classification within the Retail trade sector. We gray out and mark with an ^{o} the coefficients that are not statistically significant at the 1% level.[24]

[24]The complete set of results is presented in Appendix B.

We find that Motor vehicle and parts dealers has a very large estimated coefficient for the US↔CA dummy variable. This might be due to the regulatory hurdles that consumers face when buying a vehicle abroad. A Canadian customer wishing to import a car from a US owner or dealership, for example, will find it very difficult. Besides the customs forms she has to fill out, the Canadian customer must submit proof to the Canadian Registrar of Imported Vehicles (RIV) that the vehicle has no outstanding recalls and that it suffered no modifications, such as being adapted for disabled access. After submitting the required paperwork, the Canadian customer must pay a RIV fee and present the vehicle for a federal standards inspection. To pass the inspection, she must ensure that the US vehicle she is bringing into the country is in compliance with Canadian requirements such as bilingual and metric labeling. Vehicles that pass the RIV inspection can then be registered and licensed in Canada. Vehicles that do not pass the RIV inspection have to be exported back to the US or destroyed under the supervision of Canadian customs officials.[25]

Table 8: Results at the NAICS3 Level: Retail Trade

	$\ln d_{ij}$	$internal_{ij}$ (exponent)	$border\ US \leftrightarrow CA_{ij}$ (exponent)
NAICS3 Sector:	(1)	(2)	(3)
Motor vehicle and parts dealers	-0.051^{o}	2.74	21.80
Furniture and home furnishings	0.013^{o}	3.42	17.89
Food and beverage stores	-0.021^{o}	1.21^{o}	14.40
Clothing and clothing accessory stores	-0.015^{o}	1.31	9.53
Building material and garden equipment	-0.038^{o}	3.11	4.73
Electronics and appliance stores	-0.008^{o}	1.13^{o}	3.86
Sporting goods, hobby, book, and music	0.006^{o}	1.41	3.82
Health and personal care stores	-0.031^{o}	1.52	3.67

Importer and exporter fixed effects included. The dependent variable is the levels of conversion counts aggregated over all conversion types. Estimations performed using PPML. The data cover four years, 2008 to 2011. Robust standard errors (clustering by region pair). Sample size is 3,481 for all regressions.
^{o} not significant at the 1% level.

[25]The steps described above are provided in detail on Transport Canada's website, at `www.tc.gc.ca`, accessed on June 15, 2015. We summarize the procedures for importing a car from the US into Canada, rather than the other way around, as cars in the United States are usually cheaper than cars in Canada, so trade in motor vehicles is more likely to initiate in the US.

Even once their vehicles are licensed, the long list of bureaucratic and regulatory hurdles that Canadian customers face when purchasing cars across the border does not end. If their cars purchased from the United States break down, Canadian customers might end up having to pay for all the repairs, as many Canadian companies do not honor U.S. factory warranties.

Given the bureaucratic and regulatory procedures that end consumers must face when purchasing a car online, it is, perhaps, not surprising that for Motor vehicle and parts dealers, trade across the US-Canada border is 21.8 times less than trade within these countries' borders. This is the first indication we have that regulation may hinder online trade and that policy aimed at lessening the regulatory burden faced by end consumers would be desirable.

In Table 8, Food and beverage stores also have a high estimated coefficient for the US↔CA dummy variable. Interestingly, food and beverages travel well within a country. Table 8 shows that the estimated coefficient for the $internal_{ij}$ dummy variable is not significant. Food and beverages, therefore, appear to cross their own home state or province border unhindered, but they encounter obstacles at the international border. The estimated border effect is 14.40, indicating that for food and beverages, intranational trade is more than 14 times higher than international trade. If the large border effect for this category of products is due to regulatory differences between the two countries, appropriate policy measures can be proposed to lower the border effect for this category of goods.

Table 8 reveals that products classified as Furniture and home furnishings encounter very large intranational and international border effects. This might be due to the bulkiness of these products. At the other end of the spectrum, we find that Electronics and appliance stores, Sporting goods, hobby, book, and music products, as well as Health and personal care stores have fairly low estimated coefficients for the US↔CA dummy variable. This might be because there is little risk involved in ordering the wrong product. An iPhone, a book, or a Crest toothpaste are likely to be the same, whether purchased in the US or Canada. By comparison, Clothing and clothing accessory stores have an estimated border effect that is much higher. 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. For goods where the possibility of returning or exchanging the product is high, trading across international borders might be too daunting of a prospect for buyers and sellers alike. If this is the case, policy measures are unlikely to be able to lower the border effect for this category of goods.

7 Conclusion

In this study, we use a proprietary data set from Google to examine the US-Canada border effect. We add to the existing literature on two important dimensions. First, we are the only study to quantify the effect that the US-Canada border has on trade in the virtual world. Second, our data allow us to obtain insights into the factors that drive the border effect in the virtual world.

We find that the US-Canada border influences online trade flows negatively and significantly. All else being equal, online trade between two US states or two Canadian provinces is 6.7 times higher than trade between a US state and a Canadian province. In the virtual world, where information costs are very low, this is a surprising finding.

Another surprising finding is that even in the absence of business-to-business transactions involving intermediate inputs, the border effect remains high. For Retail trade, a classification that encompasses cross-border transactions in final goods, the estimated US-Canada border effect is large and significant. Holding everything else constant, intranational online Retail trade is 6.2 times its international equivalent. In a world where multistage production processes do not determine the pattern of trade, the border effect remains stubbornly high.

Online trade does seem to be affected, to some extent, by trade barriers. Disaggregating Retail trade at the NAICS3 level shows that goods like motor vehicles, which face heavy regulatory and bureaucratic hurdles when crossing the border between the United States and Canada, encounter very high international border effects. This result suggests

that there is a role for policy to reduce the online border effect for these products.

Similar to Hortaçsu et al. (2009), our results indicate that some services traded online encounter high border effects due to their consumption being tied to a particular location. For these service categories, there is no role for policy in reducing the border effects we find in the data.

Finally, our analysis supports previous research that argues for the use of disaggregated data when studying border effects. When disaggregating our data to the NAICS2 level, we find that the effect of the border between the US and Canada varies widely across sectors. The US-Canada border affects the sector with the highest coefficient 7.4 times more than it affects the sector with the lowest coefficient.

The emergence of online trade creates new opportunities and challenges for the economics of trade. Technology creates the opportunity for buyers and sellers in different countries to reduce their search and, in some cases, transaction 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 datasets relating to online trade. As better data become available, we expect that more researchers will join this agenda to uncover the geographical trade patterns in the online world.

References

Anderson, James E and Eric van Wincoop, "Gravity with gravitas: A solution to the border puzzle," *American Economic Review*, 2003, *93* (1), 170–192.

_ **and Yoto V Yotov**, "The changing incidence of geography," *American Economic Review*, 2010, *100* (5), 2157–2186.

_ **, Catherine A Milot, and Yoto V Yotov**, "How much does geography deflect services trade? Canadian answers," *International Economic Review*, 2014, *55* (3), 791–818.

Boivin, Jean, Robert Clark, and Nicolas Vincent, "Virtual borders," *Journal of International Economics*, 2012, *86* (2), 327–335.

Broda, Christian and David E Weinstein, "Globalization and the gains from variety," *Quarterly Journal of Economics*, 2006, *121* (2), 541–585.

_ **and** _ , "Understanding international price differences using barcode data," Technical Report, National Bureau of Economic Research 2008.

Brown, W Mark and William P Anderson, "Spatial markets and the potential for economic integration between Canadian and US regions," *Papers in Regional Science*, 2002, *81* (1), 99–120.

Chen, Natalie, "Intra-national versus international trade in the European Union: Why do national borders matter?," *Journal of International Economics*, 2004, *63* (1), 93–118.

_ **and Dennis Novy**, "Gravity, trade integration, and heterogeneity across industries," *Journal of International Economics*, 2011, *85* (2), 206–221.

Choi, Hyunyoung and Hal Varian, "Predicting the present with Google Trends," *Economic Record*, 2012, *88* (s1), 2–9.

Combes, Pierre-Philippe, Miren Lafourcade, and Thierry Mayer, "The trade-creating effects of business and social networks: Evidence from France," *Journal of International Economics*, 2005, *66* (1), 1–29.

Coughlin, Cletus C and Dennis Novy, "Is the international border effect larger than the domestic border effect? Evidence from US trade," *CESifo Economic Studies*, 2013, *59* (2), 249–276.

Eaton, Jonathan and Samuel Kortum, "Technology, geography, and trade," *Econometrica*, 2002, *70* (2), 1741–1779.

Einav, Liran, Dan Knoepfle, Jonathan Levin, and Neel Sundaresan, "Sales taxes and internet commerce," *American Economic Review*, 2014, *104* (1), 1–26.

Evans, Carolyn L, "The economic significance of national border effects," *American Economic Review*, 2003, *93* (4), 1291–1312.

Feenstra, Robert C, "Border effects and the gravity equation: Consistent methods for estimation," *Scottish Journal of Political Economy*, 2002, *49* (5), 491–506.

Garmendia, Aitor, Carlos Llano, Asier Minondo, and Francisco Requena, "Networks and the disappearance of the intranational home bias," *Economics Letters*, 2012, *116* (2), 178–182.

Gopinath, Gita, Pierre-Olivier Gourinchas, Chang-Tai Hsieh, and Nicholas Li, "International prices, costs, and markup differences," *American Economic Review*, 2011, *101* (6), 2450–2486.

Gorodnichenko, Yuriy and Linda L Tesar, "Border effect or country effect? Seattle may not be so far from Vancouver after all," *American Economic Journal: Macroeconomics*, 2009, *1* (1), 219–241.

Head, Keith and Thierry Mayer, "Non-Europe: The magnitude and causes of market fragmentation in the EU," *Weltwirtschaftliches Archiv*, 2000, *136* (2), 284–314.

_ **and** _ , "Illusory border effects: Distance mismeasurement inflates estimates of home bias in trade," Technical Report, CEPII Research Center 2002.

_ **and** _ , "Gravity equations: Workhorse, toolkit, and cookbook," *Handbook of International Economics*, 2014, *4*, 131–195.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein, "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics*, 2008, *123* (2), 441–487.

Hillberry, Russell and David Hummels, "Intranational home bias: Some explanations," *Review of Economics and Statistics*, 2003, *85* (4), 1089–1092.

_ **and** _ , "Trade responses to geographic frictions: A decomposition using micro-data," *European Economic Review*, 2008, *52* (3), 527–550.

Hillberry, Russell H, "Aggregation bias, compositional change, and the border effect," *Canadian Journal of Economics*, 2002, *35* (3), 517–530.

Hortaçsu, Ali, F Asís Martínez-Jerez, and Jason Douglas, "The geography of trade in online transactions: Evidence from eBay and MercadoLibre," *American Economic Journal: Microeconomics*, 2009, *1* (1), 53–74.

Imbs, Jean and Isabelle Mejean, "Elasticity optimism," *American Economic Journal: Macroeconomics*, 2015, *7* (3), 43–83.

Lendle, Andreas, Marcelo Olarreaga, Simon Schropp, and Pierre-Louis Vézina, "There goes gravity: eBay and the death of distance," *Economic Journal*, 2015.

McCallum, John, "National borders matter: Canada-US regional trade patterns," *American Economic Review*, 1995, *85* (3), 615–623.

Millimet, Daniel L and Thomas Osang, "Do state borders matter for US intranational trade? The role of history and internal migration," *Canadian Journal of Economics*, 2007, *40* (1), 93–126.

Obstfeld, Maurice and Kenneth Rogoff, "The six major puzzles in international macroeconomics: Is there a common cause?," in "NBER Macroeconomics Annual 2000, Volume 15," MIT press, 2001, pp. 339–412.

Ōmae, Kenichi, *The end of the nation state: The rise of regional economies*, Simon and Schuster, 1995.

Rauch, James E, "Networks versus markets in international trade," *Journal of International Economics*, 1999, *48* (1), 7–35.

Rossi-Hansberg, Esteban, "A spatial theory of trade," *American Economic Review*, 2005, *95* (5), 1464–1491.

Silva, JMC Santos and Silvana Tenreyro, "The log of gravity," *Review of Economics and statistics*, 2006, *88* (4), 641–658.

Wolf, Holger C, "Intranational home bias in trade," *Review of Economics and Statistics*, 2000, *82* (4), 555–563.

Yi, Kei-Mu, "Can multistage production explain the home bias in trade?," *American Economic Review*, 2010, *100* (1), 364–393.

Appendix

A NAICS2: Full Set of Results

Table 9: Results at the NAICS2 Level

	$\ln d_{ij}$	$contig$ $US\text{-}CA_{ij}$	$contig$ CA_{ij}	$contig$ US_{ij}	$internal_{ij}$	$border$ $US \leftrightarrow CA_{ij}$
Admin., support, and waste mgmt.	0.056 (0.037)	0.989*** (0.197)	0.064 (0.219)	0.411*** (0.096)	0.699*** (0.117)	-3.282*** (0.098)
Public administration	-0.117*** (0.041)	0.161 (0.366)	-0.665** (0.265)	-0.002 (0.094)	0.553** (0.238)	-3.067*** (0.160)
Real estate and rental and leasing	-0.221*** (0.079)	0.684* (0.366)	-1.004*** (0.340)	-0.083 (0.147)	0.391* (0.213)	-2.987*** (0.156)
Health care and social assistance	0.102** (0.051)	1.324*** (0.282)	-0.404 (0.350)	0.210** (0.102)	0.904*** (0.168)	-2.581*** (0.157)
Accomodation and food services	-0.119 (0.077)	-1.740** (0.791)	-0.280 (0.609)	-0.087 (0.131)	1.007*** (0.253)	-2.556*** (0.199)
Transportation and warehousing	-0.039 (0.032)	0.792*** (0.205)	1.044*** (0.384)	0.074 (0.070)	0.533*** (0.108)	-2.385*** (0.127)
Finance and insurance	0.023 (0.015)	-0.156 (0.121)	0.212 (0.261)	0.142*** (0.030)	0.614*** (0.068)	-2.280*** (0.105)
Education services	0.057* (0.034)	0.021 (0.335)	-1.102** (0.448)	0.263*** (0.068)	1.152*** (0.116)	-2.188*** (0.160)
Retail trade	0.008 (0.022)	0.011 (0.149)	-0.030 (0.275)	0.186*** (0.046)	0.514*** (0.116)	-1.825*** (0.084)
Info. and cultural industries	0.059** (0.027)	-0.128 (0.179)	0.261 (0.413)	0.165*** (0.049)	0.701*** (0.121)	-1.618*** (0.121)
Arts, entertainment, and recreation	0.064 (0.075)	0.020 (0.365)	-0.827** (0.379)	0.453*** (0.151)	1.828*** (0.343)	-1.589*** (0.122)
Wholesale trade	-0.070*** (0.016)	0.233 (0.293)	0.526** (0.218)	0.031 (0.032)	0.317*** (0.060)	-1.523*** (0.109)
Other services	0.007 (0.020)	0.053 (0.107)	-0.542** (0.254)	0.089* (0.049)	0.676*** (0.104)	-1.431*** (0.088)
Prof., scientific, and technical services	-0.102*** (0.029)	-0.014 (0.201)	-0.091 (0.642)	-0.042 (0.051)	0.268*** (0.103)	-1.399*** (0.137)
Utilities	-0.124** (0.052)	-0.297 (0.198)	-0.600 (0.407)	0.381*** (0.103)	1.438*** (0.199)	-1.278*** (0.140)

Importer and exporter fixed effects included. The dependent variable is the levels of conversion counts aggregated over all conversion types. Estimations performed using PPML. The data cover four years, 2008 to 2011. Robust standard errors (clustering by region pair). Sample size is 3,481 for all regressions.
*** Significant at 1%, ** significant at 5%, * significant at 10%

B NAICS3 – Retail Trade: Full Set of Results

Table 10: Results at the NAICS3 Level: Retail Trade

	$\ln d_{ij}$	*contig* $US\text{-}CA_{ij}$	*contig* CA_{ij}	*contig* US_{ij}	$internal_{ij}$	*border* $US \leftrightarrow CA_{ij}$
Motor vehicle and parts dealers	-0.051	0.164	-0.515	0.187*	1.006***	-3.082***
	(0.055)	(0.497)	(0.337)	(0.110)	(0.184)	(0.169)
Furniture and home furnishings	0.013	-0.091	0.470	0.486***	1.230***	-2.884***
	(0.093)	(0.744)	(0.318)	(0.146)	(0.467)	(0.166)
Food and beverage stores	-0.021	-2.243***	0.233	0.286	0.191	-2.667***
	(0.088)	(0.615)	(0.260)	(0.196)	(0.351)	(0.189)
Clothing and clothing accessory stores	-0.015	0.197	-0.651***	0.079**	0.270***	-2.254***
	(0.017)	(0.157)	(0.129)	(0.040)	(0.068)	(0.054)
Building material and garden equipment	-0.038	0.428	-0.037	0.185***	1.134***	-1.553***
	(0.029)	(0.444)	(0.396)	(0.055)	(0.099)	(0.110)
Electronics and appliance stores	-0.008	-0.020	-0.352	0.049	0.125	-1.351***
	(0.020)	(0.178)	(0.258)	(0.036)	(0.082)	(0.115)
Sporting goods, hobby, book, and music	0.006	-0.392**	-0.773***	0.116***	0.340***	-1.341***
	(0.017)	(0.169)	(0.283)	(0.038)	(0.075)	(0.119)
Health and personal care stores	-0.031	-0.069	-0.221	0.017	0.420***	-1.299***
	(0.022)	(0.068)	(0.300)	(0.045)	(0.078)	(0.101)

Importer and exporter fixed effects included. The dependent variable is the levels of conversion counts aggregated over all conversion types. Estimations performed using PPML. The data cover four years, 2008 to 2011. Robust standard errors (clustering by region pair). Sample size is 3,481 for all regressions.
*** Significant at 1%, ** significant at 5%, * significant at 10%

CHAPTER 3

Worldwide Gravity in Online Commerce

Cosmina Dorobanțu and Bo Cowgill*

Abstract

This paper analyzes geographical patterns of cross-country Internet transactions using proprietary data from Google. Covering 146 countries from 2008-2011 and over 10 billion online transactions, the data allow us to examine a fast growing area of trade. The paper finds the effect of distance on trade to be around -0.53, indicating that distance has a negative effect on trade even in the virtual marketplace. The study also finds that cultural characteristics, such as a shared language, have a large impact on e-commerce, while economic ties, such as a common currency, have an insignificant effect. The paper underlines the importance of accounting for selection into trade in worldwide gravity estimations and identifies two exclusion restrictions that can be used when examining online trade flows.

1 Introduction

Few empirical results in economics have been more widely confirmed than the negative relationship between distance and trade. Since the 1960s, economists have estimated gravity equations to measure the relationship of distance and trade between countries. Head and Mayer (2014) collected over 2,500 of these estimates and calculated the mean of the estimated coefficient on distance. They find that the average of the estimated distance elasticity is -0.93, indicating that a 10% increase in the distance between two countries leads to a 9.3% decrease in trade between them, all else being equal.[1]

*The authors thank Google for sharing the data used in this paper. We also thank Avinash Dixit, Keith Head, Beata Javorcik, Bertin Martens, Marc Melitz, Dennis Novy, Ferdinand Rauch, Andrés Rodríguez-Clare, Tony Venables, Pierre-Louis Vézina, Noam Yuchtman, and seminar participants at Oxford, the European Trade Study Group, and the European Commission for helpful comments. Dorobanțu acknowledges financial support from the Economic and Social Research Council.

[1]Head and Mayer (2014) expands Disdier and Head (2008)'s meta-analysis of 1,467 estimates of the distance coefficient in gravity settings. The authors generate the -0.93 average distance coefficient by averaging over all estimates included in their meta analysis, regardless of whether these estimates were obtained using a naive or a structural gravity approach.

Our paper's main contribution is to advance the nascent literature on the role of distance in the virtual world. Researchers are interested in understanding how distance affects online trade for several reasons. First, information costs are lower online. A company that has an online presence can sell to a foreign country without sending its analysts there to do extensive market research. Similarly, consumers wishing to buy a product from a foreign company face close to zero search costs, as all they need to do in order to connect with a foreign seller is conduct an Internet search. Allen (2014) shows that information costs account for a significant portion of the distance effect in the offline world. In the online environment, where information costs are lower, how much does distance matter for trade?

Second, online transactions are no longer a negligible part of the global economy. Online commerce recently passed the $1 trillion mark and its growth shows no signs of abating. An obvious question is whether the rules governing trade in the offline world also apply online: does distance still matter in a world that is more connected every day? The business press, represented by books such as Frances Cairncross' *The Death of Distance: How the Communication Revolution is Changing our Lives* or Thomas Friedman's *The World is Flat*, has an answer: distance is no longer relevant. Are these authors right? Is distance irrelevant for this new – and fast growing – form of commerce?

Despite the academic and business interest in this topic, only four other papers estimate the elasticity of distance for online trade flows in a worldwide gravity setting: Lendle et al. (2016), Lendle and Vézina (2015), Hortaçsu et al. (2009), and Blum and Goldfarb (2006). Their findings vary significantly. The lowest estimate the previous researchers obtain is -0.31, while the highest is -1.17.

In our paper, we improve the precision of these estimates by using a comprehensive data set and a methodology that corrects known econometric biases. In terms of the data, our sample reflects over 10 billion transactions conducted online over a four year timeframe, 2008-2011. Notably, our data set covers 146 countries in the world, which is a significant improvement over the other four studies. Lendle et al. (2016), the most extensive online worldwide gravity analysis to date, covers only 61 countries. We show that such limited coverage biases the distance estimate downwards.

In terms of methodology, we improve on previous online gravity work by using the methodology proposed by Helpman et al. (2008). At a worldwide level, in online trade flows, we observe many zeros. In our data, only 25% of all country pairs trade in both directions, about 30% trade in one direction only, and 45% do not trade at all. If the high number of zero trade flows between countries is the result of an economically meaningful selection process, estimators that do not account for this selection process suffer from selection and omitted variable biases.

Prior to us, no other researchers that examined the effect of distance on online trade accounted for selection into trade. We show that a simple OLS estimation that does not account for selection into trade biases the coefficient of distance upwards by as much as 17%.

One of the reasons why researchers do not account for selection into trade is that the approach requires finding an exclusion restriction. Our paper's second contribution is identifying two exclusion restrictions that can be used in the context of online trade. The first one measures the difficulty of building warehouses, which is the primary physical asset that online retails need when expanding abroad. The second one relates to countries using a common currency. We show that for online trade flows – unlike for traditional trade flows – a common currency affects the fixed, but not the variable, cost of exporting.

Our analysis reveals that conditional on exporting, a 10% increase in the distance between two countries reduces the volume of trade between them by approximately 5.3%. We conclude that at the worldwide level, distance is certainly still alive in the online world.

Section 2 presents an overview of the relevant literature. It describes the studies done to date examining the effects of distance both in the online and the offline world and provides an overview of Helpman et al. (2008). Section 3 introduces the data we use for our analysis. Section 4 describes the empirical approach. Section 5 presents the results and Section 6 concludes.

2 Relevant Literature

2.1 Gravity Equations

In order to identify the role of distance, trade economists typically resort to gravity equations. These equations have been estimated since the 1960s, based on a simple concept borrowed from physics: trade between two countries depends positively on their sizes and negatively on the distance between them.[2] Although gravity equations perform very well empirically, they had no theoretical foundations until a little over a decade ago.

Anderson and van Wincoop (2003) used a demand-side structure and relied on the simplifying Armington assumption, namely that goods are differentiated by place of origin, to derive the gravity equation. Around the same time, Eaton and Kortum (2002) used a supply-side structure to arrive at the gravity equation, in a model employing Ricardian technology with heterogeneous productivity for each country and good. A few years later, Chaney (2008), Helpman et al. (2008), and Melitz and Ottaviano (2008) integrated gravity within Melitz (2003) style heterogeneous firm models. More recently, Arkolakis et al. (2012) showed that a wide class of models, with or without heterogeneous firms, can serve as the theoretical foundation from which the gravity equation derives.

Head and Mayer (2014) provide an excellent summary of the theoretical, as well as the empirical results, of the gravity literature.[3] The variables included in empirical gravity analyses typically measure the effect of the geographic, cultural or economic closeness between trading partners. Although our primary interest relates to the role of distance, we also examine the effect on online trade of other geographical, cultural, and economic factors. We summarize below the results Head and Mayer (2014) obtain for each of these factors.

[2] Rauch (2015) provides a comprehensive overview of the similarity between gravity in physics and gravity in trade.

[3] Similar summaries are provided in Anderson (2011) and Baltagi et al. (2014).

2.1.1 Distance

The physical distance between two countries is a ubiquitous variable in any gravity equation. Head and Mayer (2014) analyze over 2,500 estimates of the distance coefficient in gravity settings and find that the average of the estimated distance elasticity is -0.93. Distance, in other words, significantly reduces trade and this empirical result holds across a wide range of samples and methodologies.

While researchers widely accept the fact that distance has a negative effect on trade, there is no consensus in the literature as to why distance reduces trade. Over the years, several theories have been proposed. We summarize the most relevant ones below.

Transport Costs

One of the most obvious explanations as to why distance has a negative effect on trade is transport costs: the further apart two countries are, the more costly it is to ship goods between them. However, although shipping costs increase with distance, researchers like Grossman (1998), Hummels (1999), and Allen (2014) convincingly argue that freight shipping costs, by themselves, cannot fully explain why trade flows decline with distance; the magnitude of the distance effect is simply too large to be explained by shipping costs alone.

Harrigan and Venables (2006) and Evans and Harrigan (2005) show that besides shipping costs, goods also incur time costs as they are transported between countries. The larger the distance, the more time it takes for goods to be delivered. Harrigan and Venables (2006) show that delivery time for intermediate goods is costly for producers. As it takes time for components to be delivered, producers must place orders in advance, while they still face uncertainties regarding final demand. Furthermore, producers might see their production processes disrupted by unsynchronized deliveries. Harrigan and Venables (2006) argue that producers' desire to minimize the uncertainties they face as a result of delivery times for intermediate goods alters the efficient spatial organization of production.

In related work, Evans and Harrigan (2005) highlight the implications of time costs for retailers. In particular, they show that timely deliveries – which can only be achieved from nearby locations – reduce retailers' costs by allowing them to respond to fluctuations in demand without having to hold costly inventories.

Although the studies highlighted above are convincing, if distance proxies only for transport costs – be they freight shipping costs or time costs – then we would expect proximity to play no role in gravity models where the goods traded do not incur transport costs. Blum and Goldfarb (2006), however, examine website visits made by American Internet users to 46 foreign countries and find a distance elasticity of -1.17. Since website visits incur no transport costs, Blum and Goldfarb (2006)'s result indicates that distance must capture other trade impediments, besides shipping or time costs. We explore some of these frictions below.

Search Costs

Several authors argue that distance also captures search costs. In a seminal work, Rauch (1999) notes that it is costly for buyers and sellers to find each other, and shows that geographical proximity has a bigger impact on matching buyers and sellers of differentiated products (which are subject to greater informational impediments) than on matching buyers and sellers of homogenous products.

In related work, Chaney (2013) proposes a theoretical model of input-output linkages between firms to highlight the importance of the costs involved in finding suppliers and consumers. In Chaney (2013)'s model, firms have two options to engage in trade. The first one is to pay a direct cost to make a foreign contact, while the second one is to interact directly with one's existing contacts and get introduced to their contacts. This need for direct contact in order for trade to take place allows Chaney (2013) to explain why distance is such a large deterrent to trade.

More recently, Allen (2014) argues that distance also captures the costs of obtaining information about market conditions elsewhere. He shows that in a perfect competition model where producers incur costs in order to obtain information about market condi-

tions in other locations, trade flows decline with distance not only because of transport costs, but also due to the fact that the probability of searching a destination decreases.

Trust

Besides transport and search costs, some researchers suggest that trust also contributes to the large distance effects researchers find in the data. Guiso et al. (2009) measure bilateral trust between countries and, using a gravity setting, show that the distance between two countries has a negative and significant effect on bilateral trust. In a related study, Huang (2007) show that distance affects countries that have high uncertainty-aversion more, and that these countries export disproportionately less to remote partners than a gravity equation would predict. Finally, in studies using data from eBay, Lendle et al. (2016) and Hortaçsu et al. (2009) show that distance has a bigger effect on transactions involving untrustworthy sellers, that have poor ratings on the site.

Localized Tastes

Ferreira and Waldfogel (2013) and Blum and Goldfarb (2006) show that in gravity settings, distance might also account for the existence of localized tastes. Ferreira and Waldfogel (2013) examine trade in recorded music, which faces low trade costs, and find that a one percent increase in distance reduces music trade by 0.42 percent. Blum and Goldfarb (2006) show that distance reduces the number of website visits for taste-dependent website categories, such as music, games, pornography or gambling. At the same time, the authors find that distance has no effect on the number of website visits for non-taste dependent website categories, such as software and general, technological or financial information sites.

2.1.2 Other Geographical Factors

Besides distance, another common geographical factor considered in gravity equations is whether two countries share a land border or not. Most researchers find contiguity to

have a positive and significant effect on trade. Head and Mayer (2014) report the average of the estimated coefficient on a dummy variable equal to 1 if two countries share a border to be around 0.5. Other geographical characteristics considered in gravity equations are whether the trading countries are landlocked or islands. Due to the limitations that these geographical characteristics place on transport options, they are generally found to negatively impact trade (see Limao and Venables (2001), Silva and Tenreyro (2006), Helpman et al. (2008)).

2.1.3 Cultural Factors

Cultural closeness between two countries has a positive effect on the amount of trade between them. Head and Mayer (2014) calculate the average estimated coefficient of a dummy variable equalling 1 if two countries speak the same language to be around 0.4. The average estimated coefficient of a dummy variable equalling 1 if two countries share a colonial link is even higher, taking a value of around 0.8. Other cultural characteristics considered by researchers in gravity settings are whether the trading partners have similar religious beliefs or whether their legal systems share the same origin. These variables are generally found to positively and significantly influence trade (see Helpman et al. (2008)).

2.1.4 Economic Factors

Free trade agreements, as well as a common currency, facilitate trade between two countries. In a gravity setting, the effect of a free trade agreement is measured by including a dummy variable that equals 1 if two countries are part of a free trade area. Head and Mayer (2014) report the average estimated coefficient of this dummy variable to be around 0.4. The effect of sharing a currency on trade is usually measured by including a dummy variable equal to 1 if two countries use the same currency. Head and Mayer (2014) calculate the mean of the estimated coefficient on this dummy variable to be around 0.2.

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 et al. (2009), Brynjolfsson et al. (2009) or Goolsbee (2001)).

Very few researchers have used online transaction data to address trade-related questions. As noted in the introduction, Lendle et al. (2016), Lendle and Vézina (2015), Hortaçsu et al. (2009), and Blum and Goldfarb (2006) are the only other studies to estimate worldwide gravity equations on data relating to online activities. However, each of these researchers have faced data limitations, some of which we are able to improve upon.

Lendle et al. (2016) work with eBay data and they have access to the most comprehensive data set out of the four papers mentioned above. The authors perform a gravity analysis focusing on 61 countries and they justify limiting their analysis to 61 countries by noting that 92% of all offline trade flows happen between these 61 states.

By omitting the countries that trade little with each other, however, Lendle et al. (2016) risk underestimating the effects of distance. If distance is part of the reason why some countries trade with each other very little, or do not trade at all, excluding the locations that do not trade much with each other will bias the coefficient on distance downward.

Lendle and Vézina (2015) replicate Lendle et al. (2016)'s analysis on an eBay data set covering sellers located in 21 emerging economies. The sellers export to 204 destinations. The authors find a distance elasticity of -0.31 for eBay trade flows, smaller than the -0.44 elasticity reported in Lendle et al. (2016). Lendle and Vézina (2015)'s study faces limitations due to the fact that the data set covers only 21 exporting countries, and also due to the authors' decision to use a linear estimator, which is known to produce biased results in a gravity setting with zero trade flows.[4]

[4]Zero trade flows are not a problem in Lendle et al. (2016), as all 61 countries trade with each other. In Lendle and Vézina (2015), however, they are, as not all 21 countries export to all 204 destinations.

Hortaçsu et al. (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 only 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.

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 are 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.

Table 1 summarizes the distance elasticities obtained by the studies summarized above. As noted in the introduction, we are able to improve upon the existing estimates, by using a comprehensive data set which covers 146 countries, and by employing a methodology that corrects known econometric biases.

Table 1: Distance Elasticities Obtained by Previous Researchers in a Worldwide Setting

Study	Distance Effect	Country Coverage	Estimator Used	Dependent Variable
Lendle et al. (2016)	-0.440*** (0.032)	61 exporters 61 importers	OLS	Value of sales (logs)
Lendle and Vézina (2015)	-0.313*** (0.048)	21 exporters 204 importers	OLS	Value of sales (logs)
Hortaçsu et al. (2009)	-0.382*** (0.030)	9 exporters 9 importers	OLS	Value of sales (logs)
Blum and Goldfarb (2006)	-1.167*** (0.114)	1 exporter 46 importers	OLS	Number of site visits (logs)

2.3 Overview of Helpman et al. (2008)

Worldwide gravity exercises have been criticized on two main grounds: first, that the estimating equation is not carefully derived from theory and second, that OLS estimates suffer from econometric biases due to ignoring the zero trade flows. Researchers like Silva and Tenreyro (2006) and Eaton et al. (2012) propose econometric methods that accommodate zero trade flows to estimate gravity equations. However, Helpman et al. (2008) is the only paper to address both criticisms outlined above by proposing a theoretical model in which zero trade arises due to an economically meaningful selection process and an empirical methodology that accounts for selection into trade.[5]

Helpman et al. (2008) start from a Melitz (2003) model with asymmetric countries. Consumers face a CES utility function. Each country has a number of firms, and each firm produces a distinctive product. A firm produces a unit of output with a cost-minimizing combination of inputs, $c_j a$. The parameter a is firm specific and it reflects differences across firms in productivity levels. The cost c_j is country specific and it reflects differences across countries in factor prices. A firm from country j can sell its products on the domestic market, or it can export to country i. If the firm wants to export to country i, however, it faces two additional costs: a fixed cost of exporting and a variable cost that is specified in the model as a "melting iceberg" transport cost. Profit maximization makes it obvious that while all firms will sell on the domestic market, exporting will be feasible only for a fraction of firms that have a productivity level a above the threshold required for exports from j to i to be profitable.

Based on the theoretical model, the authors derive the following estimating equation:

$$m_{ij} = \beta_0 + \lambda_j + \chi_i - \gamma d_{ij} + w_{ij} + u_{ij}$$

[5]Besides the theoretical appeal of Helpman et al. (2008)'s proposed estimated method, there is some evidence that it performs better than the PPML estimator proposed by Silva and Tenreyro (2006) for data sets where zero trade values are frequent. See Martin and Pham (2008).

where:

$$m_{ij} = \text{country } j\text{'s exports to country } i$$

$$\lambda_j = \text{fixed effect of the exporting country, } j$$

$$\chi_i = \text{fixed effect of the importing country, } i$$

$$d_{ij} = \text{symmetric distance between } i \text{ and } j$$

$$w_{ij} = \text{fraction of firms (possibly zero) that export from } j \text{ to } i$$

$$u_{ij} = \text{i.i.d. unmeasured trade frictions}$$

Traditional estimates are biased due to two econometric problems. First, omitting the fraction of firms that export from j to i creates an omitted variable bias. This fraction is endogenous; it is a function of the productivity threshold required for exports from j to i to be profitable, which in turn is a function of the variables controlling for the trade barriers between the two countries. Omitting w_{ij} from the estimating equation induces an upward bias in the coefficient measuring the effect of the symmetric distance between i and j, as this coefficient picks up the effect of trade barriers on firm-level trade and on the proportion of exporting firms.

Second, omitting zero trade flows creates a selection bias. Helpman et al. (2008) underline the fact that there are many zero trade flows observed in the real world. In 1997, only about 40% of country pairs traded in both directions. An additional 12% traded in one direction only, and a surprising 48% of country pairs did not trade at all. We ran a similar analysis for our data from Google and found that between 2008-2011, 25% of all country pairs traded in both directions, about 30% traded in one direction only, and 45% did not trade at all. Given the high percentage of zero trade flows in the online world, it is just as important as it is in the traditional, offline world, to correct for the selection bias created by omitting the zero trade flows. Excluding the country pairs with zero trade flows places a downward bias on γ, the trade barrier coefficient.

To correct these biases, Helpman et al. (2008) start by defining a latent variable, Z_{ij}:

$$z_{ij} = \ln Z_{ij} = \gamma_0 + \xi_j + \zeta_i - \gamma d_{ij} - \kappa \phi_{ij} + \eta_{ij}$$

where:

$z_{ij} =$ ratio of variable export profits for the most productive firm to the fixed export costs for exports from j to i

$\xi_j =$ fixed effect of the exporting country, j

$\zeta_i =$ fixed effect of the importing country, i

$d_{ij} =$ symmetric distance between i and j

$\phi_{ij} =$ observed measure of any additional country-pair specific fixed trade costs

$\eta_{ij} = u_{ij} + v_{ij}$, where v_{ij} is IID $\mathrm{N}(0, \sigma_{ij})$

The latent variable z_{ij} is not observed, but positive trade between j and i is observed when $z_{ij} > 0$ and zero trade is observed when $z_{ij} = 0$. Let T_{ij} be an indicator variable equal to 1 when country j exports to i and 0 otherwise, and ρ_{ij} be the probability that j exports to i. We can then specify the Probit equation:

$$\rho_{ij} = \Pr(T_{ij} = 1) = \Pr(z_{ij} > 0) = \Phi(\gamma_0^* + \xi_j^* + \zeta_i^* - \gamma^* d_{ij} - \kappa^* \phi_{ij})$$

If $\hat{\rho}_{ij}$ is the predicted probability of exports from j to i, we can easily obtain $\hat{z}_{ij}^* = \Phi^{-1}(\hat{\rho_{ij}})$, the predicted value of the latent variable z_{ij}^*. Since W_{ij}, the fraction of firms that export from j to i, is a monotonic function of Z_{ij}, we can use $\hat{z}_{ij}^* = \Phi^{-1}(\hat{\rho}_{ij})$ to obtain a consistent estimate for $w_{ij} = \ln W_{ij}$, the omitted variable in the gravity equation specified above. The proposed methodology, thus far, corrects the first of the two biases identified above, namely the omitted variable bias.

To correct the second bias identified above, Helpman et al. (2008) propose to use the standard Heckman (1979) correction for sample selection. For this, they use the same Probit equation specified above to obtain the inverse Mills ratio: $\hat{\eta}_{ij}^* = \phi(\hat{z}_{ij}^*)/\Phi(\hat{z}_{ij}^*)$.

Accounting for this term in the gravity equation addresses the selection bias created by omitting zero trade flows from the estimating equation.

We use the two-step estimation procedure proposed by Helpman et al. (2008) and described above to estimate the effect of distance on online trade flows. Section 3 describes the data we use, while Section 4 outlines the precise empirical specification.

3 Data

3.1 Google Data

3.1.1 Data Source

The data we use come from Google's online advertising platforms and they were generated by the same conversion tracking software that enabled us to perform the US-Canada border effects analysis presented in Chapter 2. The software tracks conversions generated as a result of users clicking on an ad placed on Google or on one of Google's partner sites.

As mentioned in Section 4.1.1 of the previous chapter, the data set has some minor limitations. In particular, the conversion tracking code reliably generates conversion counts (rather than values). Additionally, a count is recorded in the data set only if a user reaches an advertiser's thank you page within 30 days of the initial click on the ad. Finally, while we believe that we are able to attribute most buyers and sellers to the correct region, it is possible that we are attributing a small part of the buyers or sellers to the wrong location.[6]

3.1.2 Coverage

For our worldwide analysis, we use data aggregated across conversion types and sectors of economic activity. The data covers a large number of businesses advertising online.

[6] As a reminder for the reader, for buyers, we use their IP address to determine their location, while for sellers, we use their self-reported mailing address.

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 a year prior to when our data coverage starts, there were already more than one million businesses opting for Google's advertising platform. In 2007, Google also announced publicly that no client accounts for more than 10% of the company's revenues, indicating that its advertising program is not dominated by one large business.

The data set we use covers all conversion counts recorded over a period of four years, from 2008 to 2011. Similar to Lendle et al. (2016), we average the observations over the four years of data that we have at our disposal.

The overall number of recorded conversion counts is in excess of 10 billion.[7] All but six countries are featured in our data set. At the time the data were collected, trade embargoes were in place between the United States and Burma, Cuba, Iran, North Korea, Sudan, and Syria. As a US-based company, Google could not generate revenues in these countries and therefore, it could not accept ads from businesses located in any of these six countries.

3.1.3 Comparability to Estimates Obtained by Helpman et al. (2008)

At the worldwide level, we are unable to obtain online trade data broken down by sector of economic activity. While we do not know the precise composition of trade flows on Google, there are likely to be important differences between the bundle of goods and services reflected in our dataset and the bundle of products reflected in the data used by Helpman et al. (2008).

For their analysis, Helpman et al. (2008) use the bilateral trade flows reported in Feenstra's "World Trade Flows" database. This database records trade figures for 34 manufacturing industry codes and covers transactions in intermediate inputs such as primary metals and industrial chemicals. By contrast, we use data from Google, which is an all-purpose search engine that promotes a wide variety of online goods and services pri-

[7] Due to confidentiality reasons, we cannot provide precise figures.

marily targeted to end users. As such, our data are more likely to reflect transactions in consumer goods (such as books, antiques, or consumer electronics) than transactions in intermediate inputs. Furthermore, our data are likely to include transactions classified under the service sectors, such as legal advice, web-hosting, banking or web design, which are not covered by the data used by Helpman et al. (2008). Finally, our dataset is also likely to include sales of digital goods, such as software or music downloads, which are not reflected in the data used by Helpman et al. (2008).

We reproduce the results that Helpman et al. (2008) obtain in the Appendix. However, due to fact that the bundle of goods and services reflected in our dataset differs from the bundle of products reflected in the data used by Helpman et al. (2008), we cannot interpret the differences between the coefficients estimated by us and the coefficients estimated by Helpman et al. (2008) as providing a direct and precise comparison between trade barriers in the online world versus the offline world.

3.2 Explanatory Variables

We include in our regressions the same explanatory variables that Helpman et al. (2008) used in their paper. For each variable, we will discuss the data source we use, as well as our a priori hypotheses.

3.2.1 Distance

In our study, we are primarily interested in uncovering the empirical relationship between distance and online trade. We use the CEPII distances database as our data source for the distance measure, $distance_{ij}$.[8] To measure distance, we follow Head and Mayer (2009)'s suggestion to calculate the distance between two countries based on the bilateral distances between the biggest cities of those two countries and weighting those inter-city distances by the share of the city in the overall country's population.[9]

We expect distance to have a negative effect on both the extensive and the intensive

[8] This database is available online at `http://www.cepii.fr/CEPII/fr/bdd_modele/presentation.asp`

[9] We also experiment with using simple (unweighted) distance measures and we find that they produce similar results.

margins of trade. In other words, we expect distance to have a negative effect both on the probability of two countries trading and on the volume of trade between them.

As far as the extensive margin is concerned, we expect to find a negative effect because a business wanting to export online will face some fixed costs. Evidently, a firm wanting to export online goes through a process that is different from that undertaken by a firm looking to export offline. On Google's platforms, a US company wanting to export to France can do so instantaneously, by ticking a box in its Google account and making its online ads visible within France. In theory, therefore, the US company wanting to export to France faces zero fixed costs. In practice, however, it is hard to imagine a profit maximizing firm making the decision to export without undertaking some initial investment to make its product appealing and accessible to foreign consumers. Before exporting to France, a US online retailer might want to translate its website and marketing materials into French. It might also want to ensure that it can process payments made with French credit cards, or that it can safely ship and store its products once they arrive in France. The costs outlined here are small compared to the investment a traditional firm needs to undertake to export offline, but they are unlikely to be zero.

As for the intensive margin, we expect to find that distance has a negative effect on the volume of online trade for several reasons. First, the physical goods sold through Google incur transport costs similar to those incurred by the manufacturing goods sold through traditional trade channels. This is because shipping physical goods to far away places – regardless of whether the order was placed online or offline – is costly and can take a long time. Second, for some services exchanged over the Internet, the convenience of working with a provider that is close by will make transactions between online shoppers and service providers less likely over long distances. Finally, the existence of localized tastes is likely to affect online exchanges much like it affects the volume of transactions conducted through the traditional trade channels.

Although we expect distance to have a negative effect on the intensive margin of online trade, this effect is likely to be smaller online. There are several reasons for this. First, search costs are lower online. On Google, buyers and sellers find each other with a simple Internet search, the cost of which is low and uncorrelated with the distance

between the transacting parties. Second, transport costs are zero for some of the transactions conducted online. As noted in section 3.1.3, Google is a generic, all-purpose search engine and the transactions reflected in our dataset cover a wide variety of digital goods and services, in addition to physical goods. For some of the digital goods and services reflected in our dataset, the transport costs are likely to be zero or close to zero. Finally, trust might have a smaller effect on online trade flows. Internet users and online businesses can usually rate and review each other online, which can act as a deterrent to dishonest trades, whether the transacting parties are nearby or far away from each other.

3.2.2 Other Geographic Characteristics

Similar to Helpman et al. (2008), we control for three other geographical characteristics besides distance: contiguity, as well as whether the trading partners are islands or landlocked countries. To do so, we build three dummy variables. The first one, $land\ border_{ij}$, takes the value of 1 if two countries share a land border and 0 otherwise. We use the CEPII distances database to construct this dummy. The second one, $island_{ij}$, takes the value of 1 if two countries are islands and 0 otherwise. We use the CIA's *World Factbook* land boundary measure to construct this variable.[10] In our study, we consider an island to be a country whose land boundary is reported by the CIA to measure 0 kilometers. Finally, the third dummy variable, $landlock_{ij}$, takes the value of 1 if two trading partners have no coastline and 0 otherwise. The CEPII distances database provides the data necessary to construct this third dummy.

In line with other studies using offline trade data, we expect contiguity to have a positive effect on online trade. A priori, it is difficult to predict whether the fact that two trading partners are landlocked or islands will have a positive or negative effect on online trade. In studies conducted using traditional trade figures, these geographical characteristics are generally found to negatively impact trade due to the limitations they place on transport options. With online trade, we might see a positive effect precisely because of these limitations: if the inhabitants of a remote island do not find in stores the foreign

[10]The CIA's *World Factbook* is available online at `https://www.cia.gov/library/publications/the-world-factbook/`

products that they want because it is not profitable for offline retailers to transport them, they might purchase these products online. Geographical isolation, in this case, would encourage, rather than discourage, online trade.

3.2.3 Cultural Characteristics

Four variables allow us to control for the cultural closeness between two countries. The first one is $legal_{ij}$, a dummy variable taking the value of 1 if two countries' legal system share the same origin and 0 otherwise. According to the CIA's *World Factbook*, the data source we use to construct this dummy variable, each country's legal system has, at its origins, one of five elements: civil law, common law, customary law, pluralistic law or religious law. If two countries' legal system share the same element, we assign a value of 1 to $legal_{ij}$. In line with studies conducting using traditional trade data, we expect to find a positive and significant effect of $legal_{ij}$ on online trade.

The second variable that allows us to control for cultural closeness is $language_{ij}$. This is a dummy variable taking the value of 1 if two countries share an official language and 0 otherwise. We use the CEPII distances database to construct this dummy variable. Our prior is that countries that speak the same language will trade more with each other in the online world. Translating the website and the marketing materials constitutes a large part of the fixed cost incurred by a company wanting to export online. After it starts exporting, the company will likely communicate with its customers via email or over the phone about the products offered or the details of the orders placed online. Speaking the same language will facilitate that communication. Our prior is that a common language, by lowering both the fixed and the variable cost of exporting, will have a positive influence on online trade.

The third variable is $colony_{ij}$, a dummy variable that takes the value of 1 if two countries were ever in a colonial relationship and 0 otherwise. Our data source for this variable is, once again, the CEPII distances database. Since countries that share a common colonial history are culturally closer, we expect to find a positive effect of $colony_{ij}$ on online trade.

Finally, the fourth variable we use to control for cultural similarity is $religion_{ij}$. Similar to Helpman et al. (2008), we use the following formula to generate this variable: $religion_{ij} = (\%\ Protestants_i \times \%\ Protestants_j) + (\%\ Catholics_i \times \%\ Catholics_j) + (\%\ Muslims_i \times \%\ Muslims_j)$. We obtain the percentage figures from CIA's *World Factbook*. A priori, we expect countries with similar religious beliefs to trade more with each other.

3.2.4 Economic Characteristics

Two variables measure the extent of the economic ties between two trading partners. The first one is $currency_{ij}$, a dummy variable taking the value of 1 it two countries use the same currency and 0 otherwise. We use the CIA's *World Factbook* to construct this variable. The second one is FTA_{ij}, a dummy variable taking the value of 1 if two countries belong to a free trade agreement and 0 otherwise. To construct this variable, we use the World Trade Organization's Regional Trade Agreements Information System,[11] which contains a list of all regional trade agreements in place, along with their date of entry into force. As common currencies and free trade agreements encourage trade between two countries, we expect these two variables to have a positive and significant effect on online trade.

3.2.5 Exclusion Restriction

An exclusion restriction is needed for the first stage of the two-step estimation procedure proposed by Helpman et al. (2008). The authors' theoretical model suggests that a valid exclusion restriction is one that affects the fixed cost of exporting, so it enters with a non-zero coefficient in the first stage Probit equation, but has no effect on the variable cost of exporting.

Helpman et al. (2008) propose two indicators variables as exclusion restrictions, the first of which takes the value of 1 when the cost of setting up a business in both countries is above the median, while the second takes the value of 1 when the number of days and procedures required to set up a business in both countries is above the median. Although

[11] Available online at `http://rtais.wto.org/ui/PublicMaintainRTAHome.aspx`

these variables affect the fixed cost of exporting in the offline trade world, they do not necessarily affect the fixed cost of exporting online. Indeed, when we test these two indicator variables on our data set, we find that they do not have enough explanatory power in the first stage estimation to act as valid exclusion restrictions. The reason for this is simple: in the virtual world, in order to start exporting, a business does not need to set up an entity abroad.

In e-commerce, warehouses, rather than business entities, are a key component of international expansion. Online retailers often store inventory in their destination markets,[12] so securing warehouse space is one of the first steps they make when expanding abroad. There are two reasons why online retailers keep inventory in warehouses overseas. First, it speeds up the delivery of the goods to the final consumers: once a user from the destination market places an order, the good is delivered to them directly from the warehouse. Second, it makes returns and exchanges easier for the buyers.

A reasonable exclusion restriction for our analysis, therefore, would capture the costs of setting up a warehouse, rather than the cost of setting up a business. The World Bank's *World Development Indicators* database contains data on the number of procedures required to build a warehouse.[13] We use this data to build an indicator variable taking the value of 1 when the number of procedures required to build a warehouse in both countries is above the median and 0 otherwise. We find that this variable does very well as an exclusion restriction for our data set and satisfies the requirement of influencing the fixed, rather than the variable, cost of exporting online.

Table 2 provides summary statistics for all the explanatory variables described above.

[12] A recent eBay report, *Overseas Warehousing: New Trend in Cross-Border Commerce*, noted that over a quarter of Hong Kong's eBay sellers with sales over $10,000 store their goods in overseas warehouses.

[13] Data on the number of procedures to build a warehouse were not available for the following countries and territories: Andorra, Aruba, Barbados, Curacao, Holy Sea, Hong Kong, Kosovo, Libya, Liechtenstein, Macau, Malta, Monaco, Montenegro, Nauru, Netherlands Antilles, Palestine, San Marino, Sint Maarten, Somalia, Taiwan, Turkmenistan, and Tuvalu.

Table 2: Summary Statistics

Variable	Nr. Obs	Mean	Std. Dev	Min	Max
$distance_{ij}$	19,425	8,121	4,524	33	19,650
$land\ border_{ij}$	19,425	0.016	0.124	0	1
$island_{ij}$	19,425	0.043	0.203	0	1
$landlock_{ij}$	19,425	0.042	0.201	0	1
$legal_{ij}$	19,425	0.370	0.483	0	1
$language_{ij}$	19,425	0.162	0.368	0	1
$colony_{ij}$	19,425	0.007	0.081	0	1
$religion_{ij}$	19,425	0.169	0.223	0	1
$currency_{ij}$	19,425	0.013	0.112	0	1
FTA_{ij}	19,425	0.058	0.234	0	1
$exclusion\ restriction_{ij}$	19,425	0.244	0.429	0	1

4 Empirical Approach

To estimate the effect of distance using online trade flows, we use the two-step estimation procedure proposed by Helpman et al. (2008). The first stage focuses on the extensive margin of online trade, and it is a Probit regression that yields a prediction for the probability that country j exports to country i. The second stage focuses on the intensive margin, and it is a gravity equation estimated on the subset of positive trade flows, which examines the volume of exports from j to i, conditional on exporting.

The Probit equation takes the following form:

$$\begin{aligned}\rho_{ij} = \beta_0 + \beta_1 \ln(distance)_{ij} + \beta_2\ land\ border_{ij} + \beta_3\ island_{ij} + \beta_4\ landlock_{ij} + \\ \beta_5\ legal_{ij} + \beta_6\ language_{ij} + \beta_7\ colony_{ij} + \beta_8\ religion_{ij} + \beta_9\ currency_{ij} \\ \beta_{10}\ FTA_{ij} + \beta_{11}\ exclusion\ restriction_{ij} + \lambda_i + \chi_j + \varepsilon_{ij}\end{aligned}$$

where:

ρ_{ij} = 1 if the number of conversions that users in country i generate for businesses located in country j is positive, 0 otherwise.

$distance_{ij}$ = population weighted distance, in km, between the largest cities of i and j.

$land\ border_{ij}$ = 1 if two countries share a land border, 0 otherwise.

$island_{ij}$ = 1 if both countries are islands, 0 otherwise.

$landlock_{ij}$ = 1 if both countries have no coastline, 0 otherwise.

$legal_{ij}$ = 1 if both countries' legal systems share the same origin, 0 otherwise.

$language_{ij}$ = 1 if both countries share an official language, 0 otherwise.

$colony_{ij}$ = 1 if the two countries were in a colonial relationship, 0 otherwise.

$religion_{ij} = (\%Protestants_i \times \%Protestants_j) + (Catholics_i \times Catholics_j) + (\%Muslims_i \times \%Muslims_j)$

$currency_{ij}$ = 1 if both countries use the same currency, 0 otherwise.

FTA_{ij} = 1 if both countries belong to a free trade agreement, 0 otherwise.

$exclusion\ restriction_{ij}$ = 1 if the number of procedures required to build a warehouse is above the median in both countries, 0 otherwise.

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

We estimate the first stage Probit regression above in order to obtain a predicted probability of export, $\hat{\rho}_{ij}$, which in turn allows us to obtain consistent estimates for $\hat{\bar{w}}^*_{ij}$ (the fraction of exporting firms) and $\hat{\bar{\eta}}^*_{ij}$ (inverse Mills ratio). These estimates are needed as additional controls in the second stage estimation, as they correct the omitted variable and sample selection biases (see Section 2.3).

Helpman et al. (2008) propose three different ways to estimate the second stage regression. The dependent variable in the second stage estimation is, invariably, m_{ij}, which represents the exports from j to i. For us, m_{ij} is the number of conversion counts that users in country i generate for businesses located in country j. Since the coefficient associated with $\hat{\bar{w}}^*_{ij}$ does not enter the model linearly, an obvious first candidate for the second stage estimation is nonlinear least squares (NLS). A second candidate for the second stage estimation is an OLS regression, which includes among the independent variables the usual gravity controls, as well as $\hat{\bar{\eta}}^*_{ij}$ and a polynomial in $\hat{\bar{z}}^*_{ij}$ (predicted values of the latent variable; see Section 2.3). The authors obtain this estimating equation by relaxing the assumption governing the distribution of firm heterogeneity, which eliminates the nonlinearity induced by $\hat{\bar{w}}^*_{ij}$, and the need for a nonlinear estimator. Finally, the third candidate for the second stage estimation is a non-parametric functional form that

emerges as a possibility after relaxing the joint normality assumption for the unobserved trade costs. Although these three estimation methods rely on different assumptions, similar to Helpman et al. (2008), we find that they yield very similar results.

5 Results

5.1 Main Results

Table 3 reports the chapter's main results. Columns (1) and (2) underline the importance of expanding the country coverage in worldwide gravity analyses done on online data. In particular, in column (1), we use a simple linear estimator and report the results we obtain when, similar to Lendle et al. (2016), we limit the analysis to the 61 countries they covered. In column (2), we present the results we obtain when we expand the data set to 146 countries. This more than doubles the number of countries considered in the Lendle et al. (2016) paper.

We find that limiting the analysis to the 61 countries with high levels of trade between them biases the coefficients of the variables intended to capture the closeness between trading partners downwards. Remoteness, be it geographic, cultural or economic, discourages trade. By excluding from a data set the country pairs that trade little with each other, one excludes the observations for which remoteness might matter the most. A comparison of the results generated using the smaller sample size of 61 countries and those generated using the larger sample size of 146 countries confirms this hypothesis.

With the increase in sample size, the elasticity of distance increases from -0.555 to -0.610. The coefficients of the other variables that measure cultural or economic closeness increase, as well. In particular, the coefficient of $language_{ij}$, $religion_{ij}$ and FTA_{ij} go up. All increases are statistically significant and point to the importance of expanding the country coverage of worldwide analyses beyond the countries that trade the most.

Table 3: Worldwide Gravity Results – Google Data

	OLS	OLS	OLS	Probit	NLS	Poly-nomial	Non-Para-metric	PPML
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$distance_{ij}$	-0.555***	-0.610***	-0.610***	-0.139***	-0.509***	-0.528***	-0.533***	-0.525***
	(0.044)	(0.029)	(0.029)	(0.011)	(0.034)	(0.038)	(0.039)	(0.137)
$land\ border_{ij}$	0.923***	0.743***	0.743***	0.221***	0.737***	0.786***	0.776***	0.510*
	(0.144)	(0.142)	(0.142)	(0.043)	(0.138)	(0.142)	(0.143)	(0.304)
$island_{ij}$	0.399***	-0.029	-0.029	-0.114***	-0.020	-0.033	0.031	-1.418***
	(0.130)	(0.104)	(0.104)	(0.035)	(0.101)	(0.103)	(0.104)	(0.409)
$landlock_{ij}$	0.007	0.284**	0.284**	-0.005	0.221*	0.208*	0.213*	1.621***
	(0.152)	(0.115)	(0.115)	(0.041)	(0.114)	(0.116)	(0.117)	(0.416)
$legal_{ij}$	0.127***	0.138***	0.139***	0.018	0.110***	0.115***	0.116***	0.150
	(0.049)	(0.038)	(0.038)	(0.014)	(0.037)	(0.037)	(0.037)	(0.179)
$language_{ij}$	0.635***	1.192***	1.192***	0.316***	1.031***	1.072***	1.070***	1.172***
	(0.108)	(0.068)	(0.068)	(0.017)	(0.086)	(0.096)	(0.097)	(0.228)
$colony_{ij}$	0.545***	0.450**	0.450**	0.017	0.451**	0.457**	0.450**	0.518
	(0.180)	(0.202)	(0.202)	(0.098)	(0.192)	(0.196)	(0.198)	(0.422)
$religion_{ij}$	0.525***	1.051***	1.051***	0.071**	1.036***	1.038***	1.034***	0.606
	(0.136)	(0.084)	(0.084)	(0.031)	(0.081)	(0.083)	(0.083)	(0.417)
$currency_{ij}$	0.104	-0.009	-0.007	0.121**	0.099	0.122	0.101	0.216
	(0.092)	(0.151)	(0.151)	(0.056)	(0.153)	(0.156)	(0.157)	(0.330)
FTA_{ij}	0.328***	0.779***	0.779***	0.206***	0.558***	0.622***	0.615***	1.214***
	(0.076)	(0.079)	(0.079)	(0.031)	(0.079)	(0.084)	(0.085)	(0.216)
$excl\ restriction_{ij}$			-0.009	-0.052***				
			(0.061)	(0.023)				
$\hat{\omega}^*_{ij}$ (predicted fraction of exporters)					0.302***			
					(0.091)			
$\hat{\eta}^*_{ij}$ (inverse Mills ratio)					1.719***	1.720***		
					(0.081)	(0.135)		
$\hat{z}^*_{ij}$ (predicted latent variable)						1.363***		
						(0.361)		
$\hat{z}^{*2}_{ij}$						-0.182*		
						(0.111)		
$\hat{z}^{*3}_{ij}$						0.008		
						(0.011)		
Obs.	3,593	10,120	10,120	19,425	10,120	10,120	10,120	19,425
R^2	0.907	0.756	0.756	0.660		0.780	0.782	0.698

Notes: Exporter and importer fixed effects. The dependent variable is number of conversion counts in levels for the PPML estimation, and in logs for the other estimations. Only 61 countries considered in column (1); 146 countries considered in all other columns. Marginal effects at sample means and pseudo R^2 reported for Probit. Bootstrapped standard errors for NLS; robust standard errors (clustering by country pair) elsewhere.

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

Column (3) in Table 3 confirms the validity of the exclusion restriction used in the first stage regression. As noted in Section 3.2.5, a valid exclusion restriction should have no impact on the amount of trade between two countries, but it should have a statistically significant effect on the probability of two countries trading. The coefficients on our exclusion restriction reported in columns (3) and (4) of Table 3 confirm that it is exactly the type of variable that we need.

Column (4) of Table 3 reports the results of the Probit regression. This is the first stage estimation and it gives insight into the extensive margin of trade. The estimated coefficients reflect the effect that the independent variables have on the probability of two countries trading. Columns (5)-(7) of Table 3 report the results obtained in the second stage regressions for each of the estimators discussed in Section 4: nonlinear least squares, OLS with additional controls for $\hat{\eta}_{ij}^{*}$ and a polynomial in $\hat{z}_{ij}^{*}$, and finally, a non-parametric estimator. In columns (5)-(7), the estimated coefficients reflect the effect of the independent variables on the volume of trade between two countries, conditional on exporting. The results in these columns give us insight into the intensive margin of trade. We discuss the estimated coefficients below.

5.1.1 Distance

Distance has a negative and significant influence on both the extensive and the intensive margins of trade. This confirms our a priori belief that distance is not dead, even online. The second stage estimation results indicate that conditional on exporting, a 10% increase in the distance between two countries reduces the volume of trade between them by approximately 5.3%. Based on these results, we can safely conclude that among Google's advertising clients, distance lives on.

Besides uncovering the role of distance for online trade, the results reported in Table 3 also underline the importance of accounting for selection into trade when estimating worldwide gravity equations using online trade data. Using OLS biases the coefficient on distance upwards. The benchmark results, reported in column (2) of Table 3, show the elasticity of distance to be around -0.61. The magnitude of this coefficient is between

13% and 17% higher than the magnitude of the coefficients reported in columns (5)-(7) of Table 3, which are estimated using the methodology suggested by Helpman et al. (2008).

5.1.2 Other Geographic Characteristics

Our results show that besides distance, sharing a land border also significantly affects the volume of trade between two countries. The magnitude of the coefficient of the contiguity dummy in columns (5)-(7) is large: it predicts that conditional on exporting, online trade between two countries that share a land border is close to 120% larger than trade between countries that do not share a border. Contiguity also affects the extensive margin on trade. Column (4) of Table 3 shows that the coefficient of the $land\ border_{ij}$ dummy variable has a positive and significant impact on the probability of two countries trading online.

Countries that are islands or landlocked, characteristics which several studies done on data from the offline world found to negatively impact trade, no longer decrease the volume of trade in the virtual world. In columns (5)-(7), the coefficient of the dummy variable equalling 1 if both countries are landlocked is positive and statistically significant at the 10% level, while the coefficient on the dummy variable equalling 1 if both countries are islands is insignificant. We find some indication that the probability of exporting is negatively affected by both trading partners being islands. The coefficient on the $island_{ij}$ dummy variable, reported in column (4) is negative and significant.

5.1.3 Language

Our prior regarding the importance of language to online trade was correct. A good proportion of the fixed cost that a company faces when exporting online comes from translating its website and marketing materials. In column (4) of Table 3, we can see that the coefficient on the dummy variable measuring the effect of two countries sharing a language has a large, positive, and statistically significant effect on the probability of two countries trading. Once it decides to export, a company needs to provide customer support services to its foreign clients. In columns (5)-(7), we can see that sharing an

official language also has a large and positive effect on the intensive margin of trade. Conditional on exporting, the volume of trade between countries that share an official language is more than 180% higher than the volume of trade between countries that do not speak the same language. These results underline the importance of customers and businesses sharing an official language and interacting via a website written in a language that they can both understand.

5.1.4 Other Cultural Characteristics

Besides language, other cultural affinities, such as similarity in religious beliefs, colonial ties and common legal systems also have a positive influence on the volume of online trade. The index measuring similarity in religious beliefs, in particular, has a positive and statistically significant effect, both on the extensive and the intensive margins of online trade. The dummy variables measuring the effect of two countries having similar legal systems or colonial ties have a positive and statistically significant effect only on the intensive margin on trade. Conditional on exporting, the volume of trade between countries that have similar legal systems is 12% larger than the volume of trade between countries that do not. For countries that have colonial ties, the volume of trade conditional on exporting is 57% higher than for countries that do not.

5.1.5 Economic Characteristics

Belonging to a free trade agreement has a positive and statistically significant effect on the probability of countries trading and on the volume of online exports between them once the decision to export is made. All else being equal, the quantity of online trade, conditional on exporting, is around 80% higher for countries that belong to a free trade agreement than for countries that do not.

Using a common currency influences the extensive, but not the intensive margin of online trade. The coefficient of the dummy variable $currency_{ij}$, which takes the value of 1 if two countries have a common currency and 0 otherwise, is positive and significant in column (4) of Table 3, indicating that the indicator influences the probability of countries

trading with each other.

This is a sensible result. Before exporting to a new country, an online company incurs costs in order to ensure that it can provide price information and accept electronic payments in the currency of the importing country. Assume a US software developer would like to start promoting his software programs to UK-based users. Currently, he only sells his programs within the United States. All prices on his website are expressed in US dollars and he can only accept credit card payments in US dollars.

Before he starts to advertise his products in the UK, he makes changes to his website in order to ensure that his UK-based users see price figures in pounds. For example, he might implement some code on his website that uses the information contained in his website visitors' IP addresses to determine their country.[14] If the code determines that the visitors are US-based, the prices on the site are expressed in US dollars. If the code determines that the visitors are UK-based, the prices on the site are expressed in pounds. These website changes are costly not only in terms of developing the code to determine users' country of residence, but also in terms of converting all prices from US dollars to pounds.[15]

Besides the changes he makes to his website, the American software developer also needs to ensure that he can accept credit card payments in British pounds. For example, he might need to set up a new merchant account or to switch from a single-currency to a multiple-currency merchant account.[16] Depending on the provider, opening a new merchant account or adding a currency to an existing merchant account might incur setup fees.

The example of the software developer illustrates the fixed costs that a business might incur in order to sell its products in a country that uses a different currency. The results reported in column (4) of Table 3 confirm the fact that $currency_{ij}$ influences the fixed cost

[14] Alternatively, he can add a drop-down menu to his website allowing users to choose their preferred country.

[15] The seller can program his website so that the US dollar to pound price conversion is done automatically, using up-to-date exchange rate information, or he can be perform the price conversion manually, by entering a 'fixed' pound price for each product.

[16] A merchant account is a bank account that enables companies, like our software developer, to accept and process debit or credit card payments.

of exporting and has a positive and statistically significant effect on the extensive margin of trade.

Once the initial setup is complete, $currency_{ij}$ is unlikely to influence the volume of online trade between two countries. In the example above, once the US-based software developer undertakes the necessary investment to make changes to his webpage, his site automatically – and costlessly – displays prices in British pounds to UK-based users. Furthermore, the new setup of his merchant account allows him to process credit card payments made in British pounds without him incurring additional costs. Due to the fact that following the initial setup, a company does not need to incur additional costs to sell to countries using different currencies, sharing a common currency should not impact the volume of trade between two locations. Our results in columns (5)-(7) of Table 3 confirm this hypothesis. The coefficient on the dummy variable $currency_{ij}$ is insignificant.

5.1.6 PPML Results

Column (8) of Table 3 reports the results we obtain when we use the PPML estimator proposed by Silva and Tenreyro (2006). Econometrically, the PPML estimator is an improvement over the linear estimator we use to generate the results we report in column (2). This is because the PPML estimator, unlike the OLS estimator, is robust to patterns of heteroskedasticity that are present in trade data. Furthermore, the PPML estimator allows for the inclusion of zero-valued trade flows in the estimation.

However, it is less clear as to whether the PPML estimator is also an improvement over the methodology proposed by Helpman et al. (2008), which we use to generate the results presented in columns (4)-(7) of Table 3. One advantage of the PPML estimator is that it directly addresses heteroskedasticity, while the methodology proposed by Helpman et al. (2008) does not. However, the PPML estimator ignores the econometric biases arising due to sample selection, while the methodology proposed by Helpman et al. (2008) does not. For datasets like ours, where approximately 30% of country pairs trade in one direction only, and 45% do not trade at all – in other words, with significant

selection into trade – Martin and Pham (2008) show that the use of the PPML estimator leads to larger biases than the use of methodologies that control for sample selection.

Several important differences arise when comparing the estimates we obtain when using a linear estimator in column (2) and the PPML estimator in column (8). First, the impact of distance on online trade is slightly larger under OLS. The estimated elasticity of distance is -0.610 when we use OLS and -0.525 when we use PPML. Although slightly smaller, the elasticity of distance we obtain by using PPML confirms, once more, that distance matters for worldwide online trade flows.

Second, other geographical characteristics, such as both trading countries being islands or landlocked, appear to play a larger role for online trade flows under PPML. The coefficient on $island_{ij}$, the dummy variable that takes the value of 1 when both countries are islands, is insignificant in column (2), yet it is large, negative and significant in column (8).[17] Similarly, the coefficient on $land\ border_{ij}$, the dummy variable that takes the value of 1 when both countries are landlocked, is significant only at the 5% level when estimated by OLS and takes the value of 0.284, while when estimated by PPML, it is significant at the 1% level and takes the value of 1.621.

Third, with the exception of language, all other variables that capture cultural closeness play an insignificant role when estimated using Poisson. In column (2), where we report the estimates we obtain when using OLS, we find that the coefficients on $legal_{ij}$, $religion_{ij}$, and $colony_{ij}$, variables which control for similarities in legal systems, religious beliefs, and the existence of colonial ties, are positive and significant. In column (8), however, where we report the estimates we obtain when using PPML, none of these coefficients is statistically significant at the conventional levels.[18]

It is important to note, however, that the PPML estimator confirms our intuition and prior results that sharing an official language in the virtual world, where customers and businesses interact via a website, matters greatly for online trade. The coefficient on

[17] When we use the methodology proposed by Helpman et al. (2008), we find that $island_{ij}$ has a negative and significant effect on the probability of two countries trading, but it does not have a significant effect on the volume of trade between them.

[18] When examining offline trade flows, Silva and Tenreyro (2006) also find that the precision of the coefficients preceding variables that measure cultural closeness disappears when moving from the OLS to the PPML estimator.

$language_{ij}$, the dummy variable that takes the value of 1 when two countries share an official language, is significant at the 1% level when estimated using PPML and it takes the value of 1.172. This implies that all else being equal, online trade between countries that share an official language is 3.2 times higher than online trade between countries that do not have a common official language.

Fourth, in terms of economic characteristics, the PPML estimator obtains a higher estimate than the OLS estimator for the coefficient of FTA_{ij}, the dummy variable that takes the value of 1 if two countries belong to a free trade agreement.[19] Importantly for us, the PPML estimator also returns an insignificant coefficient for $currency_{ij}$, the dummy variable that takes the value of 1 if two countries have a common currency. In the next section, we show that this variable can be used as an alternative exclusion restriction when using the methodology proposed by Helpman et al. (2008).

5.2 Alternative Exclusion Restriction

Given the importance of accounting for selection into trade when assessing the role of distance for online trade, we are interested in suggesting other potential exclusion restrictions that researchers can use in their analyses. The results presented in Section 5.1 reveal another solid candidate for an identifying restriction, namely $currency_{ij}$.

As we have seen, a common currency affects the fixed – and not the variable – cost associated with exporting online. Furthermore, this variable has the added advantage of being readily available at the worldwide level and easy to generate: it simply takes the value of 1 if two countries share a common currency.

In Table 4, we replicate the analysis we used to generate the results for Table 3, but we drop the indicator variable signaling whether the number of procedures required to build a warehouse is above the mean in both countries and we instead use $currency_{ij}$ as our exclusion restriction.

The results we obtain in Table 4 are very similar to those we report in Table 3, thus con-

[19] This is, once again, similar to Silva and Tenreyro (2006), who find that the PPML estimator identifies a larger role for free trade areas than OLS.

firming the appropriateness of the common currency dummy as an alternative exclusion restriction.

Table 4: Results Using an Alternative Exclusion Restriction – Google Data

	OLS	Probit	NLS	Polynomial	Non-Param.
	(1)	(2)	(3)	(4)	(5)
$distance_{ij}$	-0.610***	-0.139***	-0.510***	-0.531***	-0.529***
	(0.029)	(0.011)	(0.034)	(0.038)	(0.039)
$land\ border_{ij}$	0.743***	0.222***	0.738***	0.788***	0.762***
	(0.142)	(0.043)	(0.138)	(0.142)	(0.143)
$island_{ij}$	-0.029	-0.114***	-0.015	-0.027	-0.026
	(0.104)	(0.035)	(0.101)	(0.103)	(0.103)
$landlock_{ij}$	0.284**	-0.004	0.219*	0.206*	0.202*
	(0.115)	(0.041)	(0.114)	(0.116)	(0.116)
$legal_{ij}$	0.138***	0.018	0.110***	0.115***	0.116***
	(0.038)	(0.014)	(0.037)	(0.037)	(0.037)
$language_{ij}$	1.192***	0.316***	1.030***	1.073***	1.061***
	(0.068)	(0.017)	(0.087)	(0.097)	(0.098)
$colony_{ij}$	0.450**	0.016	0.452**	0.457**	0.460**
	(0.202)	(0.098)	(0.192)	(0.196)	(0.197)
$religion_{ij}$	1.051***	0.072**	1.035***	1.038***	1.033***
	(0.084)	(0.031)	(0.081)	(0.083)	(0.083)
FTA_{ij}	0.779***	0.205***	0.567***	0.635***	0.611***
	(0.079)	(0.031)	(0.078)	(0.084)	(0.084)
$currency_{ij}$	-0.009	0.115**			
	(0.151)	(0.056)			
$\hat{\omega}^*_{ij}$ (predicted fraction of exporters			0.303***		
			(0.092)		
$\hat{\eta}^*_{ij}$ (inverse Mills ratio			1.719***	1.709***	
			(0.082)	(0.136)	
$\hat{z}^*_{ij}$ (predicted latent variable)				1.321***	
				(0.363)	
$\hat{z}^{*2}_{ij}$				-0.169	
				(0.112)	
$\hat{z}^{*3}_{ij}$				0.006	
				(0.011)	
Obs.	10,120	19,425	10,120	10,120	10,120
R^2	0.756	0.660		0.780	0.782

Notes: Exporter and importer fixed effects. The dependent variable is the number of conversion counts in logs. Common currency is the excluded variable. Marginal effects at sample means and pseudo R^2 reported for Probit. Bootstrapped standard errors for NLS; robust standard errors (clustering by country pair) elsewhere.
*** Significant at 1%, ** significant at 5%, * significant at 10%

6 Conclusion

In this study, we challenge the popular view that distance is dead in the virtual world. To do so, we use a proprietary data set from Google to perform the most extensive online worldwide gravity analysis to date. We find that conditional on exporting, a 10% increase in the distance between two countries lowers the volume of online trade between them by 5.3%. We conclude that distance continues to negatively affect trade, even for transactions done over the Internet.

Besides the negative effect of distance on online trade, our study also uncovers a few other interesting facts relating to online trade. We find that certain cultural affinities matter greatly in the virtual world. A common official language, in particular, has a large effect on the volume of online trade between two countries. We also find that economic ties, such as a common currency, which facilitate trade in the offline world, have an insignificant effect on the intensive margin of trade in the virtual world. Our results indicate that the economic policies that encourage trade in the offline world might not be as successful at increasing the number of cross-border transactions online.

Our study also points to the importance of accounting for selection into trade when performing gravity analyses at the worldwide level. To aid future research, we identify two variables that can be used as exclusion restrictions in a selection model designed to uncover trade patters in the online world. The first variable measures the difficulty of building a warehouse, while the second one captures the benefits of sharing a common currency in a virtual environment.

In this study, we answer the important question of how much distance matters online. However, our estimates do not provide a precise and exact point of comparison between the role of distance online versus offline. Our conversion count measure does not have a direct equivalent in the offline world. Also, the composition of online and offline trade flows differs. In our future work, as more detailed data regarding the composition of the products traded online become available, we hope to be able to bring more preciseness to the online-offline comparison.

References

Allen, Treb, "Information frictions in trade," *Econometrica*, 2014, *82* (6), 2041–2083.

Anderson, James E, "The gravity model," *Annual Review of Economics*, 2011, *3* (1), 133–160.

_ **and Eric van Wincoop**, "Gravity with gravitas: A solution to the border puzzle," *American Economic Review*, 2003, *93* (1), 170–192.

Arkolakis, Costas, Arnaud Costinot, and Andrés Rodríguez-Clare, "New trade models, same old gains?," *American Economic Review*, 2012, *102* (1), 94–130.

Baltagi, Badi H, Peter Egger, and Michael Pfaffermayr, "Panel data gravity models of international trade," 2014.

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

Brynjolfsson, Erik, Yu Hu, and Mohammad S Rahman, "Battle of the retail channels: How product selection and geography drive cross-channel competition," *Management Science*, 2009, *55* (11), 1755–1765.

Chaney, Thomas, "Distorted gravity: The intensive and extensive margins of international trade," *American Economic Review*, 2008, *98* (4), 1707–1721.

_ , "The gravity equation in international trade: An explanation," Technical Report, National Bureau of Economic Research 2013.

Disdier, Anne-Célia and Keith Head, "The puzzling persistence of the distance effect on bilateral trade," *Review of Economics and Statistics*, 2008, *90* (1), 37–48.

Eaton, Jonathan and Samuel Kortum, "Technology, geography, and trade," *Econometrica*, 2002, *70* (2), 1741–1779.

_ **, Samuel S Kortum, and Sebastian Sotelo**, "International trade: Linking micro and macro," Technical Report, National Bureau of Economic Research 2012.

Evans, Carolyn L and James Harrigan, "Distance, time, and specialization: Lean retailing in general equilibrium," *American Economic Review*, 2005, *95* (1), 292–313.

Ferreira, Fernando and Joel Waldfogel, "Pop internationalism: Has half a century of world music trade displaced local culture?," *Economic Journal*, 2013, *123* (569), 634–664.

Forman, Chris, Anindya Ghose, and Avi Goldfarb, "Competition between local and electronic markets: How the benefit of buying online depends on where you live," *Management Science*, 2009, *55* (1), 47–57.

Goolsbee, Austan, "Competition in the computer industry: Online versus retail," *Journal of Industrial Economics*, 2001, *49* (4), 487–499.

Grossman, Gene, "Comment," in "The Regionalization of the World Economy," University of Chicago Press, 1998.

Guiso, Luigi, Paola Sapienza, and Luigi Zingales, "Cultural biases in economic exchange?," *Quarterly Journal of Economics*, 2009, *124* (3), 1095–1131.

Harrigan, James and Anthony J Venables, "Timeliness and agglomeration," *Journal of Urban Economics*, 2006, *59* (2), 300–316.

Head, Keith and Thierry Mayer, "Illusory border effects: distance mismeasurement inflates estimates of home bias in trade," in "The Gravity Model in International Trade: Advances and Applications. Editors: Bergeijk and Brakman" Citeseer 2009.

_ **and** _ , "Gravity equations: Workhorse, toolkit, and cookbook," *Handbook of International Economics*, 2014, *4*, 131–195.

Heckman, James J, "Sample selection bias as a specification error," *Econometrica*, 1979, *47* (1), 153–161.

Helpman, Elhanan, Marc Melitz, and Yona Rubinstein, "Estimating trade flows: Trading partners and trading volumes," *Quarterly Journal of Economics*, 2008, *123* (2), 441–487.

Hortaçsu, Ali, F Asís Martínez-Jerez, and Jason Douglas, "The geography of trade in online transactions: Evidence from eBay and MercadoLibre," *American Economic Journal: Microeconomics*, 2009, *1* (1), 53–74.

Huang, Rocco R, "Distance and trade: Disentangling unfamiliarity effects and transport cost effects," *European Economic Review*, 2007, *51* (1), 161–181.

Hummels, David L, "Toward a geography of trade costs," *Available at SSRN 160533*, 1999.

Lendle, Andreas and Pierre-Louis Vézina, "Internet technology and the extensive margin of trade: Evidence from eBay in emerging economies," *Review of Development Economics*, 2015, *19* (2), 375–386.

_ , Marcelo Olarreaga, Simon Schropp, and Pierre-Louis Vézina, "There goes gravity: eBay and the death of distance," *Economic Journal*, 2016, *126* (592), 406–441.

Limao, Nuno and Anthony J Venables, "Infrastructure, geographical disadvantage, transport costs, and trade," *World Bank Economic Review*, 2001, *15* (3), 451–479.

Martin, Will and Cong S Pham, "Estimating the gravity equation when zero trade flows are frequent," *MPRA Paper*, 2008.

Melitz, Marc J, "The impact of trade on intra-industry reallocations and aggregate industry productivity," *Econometrica*, 2003, *71* (6), 1695–1725.

_ and Gianmarco IP Ottaviano, "Market size, trade, and productivity," *Review of Economic Studies*, 2008, *75* (1), 295–316.

Rauch, Ferdinand, "The geometry of the distance coefficient in gravity equations in international trade," 2015.

Rauch, James E, "Networks versus markets in international trade," *Journal of International Economics*, 1999, *48* (1), 7–35.

Silva, JMC Santos and Silvana Tenreyro, "The log of gravity," *Review of Economics and Statistics*, 2006, *88* (4), 641–658.

Appendix

Helpman et al. (2008)'s Results (Reproduced)

	OLS	Probit	NLS	Polynomial	Non Param.
	(1)	(2)	(3)	(4)	(5)
$distance_{ij}$	-1.167***	-0.213***	-0.813***	-0.847***	-0.789***
	(0.040)	(0.016)	(0.049)	(0.052)	(0.088)
$land\ border_{ij}$	0.627***	-0.087	0.871***	0.845***	0.863***
	(0.165)	(0.072)	(0.170)	(0.166)	(0.170)
$island_{ij}$	-0.553**	-0.173**	-0.203	-0.218	-0.197
	(0.269)	(0.078)	(0.290)	(0.258)	(0.258)
$landlock_{ij}$	-0.432**	-0.053	-0.347**	-0.362*	-0.353*
	(0.189)	(0.050)	(0.175)	(0.187)	(0.187)
$legal_{ij}$	0.535***	0.049***	0.431***	0.434***	0.418***
	(0.064)	(0.019)	(0.065)	(0.064)	(0.065)
$language_{ij}$	0.147*	0.101***	-0.030	-0.017	-0.036
	(0.075)	(0.021)	(0.087)	(0.077)	(0.083)
$colony_{ij}$	0.909***	-0.009	0.847***	0.848***	0.838***
	(0.158)	(0.130)	(0.257)	(0.148)	(0.153)
$religion_{ij}$	0.281**	0.141***	0.120	0.139	0.100
	(0.120)	(0.034)	(0.136)	(0.120)	(0.128)
$currency_{ij}$	1.534***	0.216***	1.077***	1.150***	1.107***
	(0.334)	(0.038)	(0.360)	(0.333)	(0.346)
FTA_{ij}	0.976***	0.343***	0.124	0.241	0.065
	(0.247)	(0.009)	(0.227)	(0.197)	(0.348)
$reg.\ costs_{ij}$	-0.146	-0.108***			
	(0.100)	(0.036)			
$days\ \&\ proc._{ij}$	-0.216*	-0.061**			
	(0.124)	(0.031)			
$\hat{\omega}^*_{ij}$			0.840***		
			(0.043)		
$\hat{\eta}^*_{ij}$			0.240**	0.882***	
			(0.099)	(0.209)	
$\hat{z}^*_{ij}$				3.261***	
				(0.540)	
$\hat{z}^{*2}_{ij}$				-0.712***	
				(0.170)	
$\hat{z}^{*3}_{ij}$				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%