VERY PRELIMINARY – DO NOT QUOTE, CITE, OR CIRCULATE

The Effect of Content on Global Internet Adoption*

Abstract

The availability of Internet access is considered important for economic productivity, political freedom, and a well-informed citizenry. It is therefore critical to understand the determinants of Internet adoption. In this paper, we test the effect of content availability on Internet adoption across countries. Controlling for the endogeneity of content with respect to the installed base of Internet users and a range of factors known to affect adoption, we find a statistically and economically significant effect. Although below the effect of per-capita GDP, content's effect is similar to or above that of telephone infrastructure, income equality, population density, primary school enrollment, and degree of civil liberties. Estimating cross-country adoption regressions, we find that a country one standard deviation above the mean level of accessible content has an Internet adoption rate at least 16% higher than the mean country in the sample. We also find evidence that content has a greater effect in more densely populated countries and in countries with more disparate languages. Since content is more easily altered in the short-term than many economic conditions or infrastructure, our results have ramifications for the diffusion of the Internet and the consequent impact on economic development and access to information.

Keywords: technology adoption, economic development, two-sided markets, network externalities, technology diffusion
JEL Classification: O30, O57, L86, L96,

V. Brian Viard
Cheung Kong Graduate School of Business
Beijing 100738
China
brianviard@ckgsb.edu.cn
Tel: 86-10-8518-8858

Nicholas Economides
Stern School of Business
New York University
New York, NY 10012
neconomi@stern.nyu.edu
Tel: 1-212-998-0864

This Draft: 3/24/2009

* We would like to thank Yuxin Chen, Stéphane Straub, Noam Yuchtman, seminar participants at CKGSB and Peking University, and conference participants at the IDEI Conference on the Economics of the Software and Internet Industries for helpful comments. We thank Wang Xin for excellent research assistance. All errors are our own.

VERY PRELIMINARY – DO NOT QUOTE, CITE, OR CIRCULATE

1. Introduction

The adoption of Internet access is considered important for economic productivity, political freedom, and a well-informed citizenry. Moreover, Internet adoption rates vary significantly across countries. As a consequence, there is a large literature examining economic and social determinants of Internet adoption across countries. However, there is no work properly assessing the role of content. In this paper, we test the effect of content availability on Internet adoption across countries. Controlling for the endogeneity of content with respect to the installed base of Internet users and a range of factors known to affect adoption, we find a statistically and economically significant effect of content on Internet adoption. Content's effect on adoption rates is similar to or above that of income equality, population density, primary school enrollment, telephone infrastructure, and the degree of civil liberties, although of lesser importance than that of per-capita GDP.

Content is more easily altered in the short-run than economic, educational, or infrastructure conditions and offers governments and non-governmental organizations (NGOs) a means to more quickly influence Internet diffusion. In addition, governments and NGOs may use content to influence Internet adoption outside of their own political jurisdictions. Many other factors affecting adoption cannot be influenced externally. As we describe later, there are several papers that include content as an explanatory variable in explaining Internet adoption but these papers do not properly control for the endogeneity of content production. Gandal (2006) avoids endogeneity issues by using data on individual-level Internet usage from Quebec but examines a different issue – the potential influence of the Internet on language standards.

There are a number of ways in which governments or NGOs may affect Internet content provision and our paper can be seen as an attempt to quantify the potential impact of these efforts. Canada's International Development Research Centre (IDRC) has funded several projects to directly develop online content in low-income countries. In a review of one project in Uganda, IDRC describes the content development effort as, ". . . perhaps the most complex area of support attempted by the [project] and will likely become increasingly important and valuable to the market."[1]

Other efforts are devoted to making the Internet work more seamlessly in different languages (e.g., the ability to use non-Latin characters in website addresses).[2] A United Nations report noted that, "Another area of growing concern is the absence of African languages on the Internet. The dominance of European languages has limited the spread of Internet use by excluding those not fully literate in those languages. African information ministers meeting in Dakar, Senegal, last year urged new programmes to

[1] *Funding and Implementing Universal Access: Innovation and Experience from* Uganda, Uganda Communications Commission, International Development Research Centre, Ottawa, Ontario (Chapter 3).
[2] An effort to internationalize domain names is currently under way under the auspices of the Internet Governance Forum ("International Net Domains 'Risky,'" *BBC News*, October 30, 2006, http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/6099370.stm.

promote African and other languages on the Internet."[3] Governments and international organizations also affect copyright policies which directly impact access to Internet information. Copyright issues have loomed large in Google's development of Google Book Search, a private-sector effort to significantly expand accessible Internet content in many different languages.[4]

Internet services is an example of a two-sided market. Adoption by users depends on the availability of content and vice-versa. This feedback poses a difficult empirical problem in isolating the effect of content on adoption. Our identification strategy relies on estimating the effect of relevant content produced by "large" countries on adoption by "small" countries. We argue, and provide empirical evidence, that content production by the "large" countries is exogenous to adoption of Internet access in the "small" countries. To identify content relevant to an adopting country we use the number of users of the language in which the content is written. Gandal (2006) provides evidence of language as an indicator of the relevance of Internet content using data on individual-level Internet browsing. The use of language provides significant exogenous variation in relevant accessible content across countries. Estimating cross-country adoption regressions, we find that a country one standard deviation above the mean level of accessible content has an Internet adoption rate significantly higher than that of the mean country in the sample.

The potential benefits of higher Internet adoption are economic, political, and informational. A number of studies attribute aggregate productivity gains in the U.S. to investments in information technology (IT) more broadly. Oliner and Sichel (2000) attribute a large fraction of U.S. labor productivity gains in the late 1990s to IT investments. Gordon (2003) argues that productivity improvements continued after 2000 even though IT investments dropped due to the delay in implementing complementary intangible capital. Van Ark et. al. (2008) attributes a significant portion of the productivity gap between the U.S. and Europe to the slower emergence of the knowledge economy in Europe due in part to lower investment in IT. Such investments may also increase consumer utility in the form of expanded choices and greater convenience which are not properly reflected in the productivity data.

There is less evidence about the specific effects of the Internet on productivity because it originated recently. Litan and Rivlin (2001) estimate relatively large productivity gains from use of the Internet from lower transaction costs, increased management efficiency, and increased competition. Gordon (2000) is more circumspect, arguing that recent productivity gains have occurred primarily in the production of computers and other durable goods but not in the use of the Internet outside these industries. Freund and

[3] "Harnessing the Internet for Development: African Countries Seek to Widen Access, Produce Content," *Africa Renewal*, United Nations, Vol. 20, No. 2, July 2006, page 14.

[4] As of March, 2009 Google had scanned over 7 million books for online searching and was placing newspaper advertisements in more than 70 languages to alert authors of a court settlement over copyrights to the books ("A Google Search of a Distinctly Retro Kind," *New York Times*, March 3, 2009 and "Google Hopes to Open a Trove of Little-Seen Books," *New York Times*, January 5, 2009).

Weinhold (2002) identify an increase in services trade with greater Internet penetration in a country. There is also evidence from industry-specific studies that the Internet increases competition. For example, Goolsbee (2002) finds that the introduction of price comparison websites led to reduced prices for health insurance purchases online, while Scott Morton, et. al. (2001) finds that car referral services on the Internet reduced prices relative to offline purchases.

Although controversial, academic articles and the popular press frequently argue that the Internet leads to greater political freedom. Benkler (2006) surveys reasons why this might be the case but also discusses some of the counter-arguments. Norris (2001) surveys a number of qualitative studies on the relationship between the Internet and democracy. Quantitative estimates of Internet's effect on democracy are scarce. Best and Wade (2005) uses regression analysis to relate Internet adoption and democracy but explores causality only through qualitative arguments. Kedzie (1997) performs a similar analysis but uses instrumental variables to control for the endogeneity of Internet adoption. Both studies find a positive effect of the Internet on democracy. More generally, the Internet leads to greater access to information. In a survey article on social inequality, DiMaggio et. al. (2004) points to the predominance of English-language Internet content as an important dimension of inequality between social and linguistic groups.

We also examine two aspects of agglomeration and the effect of content. We find evidence consistent with awareness of Internet content spreading through social interaction. Content has a greater effect in more densely populated countries. This suggests that it is important to make potential adopters aware of the availability of content in order to spur adoption. In this sense, Internet content and cities are a complement. These results provide an interesting supplement to a study by Sinai and Waldfogel (2004), using intra-country data and focusing on local content. They find that cities attract more local content but that controlling for local content the Internet and cities are substitutes. Thus, we find the opposite for non-local content. We also find that content has a greater effect in countries with more disparate languages. This is consistent with the use of the Internet to overcome linguistic isolation, similar to the effect on racial isolation that Sinai and Waldfogel (2004) finds.

The remainder of the paper is organized as follows. Section 2 discusses our identification strategy and Section 3 presents our model of adoption. Section 4 discusses the data and Section 5 our estimation results. We conclude in Section 6.

2. Identification Strategy

Internet services is an example of a two-sided market as formalized by Rochet & Tirole (2003). In a two-sided market, network externalities for two products interact in a common platform so that "hardware" adoption depends on "software" adoption and vice-versa. In Internet services, access is hardware and content is software. A greater supply of

content drives adoption and a larger installed base drives content creation. Empirically, this feedback creates a difficult identification problem. Simply relating adoption and content will overstate content's effect as it will conflate the effect of content on adoption with the feedback effect of adoption on content.

To disentangle content's effect on adoption we estimate the effect of content created by "large" countries on the Internet adoption rates in "small" countries. We omit the large countries, which we call the content-generating countries, from our estimation of the small countries, which we will refer to as the adopting countries. Our identification relies on the assumption that the content created by the large countries is exogenous to the rate of adoption by small countries. Intuitively, we assume that the number of adopters in the small countries is small enough relative to the number of adopters in the large countries that content creators ignore adoption by the small countries. When we present our data and results we will provide empirical evidence that this is the case. At the same time, we assume that adoption in small countries is influenced by relevant content in large countries because Internet content is ubiquitous.

We define relevant content and therefore "small" and large countries based on language. Specifically, for a language we identify a country or countries that comprise a large percentage of the worldwide users of that language as content-producers. The remaining countries with a smaller population using that language we identify as adopting countries. Therefore, our identification strategy requires languages with a skewed distribution of users – a few countries represent most of the worldwide users while a large number of smaller countries have a negligible percentage of the users. This provides a large amount of data for estimation while satisfying the exogeneity assumption. Since each country's population uses a mixture of languages we construct a weighted-average measure of the relevant content based on the fraction speaking each language. As a byproduct of this, the distribution of languages provides significant cross-sectional variation in relevant content. This also significantly reduces the possibility that there are other factors that affect both content creation in the large countries and adoption rates in the small countries. To do so, these other factors would have to be language-specific. We consider this possibility when we discuss our results.

Our identification strategy is related to that in Gowrisankaran and Stavins (2004), which estimates network externalities in the adoption of the automated clearinghouse system (ACH) by U.S. banks. In estimating the network externalities in adoption, the authors face an identification issue similar to ours. Clusters of banks adopting ACH may be due to network externalities but may also be due to a strong preference for ACH in that locale. To isolate the two effects, the authors use the effect of adoption by small branches of large banks on the adoption decisions of rival banks located in the same local markets as the small branches. The identification argument is based on the fact that a bank must implement ACH at all its branches simultaneously. Our identification strategy differs in that we use the distribution of languages across countries as exogenous variation in the content relevant to each adopting country. This distribution is exogenous since people do

not move to access Internet content, which is ubiquitous in the absence of government restrictions.

Ideally we would also estimate the effect of Internet adoption on content production using a similar identification approach. Unfortunately, this is not feasible given the data as described in Appendix 1.

3. Adoption Model

In this section, we present our model of the relationship between content created by content-producing countries and adoption by adopting countries. Let i index adopting countries, j index languages, and t index years. We model the fraction of people in country i speaking language j who adopt the Internet at time t as:

$$\frac{\text{Adopters}_{ijt}}{\text{Speakers}_{ij}} = \alpha + \beta Z_{it} + \gamma \text{Content}_{jt}, \qquad (1)$$

where Adopters_{ijt} is the number of Internet adopters among speakers of language j in country i at time t, Speakers_{ij} is the number of users of language j in country i which does not vary over time in our data, Content_{jt} is the content available in language j at time t, Z_{it} includes other control variables thought to affect Internet adoption, and $\{\alpha, \beta, \gamma\}$ are parameters to be estimated. We expect $\gamma > 0$. This specification assumes that the effect of content across languages is the same. While in theory we could allow the effect of content on adoption to vary by language, in practice there is insufficient data to identify these separate effects.

Since we only observe the aggregate number of Internet adopters in each county, we transform this equation into one which we can estimate. Multiplying through by the number of speakers of language j and then summing across all languages we obtain:

$$\sum_j \text{Adopters}_{ijt} = \alpha \sum_j \text{Speakers}_{ij} + \beta Z_{it} \sum_j \text{Speakers}_{ij} + \gamma \sum_j \left[\text{Speakers}_{ij} \text{Content}_{jt} \right]. \qquad (2)$$

In order to preserve enough degrees of freedom in our regression we restrict the analysis to include a small set J_I of the most pervasive languages. Including additional languages reduces the number of countries available for analysis since the content-producing

countries for each language must be excluded to maintain the exogeneity assumption. If we include only countries for which:

$$\sum_{j \in J_I} \text{Speakers}_{ij} = \text{Population}_{it}, \quad (3)$$

then:

$$\frac{\sum_{j \in J_I} \text{Adopters}_{ijt}}{\sum_{j \in J_I} \text{Speakers}_{ij}} = \frac{\sum_{j \in J_I} \text{Adopters}_{ijt}}{\text{Population}_{it}} = \text{Penetration}_{it}, \quad (4)$$

where Penetration_{it} is the fraction of the population in country i that have adopted the Internet at time t, which is the data we observe. Dividing both sides of Equation (2) by Population_{it} we then get:

$$\text{Penetration}_{it} = \alpha + \beta Z_{it} + \gamma \frac{\sum_{j \in J_I} \left[\text{Speakers}_{ij} \text{Content}_{jt} \right]}{\text{Population}_{it}} + \varepsilon_{ijt}, \quad (5)$$

where ε_{ijt} is an independent and identically distributed Normal error representing idiosyncratic factors affecting adoption by speakers of language j in country i at time t.

In our data, Equation (3) will not hold precisely because we do not include all the world's languages in our analysis. Fortunately, the approximation will likely bias γ toward zero, understating the true effect of content. To see this, let J_E represent the set of languages excluded from the analysis. When J_E is non-empty, Equation (5) becomes:

$$\text{Penetration}_{it} = \alpha + \beta Z_{it} + \gamma X_{it}^{I} + \hat{\varepsilon}_{ijt}, \quad (6)$$

where $X_{it}^{k} = \frac{\sum_{j \in J_k} \left[\text{Speakers}_{ij} \text{Content}_{jt} \right]}{\text{Population}_{it}}$ $k = I, E$, and $\hat{\varepsilon}_{ijt} = \gamma X_{it}^{E} + \varepsilon_{ijt}$. Applying the omitted variable formula (see Greene (2003), page 148 – 149) we obtain:

$$E\left[\hat{\gamma} \middle| Z_{it}, X_{it}^{I} \right] = \gamma \left(1 + \left(\left[\mathrm{Z}_{it}, \mathrm{X}_{it}^{I} \right]' \left[\mathrm{Z}_{it}, \mathrm{X}_{it}^{I} \right] \right)^{-1} \left[\mathrm{Z}_{it}, \mathrm{X}_{it}^{I} \right]' X_{it}^{E} \right), \quad (7)$$

so that the bias of γ depends on the correlation between the included language-content interaction and the excluded language-content interaction net of the effect of all the other control variables. While the sign of this is theoretically indeterminate, it is likely that this correlation is negative since the correlation between the included and excluded content-language interaction measures is negative. In this case, γ will be biased toward zero. We show in Section 5 that this is consistent with our results when we increase the tolerance with which Equation (3) holds.

4. Data

Our sample includes data on 163 adopting countries and 43 content-generating countries from 1998 to 2004. We also include as "countries" non-self-governing territories for which we have data.[5] We include these in our main analysis because we believe the Internet policies of these territories differ substantially from their governing countries so that they represent independent observations. However, as a robustness check we also check our estimates excluding these observations. [Insert results here.] In this section we describe each of our main variables. Table 1 contains summary statistics on all variables.

Internet Users

Our dependent variable is the number of adopters of Internet access per 100 people in country i at time t. This data is collected by the International Telecommunications Union (ITU) and does not distinguish speeds or modes of Internet access. During the time period of our study, virtually all Internet access was through one of three modes: narrowband (or dial-up) access through a phone line, broadband (or digital subscriber line) access through a phone line, and broadband access through cable lines. The ITU data attempts as best as possible to capture all Internet users regardless of their location.[6] It is unfortunate that the data do not allow us to control for the speed of access since content may drive adoption of higher-quality access. However, during the time of our study most relevant content is text or audio and not visual so that this omission may not be great. There is significant variation in Internet adoption rates as shown in Figure 1.

[5] The non-self-governing territories in our data include overseas territories (Bermuda), overseas regions (French Guiana, Guadeloupe, Martinique), overseas collectivities (French Polynesia, Mayotte), sui generic collectivities (New Caledonia), special administrative regions (Hong Kong, Macao), disputed territories (Palestinian West Bank and Gaza), unincorporated organized commonwealths (Puerto Rico), overseas departments (Reunion), and unincorporated organized territories (Guam, U.S. Virgin Islands). Content measures are not available for Hong Kong, Macao, and Mayotte so they are not used in identifying the effect of content.

[6] ITU's data collection distinguishes between "Estimated Internet Users" and "Internet Subscribers" ("Key Indicators of the Telecommunication/ITC Sector," International Telecommunications Union, 2005). Users of Internet cafes, for example, would be included in the former, which is our variable, but not in the latter.

Language Users

As discussed in Sections 2 and 3, an ideal candidate language satisfies two criteria: 1) it is spoken in many countries and 2) the distribution of its usage is skewed with a few countries comprising a significant fraction of its total users. Such languages simultaneously generate significant data for analysis, while maintaining the exogeneity assumption necessary for identification. [Insert distributions of usage across countries for a couple languages.]

Based on these criteria we choose the following set of included languages:

$$J_I = \left\{ \begin{array}{l} Chinese, Spanish, English, Arabic, Hindi, Portuguese, \\ Bengali, Russian, Japanese, German, French, Hausa, \\ Somali, Zulu, Nyanja, Pulaar, Pular, Swahili \end{array} \right\} \quad (8)$$

The first ten are the most-spoken languages in the world based on *Ethnologue* (2005). French is the seventeenth most-spoken language. The usage of the languages between the tenth and seventeenth (Javanese, Telugu, Marathi, Vietnamese, Korean, and Tamil) is either not widely dispersed or is fairly uniformly distributed across countries, making them unattractive for estimation. The last seven languages were chosen to increase representation of African languages based on the extent to which they meet our criteria. Each of these seven languages is spoken in at least four countries and the two most populous countries speaking the language represent at least sixty percent of the world's population speaking that language.

During the time period of our study Internet content was primarily textual. Therefore, we ideally would use a measure of the number of literate users of each language in creating our relevant content measure. Since we do not know of any source for the number of literate users by language we rely instead on a measure of the number of speakers of each language and include the country's overall literacy rate as a control variable.[7] We are careful, however, to group together spoken dialects whose users employ the same written language.[8] For example, we combine speakers of the many Chinese dialects since they all utilize simplified Chinese as their written language. Similarly, we combine speakers of different Arabic dialects since they all use Standard Arabic as their written language.

Ethnologue (2005) provides detailed estimates of the number of speakers of each of the world's languages by country. It contains comprehensive data on the number of first-

[7] As a robustness check we reran the regression using a content measure that directly adjusts the number of speakers by the overall literacy rate in constructing our content measure. [Incorporate results here.]

[8] The treatment of Creoles is more difficult since *Ethnologue* does not have consistent information about the written language used by Creole speakers. Therefore, we estimate our main results excluding Creole speakers. As a robustness check, we reran the results including Creole speakers. [Incorporate results here.]

language speakers.[9] Therefore, our main estimates use first-language speakers. Its data on second-language speakers is not as complete so we include estimates including second-language speakers only as a robustness check. [Incorporate results here.][10]

Our eighteen included languages constitute 3.07 billion people or 48% of the world's population of 6.40 billion in 2004. To choose the content producers for each language we used the following procedure. Sort the countries in descending order according to the number of speakers of that language. Keep adding countries such that the last country added brings us above 75% of the total speakers of the language. There were three exceptions to this procedure when we kept adding above 75% of the speakers and one exception where we stayed below 75%.[11] Column 3 of Table 2 shows the content-producing countries chosen for each of these languages, while Columns 4 and 5 show the total number of speakers in those countries and the total number of speakers worldwide. Column 6 shows the percentage of the worldwide speakers residing in the content-producing countries. Our identification assumption relies on this being a large number so that content producers are unaffected by smaller countries. For most of the languages content producers represent eighty percent or more of the world's speakers of that language. The only exceptions to this are Bengali, French, and Nyanja.

The last two columns of Table 2 show the number of speakers in the largest adopting country and as a fraction of the total number of speakers of that language. Our identification strategy depends on the latter being a small percentage so that Internet adoption by these countries does not affect content production by the content-producing countries. For most of the languages the largest adopting country contains a small fraction of the total number of speakers of the language. Nyanja, Russian, and French meet the criteria least well with 17.1% of Nyanja speakers in Zambia, 7.8% of Russian speakers in Ukraine, and 6.2% of French speakers in Belgium.

Content

We measure content by the number of host computers connected to the Internet in each year for each content-generating country.[12] Host computers contain the content that users

[9] *Ethnologue* does not distinguish between native and primary first-language speakers. This should be considered in interpreting our results.

[10] Gandal (2006) provides some evidence on second language use and Internet content.

[11] The three exceptions above 75% were because there was an obvious large drop between two countries. For Mandarin Chinese, mainland China alone would bring us above 75% but we added Taiwan because it had 5.2 times as many Mandarin speakers as the next largest country, Malaysia. For English, the U.S. and the U.K. alone would bring us above 75% but we added Canada and Australia because Australia was 4.5 times as large as the next largest country, South Africa. For Portuguese, Brazil alone would bring us above 75% but we added Portugal because it is 15.7 times as large as the next largest country, Paraguay. The one exception below 75% was Bengali. Although Bangladesh represents only 59% of Bengali speakers we did not add more countries because the next largest country, the United Arab Emirates, would add only 0.04%.

[12] Host computers are those computers connected to the Internet with a host name. There are many more computers connected indirectly to the Internet through local area networks (intranets). Importantly for our

can access and the number of computers is proportional to the total available content. This measures only the quantity not quality of content. However, for our estimates to be unbiased it need only be true that quality is proportional to quantity of storage across different languages.

Our measure of the number of Internet host computers is based on data from the Internet Systems Consortium, Inc. (http://www.isc.org/) (ISC). During the time period of our study, ISC takes an annual census of all the host computers connected to the Internet. The technical details are complex due to the sheer size of the Internet but ISC essentially counts the number of Internet Protocol (IP) addresses that have been assigned a Uniform Resource Locator (URL) or name.[13] Each computer on the Internet is assigned an IP address between 1 and 4.3 billion but not every address is used.[14] Only those that have been assigned a Uniform Resource Locator (URL) are actually in use. To determine which have been assigned a URL, ISC must send a query to that address. Since it would take too long for ISC to query every possible number in use, it uses a sophisticated sampling algorithm to reduce the time.[15]

In the process of doing this, ISC also gathers the address, or URL, of each host computer connected to the Internet.[16] This address is used to allocate the hosts to a country. ISC maintained the same sampling procedure throughout the time period of our study, making the number of hosts comparable across years. However, since the storage capacity of computers changed over time, the content measures are not necessarily comparable over time. Another issue is that the storage capacities of computers may vary systematically across countries. However, our results from first-differences regressions will be robust to this problem.

Host addresses are identified by either a two-digit country code (e.g., .za for New Zealand, .uk for United Kingdom, and .ca for Canada) called a country-code Top Level Domain (ccTLD) or a three-digit generic Top Level Domain (gTLD) (e.g., .com, edu, and .org) in their URL. Although the two-digit ccTLD does not necessarily imply that the computer is physically located within the country this is fine for our purposes as long as

---

purposes, content is only accessible on the Internet if it is stored on a host computer. Other computers attached to the intranet can access the Internet but can only store content that is internally available.

[13] An IP address is associated with a single host which is how ISC finds the host names. A request is sent to each active IP address requesting the unique host name. A host may have more than one IP address associated with it so ISC resolves these duplicates.

[14] Technically, the addresses are between 1 and 2^{32}.

[15] More details can be read at http://www.isc.org/index.pl?/ops/ds/.

[16] The relationship between hosts and addresses (URLs) is more complicated. All web pages have a unique URL and are part of a sub-domain which is in turn part of a domain. A domain name such as "google.com" can have many sub-domains such as "www.google.com," "video.google.com," "appengine.google.com," and "investor.google.com". In the early days of the Internet a host commonly had a single sub-domain name. However, sub-domains now commonly map to multiple IP addresses and therefore multiple hosts. The domain naming system is not critical to ISC's host counting since the hosts are uniquely named and have a unique IP address. ISC identifies the sub-domain associated with each host for purposes of allocating content to countries.

the computer contains content created by that country. The rules for assigning ccTLDs differ slightly across countries but generally most countries require a local presence requirement such as citizenship, resident address, or administrative contact residence within the country of an applicant. Using the ccTLD's, ISC assigns each domain to a country. One complication is that there are also Generic Top Level Domains (gTLDs), such as .com, .net, and .org, which can pertain to any country. ISC allocates these based on each country's share of ccTLDs. Thus, the hosts measure across countries is proportional to the hosts measure based on ccTLDs.

Our measure of content for each of the eighteen languages included in our analysis is then based on the total number of hosts for the content-generating countries in that year. For example, for Portuguese-language content we sum the number of host computers in each year for Brazil and Portugal. In our baseline results we assume that all the host computers in content-generating countries pertain to the dominant language in that country. For example, although only 70.5% of Syria speaks Arabic we assume all host computers in Syria contain Arabic content. As a robustness check we re-estimate our regressions weighting the content in each content-producing country by the fraction of the population speaking the dominant language.[17] [Incorporate results here.]

We also create a measure of each country's own content based on its number of hosts using the same methodology as for the content-generating countries. While we do not use this measure directly, we use it to test our exogeneity assumption as we describe further below. Now that we have defined hosts and the measure of content, it is necessary to discuss one issue with ITU's estimates of the number of Internet users described earlier. Prior to 1999, if ITU could not find independent estimates of the number of users it based it on a multiple of the number of hosts in the country. Although the post-1998 ITU reports do not directly address this issue there is a 1999 ITU analyst presentation which states that Internet users were previously sometimes measured by multipliers but are now measured by surveys only.[18] Although the number of hosts and users should be related due to the two-sided nature of the market discussed earlier, it would be problematic for our estimates if there were few countries with independent estimates of hosts and users. To check the accuracy of the ITU analyst's report, we regress the number of Internet users on the number of hosts for the countries in our sample. This yields an R^2 of 0.216, indicating that for the vast majority of the countries there is an independent estimate of users and hosts. In fact, the R^2 is virtually identical to the 0.215 obtained from regressing the number of Internet users on the number of telephone lines for the countries in our sample. Thus, the relationship between number of Internet users and hosts is no more systematic than that between number of Internet users and number of telephone lines.

[17] Doing so requires us to includes two other African languages: Tswana and Central Kanuri.
[18] Report accessible at: www.itu.int/ITU-D/ict/papers/1999/MM-Inet99-Jun99.ppt.

Control Variables

Our goal is to include as many control variables used in previous studies of Internet adoption as possible so as to isolate the effect of content. Therefore, subject to preserving enough degrees of freedom to discern the effect of content, our goal is to maximize the variance explained by our regression rather than the significance of coefficients on individual control variables. When we discuss our results we show that we achieve a high R^2 in our regressions. Thus, the statistical significance of variables in our estimation may be lower than those in previous studies because we include more variables. To identify candidate control variables we rely on previous papers which estimate broadband adoption regressions across countries. The basis of these papers is usually an adoption regression in which the dependent variable is a measure of Internet adoption rates across countries and the dependent variables are social and economic factors thought to influence adoption.

We include a variety of economic, social, and infrastructure variables. Per-capita gross domestic product (GDP) is included as a measure of a country's wealth. We expect this to have a positive effect on adoption. To control for the distribution of wealth within countries we include the Gini coefficient of income. We expect higher income inequality (higher Gini coefficient) to negatively affect adoption. Content is likely more highly valued in countries with more literate and educated populations. To control for this we include the literacy rate and the fraction of school-age children enrolled in primary school. We expect both of these to positively affect Internet adoption. Since familiarity with computing and Internet services is age-dependent, we control for the fraction of the population below twenty years of age, between twenty and 65 years of age, and above 65 years of age.

We include several measures of infrastructure related to providing Internet access. The fraction of the population with fixed phone lines is included as a measure of the telecommunications infrastructure. While there are other means of gaining access to the Internet during this time, these were either rare (satellite and wi-fi) or likely highly correlated with telephone infrastructure (cable television). We also include the cost of a three-minute phone call to the U.S. as a measure of the regulatory and competitive environment of the telecommunications industry. Two other variables, population density and fraction of population living in urban areas, are included to measure infrastructure or demand or both. More densely-populated areas can be served more cheaply on a per-customer basis than more disperse populations. At the same time, it may be that urban residents have different demand for Internet services than rural residents. Finally, we include *Freedomhouse*'s measure of the civil liberties in each country. This measure rates the freedom of citizens to engage in freedom of expression on a scale of one to seven with seven being the most free.[19]

[19] *Freedomhouse* defines seven as the least free. We reverse the order for ease in interpreting our results.

These control variables are drawn from a variety of papers. Wallsten (2006) explains broadband penetration for OECD countries. We include all explanatory variables from this study except for a measure of whether and how non-incumbent phone companies have access to household phone lines because it is proprietary data and available only for OECD countries. Including the cost of telephone calls should serve as a good proxy for this. Wallsten (2005) assesses the impact of regulation on Internet adoption and prices in developing countries. We include all explanatory variables from this study except a regulatory variable which is not publicly available and is not available for developed countries. [The study also includes trade and personal computer density which we will include in a future version of the paper but in a different capacity.] Ford, et. al. (2007) produce a broadband performance index for OECD countries based on the predicted values from an adoption regression. We include all the explanatory variables from this study but use different instruments for Internet access price. [The study also includes household size which we will include in a future version of the paper.]

Chinn and Fairlie (2006) explain cross-country Internet and computer adoption rates, decomposing the differences into variation in observable characteristics and contributions of those characteristics. We include all the explanatory variables in this study that are related to Internet adoption. [This study also includes trade which we will include in a future version of the paper but in a different capacity as explained below.] [Incoporate description of Cava-Ferreruela et. al. (2006).]

There are three papers that include the effect of English language usage on adoption, although they all include only a single language and none of them addresses the endogeneity of content. Hargittai (1999) explains Internet adoption by OECD countries. We include all the explanatory variables in this study except a competitiveness measure because it is a coarser measure of competition than telephone call cost. The author also includes a measure of pervasiveness of English language usage in the country because of the predominance of its use in the media and computing fields. The effect of language is not significant. Similarly, Kiiski and Pohjola (2002) estimate a diffusion model of Internet adoption by OECD countries. We include all the explanatory variables in this study except a measure of telecommunications competition because it is a coarser measure than our measure of telephone call cost. The authors also include a measure of proficiency with the English language for the same reason as Hargittai (1999) but estimate a negative effect. Wunnava and Leiter (2008) also estimate a diffusion model of Internet adoption but with more countries than Kiiski and Pohjola (2002). We include all the explanatory variables in this study. The authors also include a measure of English-language proficiency as a measure of the accessibility of English-language content. They find a positive and significant effect of English language on adoption consistent with our results, although they do not address the endogeneity of content and include only a single language.

More details on all the variables and their sources are provided in Appendix 2. Table 3 shows the correlation between the explanatory variables used in our study. Some

variables are highly correlated with each other consistent with our approach of including as many control variables as possible. Table 4 shows the correlations between the key variables in our study. A country's own content based on the number of host computers is highly correlated with the fraction of Internet users consistent with the two-sided nature of the market. Our measure of relevant content, the language-content interaction measure, is highly correlated with the fraction of Internet users but is not correlated with a country's own content. The latter is nearly zero and insignificant. This is consistent with relevant content influencing Internet adoption separately from the feedback between a country's own content and the number of adopters. This provides additional reassurance that our measure of content is exogenous to adopting countries' adoption rates. This also highlights that a key source of variation in our data is in the distribution of languages across countries.

5. Results

We find that content creation has a positive and significant effect on adoption. The effect is below that of per-capita GDP but is on par with or above that of income inequality, population density, primary school enrollment, telephone infrastructure and the degree of civil liberties. We find evidence that awareness of content spreads through social interaction. Content has a greater effect in countries with denser populations. This is consistent with content being a complement to cities. We also find that content has a greater effect in countries with more disparate languages, consistent with content as a tool to overcome linguistic isolation.

Baseline Results

We estimate Equation (6) with increasingly tighter criteria for the approximation in Equation (3) to assess the tradeoff between more data and the precision of the approximation. We include all countries and all years in the regression but include the language-content interaction data only for those country-year pairs that meet the criteria. For example, in 1998 Singapore has a population of 3.92 million people. Out of the eighteen languages we include, 0.227 million speak English, 0.005 million speak Hindi, 0.0006 million speak Bengali, 0.02 million speak Japanese, and 1.806 million speak Mandarin. Therefore, 2.06 million people or 52.5% of the population in Singapore speak one of the eighteen languages in our included set. Data on Singapore would be included in all regressions but its language-content interaction variable would be included only when we apply a threshold for the approximation in Equation (3) of 52.5% or less. Otherwise, its language-content interaction will be set to zero and a dummy variable for missing language-content interaction set to one.

In all of our regressions we use the Huber/White/sandwich estimator of variances. Heteroskedasticity is possible because Internet adoption rates are generally increasing within a country over time. Thus, we might expect measurement error to increase over

time. Correlation of residuals within a country over time is also quite possible because several variables in our data are the same across years, including the literacy rate, Gini coefficient, population density, and population age.[20]

We first estimate Equation (6) for all countries in the sample without applying any restriction on Equation (3). Therefore, based on the analysis in Section 3 the effect of content on adoption is likely to be downward-biased. Column 1 of Table 5 shows the results. The regression yields an R^2 of 0.68, indicating that a fair amount of the variance has been explained and that there are hopefully few omitted variables. This reflects the fact, discussed earlier, that we include virtually every variable found in previous studies predicting Internet adoption.

Only a few of the control variables are significant although, as noted earlier, this may be because we include more control variables than previous studies. Those that are significant generally have the expected sign. GDP per capita has a positive and very significant effect on adoption. An additional $1,292 in annual per capita income is associated with a one percentage point higher adoption at the mean values of the variables. A country one standard deviation above the mean in the sample has, on average, 7.40 percentage points higher adoption than the mean country. This is a large effect given the mean adoption level of Internet access in the sample is about eleven percent.

Telecommunications infrastructure, as measured by fraction of the population with fixed phone lines, has a positive effect but it is not significant. Greater income inequality is associated with lower Internet adoption, consistent with a concentration of wealth providing access for the few but not the many. The effects are much smaller than for GDP, however, with a country one standard deviation above the mean having, on average, 0.48 percentage points lower adoption. Population density is highly significant with a positive effect on adoption, consistent with easier construction of Internet infrastructure in more densely populated areas or greater demand for Internet in more populous areas. An additional one thousand people per square kilometer is associated with 0.57 percentage points higher adoption. A country one standard deviation above the mean has, on average, 1.18 percentage points higher adoption.

The age category variables are not significant. Countries with more population residing in urban areas have higher adoption rates but the effect is not significant perhaps because it is highly correlated with population density. School enrollment is highly statistically significant. A one percentage point increase in net primary school enrollment is associated with 0.06 percentage points higher adoption. A country one standard deviation above the mean has 0.98 percentage points higher adoption. The degree of civil liberties in a country has a significant effect on Internet adoption, consistent with freedom of

[20] It is also possible that errors are correlated across countries since the language-content interaction variable for each country is a function of the numbers of hosts in the content-generating countries which is common to multiple countries. However, this is not likely since these hosts numbers are interacted with the number of speakers of the languages in each country.

expression leading to a greater flow of information via the Internet. A country one standard deviation above the mean in the index has, on average, 0.55 percentage points higher adoption. Finally, the cost of an international telephone call has a negative but insignificant effect on adoption.

Two variables do not have the expected sign, although only one of them is significant. The literacy rate has a negative and significant effect on the adoption rate. A one percent increase in the literacy rate is associated with 0.06 percentage points lower adoption and a country one standard deviation above the mean has, on average, 1.16 percentage points lower adoption. Since we control for school attendance the literacy variable captures any effect beyond education. The normalized Internet prices are not significant for any of the three years, but they have a positive sign in two of the three years. The insignificance of the result may be due to the fact that Internet prices are only widely available in a single year of our data. The unexpected sign on Internet prices could be due to the endogeity of prices – countries with greater unobserved demand for Internet access are likely to also experience higher Internet access prices. [In a future version of the paper we plan to instrument Internet price for endogeneity.] The coefficients on the year dummies and the constant are consistent with higher Internet adoption rates over time; however, as discussed earlier, this should be interpreted with caution since the content measure that we employ is not guaranteed to be consistent over time.

Magnitude of Effects

We can now compare content's effect on adoption to those of the control variables. Content has a positive and very significant (a t-statistic of 4.14) effect on adoption. A country one standard deviation above the mean in the language-content interaction variable has, on average, 1.75 percentage points higher adoption. This is 16.5% of the mean adoption level of 10.6%. Therefore, countries with a larger number of speakers of languages with more Internet content available in that language have higher Internet adoption rates. We can compare this to the effect of a one standard deviation increase in each of the statistically significant control variables. For per-capita GDP the effect is a 69.9% increase, for income inequality a 4.5% decrease, for population density an 11.1% increase, for school enrollment a 9.3% increase, for civil liberties a 5.2% increase, and for literacy an 11.0% decrease.

The effect of content is much smaller than that of GDP but is comparable to or above the other significant control variables in the regression. The implications of this are important for countries who wish to stimulate adoption of the Internet. If a country can increase GDP then Internet access will increase dramatically, but this is difficult. Subsidizing or directly creating relevant Internet content may be easier and less costly. In addition, governments or non-governmental development organizations can indirectly influence the adoption of Internet access in other countries by creating content relevant to other countries and in the appropriate language.

There are two issues with the above comparison. First, moving any of these independent variables by one standard deviation is a large change. Therefore, it would be useful to estimate the effects from reasonable changes in the independent variables. Second, this comparison assumes that it is equally easy to move any of the independent variables by one standard deviation. Therefore, it would be useful to get some gauge of the difficulty of changing these variables. One way to do this is to compute how quickly each of the independent variables has changed over the time period of our data and compute the effects such average changes would have on adoption. The top panel of Table 6 summarizes the average annual changes in each of the time-varying variables for the set of adopting countries. For example, adoption increased on average 2.3 percentage points per year for the adopting countries in the sample.[21] The two rightmost columns compute the effect that the average change in each of the independent variables would have on adoption evaluated at the mean of all the variables. For example, per-capita GDP increased $416 per year. Evaluated at the sample means, this would increase adoption by 0.32 percentage points for the average adopting country in the sample. This is 14.1% of the average yearly increase of 2.3 percentage points in adoption.

The bottom panel of Table 6 summarizes the average annual change in the time-varying variables for content-producing countries. The far right column represents the increase in the adoption rate as a percentage of the average annual increase in adoption by the adopting countries. For example, per-capita GDP increased $405 per year on average for these countries. Such an increase would stimulate adoption by 0.31 percentage points for the mean country in the sample. This represents 13.7% of the 2.3 percentage point annual increase in Internet usage for the adopting countries. The annual increase in content is substantial for these countries – 857 thousand new hosts per country per year – and this implies a 0.10 percentage point increase in Internet usage. Comparing this to "normal" increases in other variables, we see that its effect is greater than that of all other variables except increases in GDP and declines in costs of telephone access in adopting countries. Content is as or more important than all other variables.

Whether the top or bottom panel of Table 6 is more appropriate in predicting content's effect depends on which you think more accurately predicts rates of change over time (except content which can only be assessed by looking at the content-producing countries). However, the historical rates of change are similar across the two groups so the implications would not be that different either way.

Alternative Specifications

Column 2 of Table 5 repeats the regression in Column 1 but includes a measure of the country's own Internet content. Since this variable is endogenous, its coefficient should be interpreted with caution. However, it is useful to include it in the regression to make

[21] The number of adopting countries is reduced by seventeen and the number of content-generating countries by five because these countries have data for only one year. [Confirm this.]

sure that our measure of content is not simply a proxy for the country's own Internet content. The results show that the effect of the language-content interaction variable actually becomes stronger, both in magnitude and significance, when a measure of the country's own Internet content is included. This is consistent with the language-content interaction independently measuring content that affects but is not affected by Internet adoption in the adopting country. The coefficients on the other variables are largely unaffected by the inclusion of the country's own Internet content. A country's own content is associated with higher adoption and is highly statistically significant. A one standard deviation increase in a country's own content at the mean of all other variables is associated with a 3.2 percentage point increase in adoption. Since the causality is indeterminate here this should be interpreted strictly as a correlation.

To make sure that our results are not driven by a few outliers, the results of a robust regression are reported in Column 3 of Table 5. The regression applies a weight to each observation equal to the reciprocal of the absolute distance of its OLS residual from the median residual and then normalized by seven times the median residual. The coefficient on language-content interaction is even larger in magnitude than in the non-robust regression. [The standard errors in this regression will be adjusted for heteroskedasticity and autocorrelation in a future version of the paper.]

Column 4 of Table 5 repeats the regression in Column 1 but includes country fixed effects. This will control for any unobserved factors at the country level which affect Internet adoption. The cost of doing so is that several factors, including literacy, income inequality, population density, population age, and fraction urban population, must be omitted from the regression because they do not vary over time. The effect of the language-content interaction variable is now diminished to about 41% of the baseline results but it remains significant at the 6% level even after the significant degrees of freedom consumed by the fixed effects.

Threshold Analysis

We now estimate Equation (6) with increasingly tighter criteria for the approximation in Equation (3). Ideally, Equation (3) would hold with equality; however this would require that we have a measure of content for every language spoken in every country included in our data. The problem with doing so is that for each additional language we must exclude additional countries from the analysis as content-producing countries. Thus, there is a tradeoff between the precision with which the approximation in Equation (3) holds and the amount of data we can include in the analysis. Our approach is to choose the languages for which content production is concentrated but use of the language is most dispersed. This allows us to exclude as few countries as possible while having enough observations for identification.

To assess the reasonableness of this approach we estimate Equation (6), applying different thresholds to Equation (3) to see if the results change dramatically. For example,

if we apply a threshold of 50% then we require that at least 50% of the population of a country speak one of our included languages for its language-content interaction measure to be included in the analysis. Otherwise, if it falls below the 50% threshold, we include a dummy variable and set its language-content interaction measure to zero. Thus, only countries with at least 50% of the population speaking the included languages will identify the content effect but all countries help in identifying the other control variables. As we increase the threshold, we lose data but reduce any omitted-variable bias in estimating the language-content interaction coefficient. Changing the threshold also affects the mix of countries included in the analysis, an effect we discuss along with the results.

As we argue earlier in the paper, when Equation (3) does not hold with equality the coefficient on the language-interaction variable is likely biased toward zero and the effect of content understated. Column 1 of Table 7 displays estimates of Equation (6) applying a threshold of zero percent (i.e., the language-content interaction variable is identified only by countries with a non-zero percentage of the population speaking one of the eighteen included languages). These estimates are based on the same data as in Column 1 of Table 5 except that we dummy out ten countries (40 observations) in which none of the included languages is spoken. There are 681 country-year pairs with language-content interaction above zero. The countries included are marked with one or more asterisks in Table 8 (those without any asterisk are those with none of the population speaking an included language). The results are virtually unchanged from our baseline analysis. A country one standard deviation above the mean has 1.83 percentage points higher adoption. This is 17.5% of the mean adoption rate of 10.5 percentage points for these countries.

Column 2 of Table 7 applies a threshold of 25% to Equation (3). That is, the effect of the language-content interaction variable is identified only from countries in which at least 25% of the population has one of the eighteen included languages as their primary spoken language. The number of country-year pairs meeting the threshold drops to 188. The countries included in the analysis are marked with two or more asterisks in Table 8. Note that one ramification of applying a stricter threshold is that the number of African and Pacific countries used to identify the effect of content drops significantly. This results from the fact that the languages spoken in countries within these regions are much less concentrated than those spoken in the rest of the world. The effect of content on adoption is somewhat higher than in Column 1. A country one standard deviation above the mean has, on average, 3.54 percentage points higher adoption. The mean adoption rate for these countries is 12.4 so a country one standard deviation above the mean has, on average, 28.6% higher adoption.

The results with a 50% threshold are shown in Column 3 of Table 7. The number of observations identifying the content effect drops to 102 and the countries included are those with three or more asterisks in Table 8. Note that the effect of content on adoption is now determined primarily by countries in the Americas. For this subset of countries, a

country one standard deviation above the mean has 4.37 percentage points higher adoption, on average. This is 37.8% of the mean adoption rate of 11.6 for this set of countries.

Finally, Column 4 of Table 7 shows the results applying a 75% threshold. That is, only countries for which at least 75% of the population speaks one of the eighteen included languages are used to identify the effect of content on adoption. The number of observations identifying the content effect drops to 76 and only the countries with four asterisks in Table 8 are used to identify the effect of content. For this set of countries, a country one standard deviation above the mean has, on average, 6.47 percentage points higher adoption. This is 57.4% of the average adoption rate of 11.3 for this set of countries.

As we increase the precision with which Equation (3) holds, content has a greater predicted effect on adoption, consistent with the predicted downward bias discussed earlier. However, this claim should be treated with some caution since the mix of countries changes as we increase the threshold. In particular, countries in the Americas are over-represented as we increase the threshold.[22] Taken as a whole, however, the results are fairly consistent and suggest a positive and significant effect of content on adoption. The statistical significance of the language-content interaction variable is also roughly constant as we increase the threshold from 0% to 50% (with a t-statistic of about 4.5) but jumps to about 10 for the 75% threshold.

First Differences Estimates

Since we have panel data, we can also estimate the effect of content on adoption using first differences and allow for unobserved country-specific factors that affect adoption. Specifically, suppose that we modify our model of adoption to:

$$\text{Penetration}_{it} = \alpha_t + \beta Z_{it} + \gamma \frac{\sum_{j \in J_I} \left[\text{Speakers}_{ij} \text{Content}_{jt} \right]}{\text{Population}_{it}} + \xi_i + \varepsilon_{it}, \qquad (9)$$

where ξ_i captures unobserved (to the econometrician) factors in country i affecting adoption. Taking the difference in Internet penetration between periods t and $t-1$ we obtain:

$$\Delta\text{Penetration}_{it} = \Delta\alpha_t + \beta \Delta Z_{it} + \gamma \Delta X^I_{it} + \eta_{it}, \qquad (10)$$

[22] Although the countries are concentrated in the Americas the languages represented are diverse with the following number of observations identifying the content effect at the 75% threshold: Spanish (49), Mandarin Chinese (48), Arabic (44), English (33), Japanese (15), German (13), Portuguese (7), Russian (7), French (7), Hindi (6), and Pulaar (4).

where $\Delta\text{Penetration}_{it}$ is the change in Internet adoption in country i between $t-1$ and t, $\Delta\alpha_t$ is the change in baseline penetration, ΔZ_{it} is the change in control variables, and ΔX_{it}^{I} is the change in the language-content interaction variable. The advantage of this approach is that it removes any spurious correlations in the levels of the variables and controls for any country-level unobservables that affect adoption. The disadvantage is that we lose one year's worth of data and we cannot identify the effect of any control variables that are measured only once during the time period of our data.

The results of the first-differences estimation applying different thresholds to Equation (3) are shown in Table 9. The effect of a change in content on the change in adoption increases both in statistical and economic significance as the threshold becomes stricter. At a 0% threshold the coefficient is significant only at the 17% level. A country with a change in language-content interaction one standard deviation above the mean has a 0.18 percentage point higher change in adoption rates. This is 6.9% of the mean change in adoption rates for this set of countries. When the threshold is increased to 25% the statistical significance increases to the 12% level and a country with a change in language-content interaction one standard deviation above the mean has a change in adoption 13.3% above the mean change for that set of countries. At a threshold of 50% the corresponding statistical significance is the 6.7% level and economic magnitude is 23.7%. Finally, with a threshold of 75% the significance level is 3.8% and the economic magnitude is 36.4%. These results are also consistent with a downward bias due to omission of some languages.

Even controlling for unobserved country-level factors affecting adoption, content has a significant effect, both statistically and economically, on Internet adoption rates at least when content is measured for most of the speakers of a language in a country.

Linguistic and Geographic Isolation

Internet content may act as a substitute or complement to isolation. On the one hand, isolated populations may use the Internet as a means to access people with similar interests or characteristics. If this is the case, content would have a greater effect on isolated groups. On the other hand, people may learn about the usefulness of the Internet through word-of-mouth and this is more likely if they are less isolated. If this is the case, content would have a smaller effect on isolated groups. We attempt to distinguish these two alternatives using two different measures of isolation: geographic isolation, as measured by population density, and linguistic isolation, as measured by linguistic heterogeneity.

To test whether the effect of content varies with geographic isolation we interact population density with the language-content interaction measure. The results are shown in Column 1 of Table 10. The baseline effect of content is insignificant but the interaction effect is very significant. A country one standard deviation above the mean population

density at the mean language-content interaction has 1.98 percentage points higher adoption rate. This is a 18.7% increase over the mean adoption rate of 10.6.

These results are consistent with awareness of Internet content spreading by word-of-mouth. People living closer to one another are more likely to disseminate information about the benefits of the Internet. Stated the other way around, Internet content appears to be less effective in driving adoption for more geographically isolated people. In this sense, our results are consistent with the Internet being a complement to cities.

To test the effect of linguistic isolation, we create a Herfindahl index of languages spoken in each of the adopting countries:

$$\text{HHI}_{it} = \sum_{j \in J_I} \left(s_{ijt} \right)^2 , \qquad (11)$$

where $s_{ijt} = \dfrac{\text{Speakers}_{ij}}{\sum_{j \in J_I} \text{Speakers}_{ij}}$. We then interact this index with the language-content interaction variable in our regression:

$$\text{Penetration}_{it} = \alpha + \beta Z_{it} + \gamma X_{it}^{I} + \theta \text{HHI}_{it} + \delta \text{HHI}_{it} * X_{it}^{I} + \varepsilon_{ijt} . \qquad (12)$$

In this case, $\delta > 0$ indicates content has a greater effect in linguistically homogeneous countries while $\delta < 0$ indicates the opposite. Column 2 of Table 10 shows the results. The baseline effect of linguistic homogeneity is insignificant but the interaction effect is significantly negative. Content has a smaller effect in countries with more homogeneous language speakers. A country one standard deviation above the mean in its language Herfindahl index at the mean language-content interaction has 0.86 percentage points lower adoption. This is a 8.1% decrease over the mean adoption rate in the sample. This result is consistent with the Internet as a tool to overcome linguistic isolation. Those speaking a less common language may use the Internet to access content in their language.

It is interesting to relate our results to those of Sinai and Waldfogel (2004). Using individual-level data from the U.S., they find that the Internet is both a complement to and a substitute for cities. They find that more populous cities have more locally-targeted Internet content which in turn drives higher Internet access. A different mechanism underlies the positive relationship between Internet adoption and population density in our data since we measure non-local content. Sinai and Waldfogel also find that the Internet and cities are substitutes in that, conditional on availability of local content, Internet access rates are lower in more populous cities. In contrast, we find that non-local content has a greater effect on adoption in more populous cities. Finally, Sinai and Waldfogel find that the Internet is used to overcome racial isolation; blacks are more

likely to adopt the Internet if they are a smaller fraction of the local population. We find a similar effect for linguistic isolation.

Alternative Explanations

Our adoption model assumes that "large" countries affect "small" countries through an indirect network effect: the production of content by large countries drives Internet adoption by small countries through the shared platform of the Internet infrastructure. An alternative explanation is that large countries influence small countries through a direct network effect: a common language between countries leads to increased economic activity and therefore more communication via the Internet, such as email or bulletin boards.

To test this alternative, we estimate the same model but include a language-weighted measure of trade between the large and small countries (a language-trade interaction) in addition to the language-content interaction measure:

$$\text{Penetration}_{it} = \alpha + \beta Z_{it} + \gamma_1 X^{I}_{it} + \gamma_2 \sum_{j \in J_I} \frac{\sum_{k \in K_j} [\text{Speakers}_{ik} \text{Trade}_{ikt}]}{\sum_{k \in K_j} \text{Trade}_{ikt}} + \varepsilon_{ijt}, \qquad (13)$$

where K_j is the set of content producing countries for language j and Trade_{ikt} is the trade between country i and country k and Trade_{it} is the total trade between country i and all other countries.

If direct network effects are driving the relationship between countries then we expect γ_2 to be significant and positive and to reduce γ_1 to be insignificant. [In a future version of the paper we plan to estimate this using both inbound and outbound trade. We also plan to estimate the same regression using measures of tourism rather than trade.]

6. Conclusion

We find that web content "exported" by large countries plays a significant role in stimulating Internet access adoption in smaller countries. Its effect is on par with or above other important social and economic factors such as population density, income inequality, education, telecommunications infrastructure and civil liberties. Thus, content can play a crucial role in encouraging diffusion of the Internet. Importantly, governments and non-governmental organizations can influence adoption in other countries through this mechanism. In fact, this is implicit in our estimation strategy. More targeted Internet content is likely to have even greater effects than we find since we treat all content in a given language as equally relevant in our estimation.

We also find that content has a greater effect in more urban areas. This is consistent with "word-of-mouth" playing a role in spreading awareness of content. Countries with more disparate languages are also more affected by content than are those with homogeneous languages. Thus, Internet content may play a role in overcoming social isolation.

Our data do not allow us to quantity the costs and benefits of content creation. Since we find that content creates a positive externality in spurring Internet adoption the benefits of content creation may not be fully internalized in the market. It would be useful to quantity the magnitudes of the spillover and compare it to the costs of content creation.

Bibliography

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*, Yale University Press, New Haven, Connecticut.

Best, M. L. and K. W. Wade (2005). "The Internet and Democracy: Global Catalyst or Democratic Dud?" The Berkman Center for Internet & Society Research Publication No. 2005-12.

Cava-Ferreruela, I. and A. Alabau-Muñoz (2006). "Broadband Policy Assessment: A Cross-National Empirical Analysis," *Telecommunications Policy*, 30, 445 – 63.

Chinn, M. D. and R. W. Fairlie (2006). "The Determinants of the Global Digital Divide: A Cross-Country Analysis of Computer and Internet Penetration," *Oxford Economic Papers*, 59, 16 – 44.

DiMaggio, P. et. al. (2004). "Digital Inequality: From Unequal Access to Differentiated Use," in *Social Inequality*, K. M. Neckerman, editor, Russell Sage Foundation, New York, NY.

Freund, C. and D. Weinhold (2002). "The Internet and International Trade in Services," *American Economic Review Papers and Proceedings*, 92, 236 – 240.

Gandal, N. (2006). "Native Language and Internet Use," *International Journal of the Sociology of Language*, 182, 25 – 40.

Ford, G., Koutsky T. and L. Spiwak (2007). "The Broadband Performance Index: A Policy-Relevant Method of Comparing Broadband Adoption Among Countries," working paper.

Goolsbee, A. (2002). "Does the Internet Make Markets More Competitive? Evidence from the Life Insurance Industry," *Journal of Political Economy*, 110, 481 – 507.

Gordon, R., editor (2005). *Ethnologue: Languages of the World*, 15th edition, SIL International, Dallas, Texas.

Gordon, R. J. (2000). "Does the 'New Economy' Measure up to the Great Inventions of the Past?" *Journal of Economic Perspectives*, 14, 49 – 74.

Gordon, R. J. (2003). "Exploding Productivity Growth: Context, Causes, and Implications," *Brookings Papers on Economic Activity*, 34, 207 – 298.

Gowrisankaran, G. and J. Stavins (2004). "Network Externalities and Technology Adoption: Lessons from Electronic Payments," *RAND Journal of Economics*, 35, 260 – 276.

Greene, W. (2003). *Econometric Analysis*. Prentice-Hall, Inc., Upper Saddle River, New Jersey.

Hargittai, E. (1999). "Weaving the Western Web: Explaining Differences in Internet Connectivity Among OECD Countries," *Telecommunications Policy*, 23, 701 – 718.

Kedzie, C. R. (1997). "Communication and Democracy: Coincident Revolutions and the Emergent Dictator's Dilemma," RAND PRGS Dissertations No. RGSD-127, accessed at http://www.rand.org/pubs/rgs_dissertations/RGSD127/index.html.

Kiishi, S. and M. Pohjola (2002). "Cross-Country Diffusion of the Internet," *Information Economics & Policy*, 14, 297 – 310.

Litan, R. E. and A. M. Rivlin (2001). "Projecting the Economic Impact of the Internet," *American Economic Review*, 91, 313 – 317.

Norris, P. (2001). *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*, Cambridge University Press, Cambridge, United Kingdom.

Oliner, S. D. and D. E. Sichel (2000). "The Resurgence of Growth in the Late 1990s: Is Information Technology the Story?" *Journal of Economic Perspectives*, 14, 3 – 22.

Rochet, J. and J. Tirole (2003). "Platform Competition in Two-Sided Markets," *Journal of the European Economic Association*, 1, 990 – 1029.

Scott Morton, F., F. Zettelmeyer, and J. Silva-Risso (2001). "Internet Car Retailing," *Journal of Industrial Economics*, 49, 501 – 519.

Sinai, T. and J. Waldfogel (2004). "Geography and the Internet: Is the Internet a Substitute or a Complement for Cities?" *Journal of Urban Economics*, 56, 1 – 24.

Van Ark, B., M. O'Mahony and M. P. Timmer (2008). "The Productivity Gap between Europe and the United States: Trends and Causes," *Journal of Economic Perspectives*, 22, 25 – 44.

Wallsten, S. (2005). "Regulation and Internet Use in Developing Countries," *Economic Development and Cultural Change*, 53, 501 – 523.

Wallsten, S. (2006). "Broadband and Unbundling Regulations in OECD Countries," working paper.

Wunnava, P. and D. Leiter (2008). "Determinants of Inter-Country Internet Diffusion Rates," working paper.

Appendix 1 Difficulties in Estimating Effect of Adoption on Content

There are two problems with applying our identification approach in estimating the effect of adoption on content creation. One is methodological and the other due to data limitations. Suppose we follow an approach parallel to that used for adoption. In this case we assume that content production by a small country is driven by Internet adoption in large countries and that the latter is exogenous with respect to the small country's content production. We exclude the large countries from the analysis and specify content production by the small countries as:

$$\text{Content}_{ijt} = \phi + \rho W_{it} + \delta \text{Content}_{-ijt} + \theta \text{Adopters}_{jt} \ \forall j ,$$

where $-i$ represents all countries except country i, Adopters_{jt} is the number of adopters in the large countries, and W_{it} includes control variables affecting content production. The first problem in estimation is the inclusion of $Content_{-ijt}$. It is necessary to control for content in language j produced by all other countries since it is a substitute for country i's content. This problem did not arise in the adoption model because adoption by users outside a country is not a substitute for adopters inside. To see the problem explicitly, solve for the reduced form of this system of equations:

$$\text{Content}_{ijt} = \tilde{\phi} + \tilde{\rho} W_{it} + \tilde{\theta} \text{Adopters}_{jt} \ \forall j .$$

The structural parameter of interest, θ, cannot be recovered from the reduced form parameters without instruments for content production by other countries. Even if we could find suitable instruments for content production in other countries there is another impediment to estimating this equation. Since we only observe content and adoption at the country rather than language level we must sum both sides of the equation across all languages (and add an error term of unobserved factors affecting content production by country i at time t):

$$\text{Content}_{it} = \hat{\phi} + \hat{\rho} W_{it} + \tilde{\theta} \text{Adopters}_{t} + \varepsilon_{it} ,$$

While this equation is theoretically estimable, there is almost no variation in the data to identify the effects. The content created in each country depends on the total number of Internet users worldwide. Therefore, the only variation in the data would be over time. In the adoption regressions we use the fact that the total number of adopters in each country must equal the sum of adopters across all languages spoken in the country. As a result, the distribution of languages provided significant variation through the weighted average of content. This relationship does not carry over here due to the ubiquity of content – content is affected by worldwide adoption rather than by a language weighted average of adopters.

Appendix 2

Variable Description and Data Sources

Variable	Description	Frequency/ Availability	Data Source
Internet Users	Number of people per 100 population with some form of Internet access.	Yearly/1998 - 2004	International Telecommunications Union
Per-Capita GDP	GDP per-capita in current U.S. dollars using purchasing power parity.	Yearly/1998 - 2004	World Bank
Fraction w/ Phone Lines	Fraction of the population with telephone main lines in use.	Yearly/1998 - 2004	International Telecommunications Union
Literacy Rate	Literacy rate of population aged 15 and above, years 2000 to 2005.	Once	*The State of the World's Children 2008* , United Nations Childrens Fund
Gini Coefficient	Gini coefficient of inequality of income distirbution, various years from 1995 to 2006	Once	2006 United Nations Human Development Report, Table 15
Normalized Internet Price	Internet monthly access price for 20 hours of off-peak use in US$ (1998 and 2000); Internet monthly access price for 30 hours of peak use in US$ (2001)	Yearly/1998, 2000 - 2001	"Challenges to the Network: Internet for Development," *ITU Internet Report,* 1999; "IP Telephony," *ITU Internet Report,* 2001; "Internet for a Mobile Generation," *ITU Internet Report,* 2002
Population Density	Thousands of people per square kilometer, year 2000.	Once	United Nations Statistics Division
Age Below 20	Fraction of population age 19 and below, year 2000.	Once	United Nations Statistics Division
Age Above 65	Fraction of population age 65 and above, year 2000.	Once	United Nations Statistics Division
Fraction Urban Population	Fraction of population living in urban areas, year 2000	Once	United Nations Statistics Division
Fraction School Enrollment	Fraction net enrollment in primary education for both sexes, years 1999 to 2004.	Yearly/1999 - 2004	United Nations Statistics Division
Civil Liberties Index	Civil liberties measured on a one-to-seven scale, with one representing the lowest degree of freedom and seven the highest, years 1998 to 2004.	Yearly/1998 - 2004	*Freedom in the World*
Cost of Telephone Call	Average cost of three-minute telephone call to the United States in US$, years 1998 to 2004.	Yearly/1998 - 2004	World Development Indicators
Language-Content Interaction	Millions of hosts of "relevant" content. See text for detailed description.	Yearly/1998 - 2004	*Ethnologue* (2007) (language) and Internet Systems Consortium (hosts)
Own Content	Millions of hosts. See text for detailed description.	Yearly/1998 - 2004	Internet Systems Consortium

VERY PRELIMINARY – DO NOT QUOTE, CITE, OR CIRCULATE

Table 1

Descriptive Statistics, 163 Countries 1998 - 2004

Variable	N	Mean	Standard Deviation	Min	Max
Internet Users	1,082	10.591	14.645	0.010	75.460
Per-Capita GDP	882	8.778	9.557	0.450	60.249
Fraction w/ Phone Lines	717	0.254	0.211	0.000	0.908
Literacy Rate	673	0.813	0.196	0.240	1.000
Gini Coefficient	626	0.411	0.110	0.247	0.743
Normalized Internet Price (1998)	25	0.003	0.003	0.001	0.011
Normalized Internet Price (2000)	25	0.002	0.002	0.001	0.007
Normalized Internet Price (2001)	101	0.027	0.041	0.001	0.207
Population Density	1,081	0.427	2.083	0.000	21.483
Age Below 20	991	0.417	0.118	0.196	0.605
Age Above 65	991	0.068	0.045	0.011	0.182
Fraction Urban Population	1,081	0.555	0.243	0.077	1.000
Fraction School Enrollment	596	0.878	0.155	0.278	1.000
Civil Liberties Index	968	4.680	1.720	1.000	7.000
Cost of Telephone Call	660	3.530	2.843	0.172	15.963
Language-Content Interaction (millions of relevant hosts)	721	2.990	15.068	0.000	193.026
Own Content (millions of hosts)	1,028	0.083	0.360	0.000	5.434
Year 1999	1,082	0.143	0.351	0.000	1.000
Year 2000	1,082	0.148	0.355	0.000	1.000
Year 2001	1,082	0.145	0.352	0.000	1.000
Year 2002	1,082	0.142	0.350	0.000	1.000
Year 2003	1,082	0.145	0.352	0.000	1.000
Year 2004	1,082	0.144	0.351	0.000	1.000
Language-Content Interaction (> 0% Threshold)	681	3.165	15.487	0.000	193.026
Internet Users (> 0% Threshold)	681	10.460	13.569	0.010	65.680
Language-Content Interaction (25% Threshold)	188	9.063	28.139	0.002	193.026
Internet Users (> 25% Threshold)	188	12.388	14.186	0.040	60.990
Language-Content Interaction (50% Threshold)	102	15.957	36.844	0.003	193.026
Internet Users (> 50% Threshold)	102	11.550	14.221	0.040	60.990
Language-Content Interaction (75% Threshold)	76	15.967	38.792	0.005	193.026
Internet Users (> 75% Threshold)	76	11.277	15.438	0.040	60.990

See Appendix 2 for a description of the variables and their sources.

Table 2

Profiles of Langauges Included in Analysis[1]

Ranking[2]	Language	Content Producing Countries	Included Speakers[3]	Total World Speakers[3]	% of World Speakers Included	Largest Adopting Country		
						Name	Number Speakers[3]	% of World Speakers
1	Mandarin Chinese	China	1,171.05	1,204.76	99.0%	Malaysia	4.39	0.36%
		Taiwan	22.69					
2	Spanish	Mexico	86.21	322.30	79.5%	Domincan Republic	6.89	2.14%
		Columbia	34.00					
		Argentina	33.00					
		Spain	28.17					
		Venezuala	21.48					
		Peru	20.00					
		Chile	13.80					
		Cuba	10.00					
		Ecuador	9.50					
3	English	United States	210.00	309.35	96.3%	South Africa	3.46	1.12%
		United Kingdom	55.00					
		Canada	17.10					
		Australia	15.68					
4	Arabic	Egypt	68.24	215.21	88.6%	Mauritania	2.48	1.15%
		Algeria	20.50					
		Morocco	18.80					
		Iraq	18.29					
		Sudan	15.02					
		Yemen	14.68					
		Saudi Arabia	14.59					
		Syria	11.47					
		Tunisia	9.00					
5	Hindi	India	180.77	180.77	99.6%	South Africa	0.89	0.49%
6	Portuguese	Brazil	163.15	177.46	97.6%	Paraguay	0.64	0.36%
		Portugal	10.00					
7	Bengali	Bangladesh	171.07	171.07	58.5%	United Arab Emirates	0.07	0.04%
8	Russian	Russia	145.03	145.03	81.3%	Ukraine	11.34	7.82%
9	Japanese	Japan	122.43	122.43	98.9%	Singapore	0.02	0.02%
10	German	Germany	75.30	95.39	86.8%	Kazakhstan	0.96	1.00%
17	French	France	64.86	64.86	78.6%	Belgium	4.00	6.17%
	Hausa	Nigeria	18.53	24.16	97.4%	Chad	0.10	0.41%
		Niger	5.00					
	Somali	Somalia	7.78	12.65	87.9%	Djibouti	0.29	2.30%
		Ethiopia	3.33					
	Zulu	South Africa	9.20	9.56	96.2%	Lesotho	0.25	2.59%
	Nyanja	Malawi	7.00	9.35	74.9%	Zambia	1.60	17.11%
	Pulaar	Senegal	2.39	3.24	81.7%	Guinea-Bassau	0.25	7.56%
		Gambia	0.26					
	Pular	Guinea	2.55	2.92	87.4%	Sierra Leone	0.18	6.12%
	Swahili	Tanzania	0.54	0.77	86.9%	Oman	0.02	2.85%
		Kenya	0.13					
		Total	2,927.60	3,071.29	91.8%		37.80	1.23%

[1] Source: Ethnologue (2007)
[2] Most-spoken languages by first-language speakers according to Ethnologue (200?)
[3] Millions
[4] As fraction of speakers in content-generating countries.

Table 3

Correlation Matrix for Characteristics of Sample Countries 1998 - 2004 (N = 1,082)

	Language-Content Interaction	Per Capita GDP	Fraction w/ Phone Lines	Literacty Rate	Gini Coefficient	Normalized Internet Price	Population Density	Age Below 20	Age Above 65	Fraction Urban Population	Fraction School Enrollment	Civil Liberties Index	Cost of Telephone Call
Per-Capita GDP	0.095												
	(0.002)												
Fraction w/ Phone Lines	0.170	0.579											
	(0.000)	(0.000)											
Literacy Rate	-0.160	-0.147	-0.255										
	(0.000)	(0.000)	(0.000)										
Gini Coefficient	-0.023	0.009	-0.189	0.334									
	(0.442)	(0.778)	(0.000)	(0.000)									
Normalized Internet Price	-0.023	-0.078	-0.101	-0.018	0.064								
	(0.450)	(0.010)	(0.001)	(0.562)	(0.035)								
Population Density	0.010	0.123	0.146	-0.137	-0.099	-0.028							
	(0.752)	(0.000)	(0.000)	(0.000)	(0.001)	(0.358)							
Age Below 20	-0.164	-0.333	-0.535	0.290	0.301	0.125	-0.186						
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)						
Age Above 65	-0.034	0.504	0.419	0.099	-0.040	-0.075	0.013	-0.045					
	(0.266)	(0.000)	(0.000)	(0.001)	(0.189)	(0.014)	(0.670)	(0.144)					
Fraction Urban Population	0.175	0.431	0.561	-0.171	-0.285	-0.110	0.267	-0.492	0.370				
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)				
Fraction School Enrollment	0.003	0.343	0.108	0.192	0.176	-0.037	-0.032	-0.002	0.247	0.073			
	(0.925)	(0.000)	(0.000)	(0.000)	(0.000)	(0.226)	(0.300)	(0.958)	(0.000)	(0.016)			
Civil Liberties Index	-0.015	0.426	0.210	0.017	0.181	-0.015	-0.084	-0.225	0.120	-0.016	0.288		
	(0.613)	(0.000)	(0.000)	(0.568)	(0.000)	(0.627)	(0.006)	(0.000)	(0.000)	(0.593)	(0.000)		
Cost of Telephone Call	-0.087	-0.135	-0.190	0.202	0.161	0.131	-0.066	0.261	-0.091	-0.216	-0.027	-0.038	
	(0.004)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.031)	(0.000)	(0.003)	(0.000)	(0.372)	(0.208)	
Own Content	0.022	0.428	0.291	-0.175	-0.003	-0.033	0.011	-0.165	0.277	0.161	0.189	0.209	-0.097
	(0.465)	(0.000)	(0.000)	(0.000)	(0.916)	(0.276)	(0.730)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.001)

Signficance levels are in parentheses.

Table 4

Adoption/Content Correlation Matrix for Sample Countries 1998 - 2004 (N = 721)

	Internet Users	Language-Content Interaction
Language-Content Interaction	0.346 (0.000)	
Own Content	0.529 (0.000)	0.013 (0.724)

Signficance levels are in parentheses.

Table 5

Effect of Content on Internet Adoption for All Sample Countries
1998 - 2004 (N = 1,082)

	Baseline	Own Content	Robust Regression	Country Fixed Effects
Per-Capita GDP	0.7739 ***	0.6478 ***	0.6689 ***	1.2896 ***
	(0.0674)	(0.0601)	(0.0238)	(0.2443)
Fraction w/ Phone Lines	2.9844	3.2265	1.3660	-16.0729 ***
	(3.1603)	(2.9578)	(1.1468)	(2.5215)
Literacy Rate	-5.9429 ***	-4.2921 ***	-2.5492 **	
	(1.5284)	(1.4082)	(1.0452)	
Gini Coefficient	-4.3779 *	-2.2594	-1.9587	
	(2.5904)	(2.4117)	(1.7301)	
Normalized Internet Price (1998)	-34.1915	-98.1802	-106.7813	-296.7054
	(380.9571)	(340.6991)	(225.8754)	(363.5217)
Normalized Internet Price (2000)	677.9763	502.5321	64.2411	315.3866
	(558.5902)	(544.8038)	(375.5117)	(292.5760)
Normalized Internet Price (2001)	11.8821	6.7391	1.7838	4.2079
	(11.0475)	(9.7345)	(9.7321)	(6.7891)
Population Density	0.5687 ***	0.5373 ***	1.7852 ***	
	(0.2161)	(0.1914)	(0.0665)	
Age Below 20	-13.6510	-8.5860	0.8145	
	(20.3906)	(19.7677)	(9.0125)	
Age Between 20 and 65	-1.1140	11.6949	2.8627	
	(16.7604)	(16.5041)	(7.1667)	
Age Above 65	7.2389	16.1203	15.1056	
	(30.7974)	(30.3744)	(13.5667)	
Fraction Urban Population	1.3098	1.5573	2.5454 ***	
	(1.8643)	(1.8055)	(0.7734)	
Fraction School Enrollment	6.3392 ***	4.0853 **	1.2630	3.3788 *
	(1.9447)	(1.7982)	(1.3111)	(2.0109)
Civil Liberties Index	0.3178 *	0.3164 *	0.2754 ***	-0.1239
	(0.1717)	(0.1663)	(0.1033)	(0.3824)
Cost of Telephone Call	-0.2041	-0.0705	0.0660	0.1168
	(0.1275)	(0.1110)	(0.0640)	(0.1022)
Language-Content Interaction	0.1159 ***	0.1301 ***	0.2613 ***	0.0472 *
	(0.0278)	(0.0236)	(0.0103)	(0.0246)
Own Content		8.8018 ***		
		(1.3262)		
R-Squared	0.6777	0.7159	0.8761	0.5233

For Columns 1 and 2, robust standard errors in parentheses. * = 10% significance, ** = 5%significance, *** = 1%significance. Standard errors in Column 3 are unadjusted [This will be done in a future version of the paper.] Year dummies and dummy variables for missing values of all variables included in all regressions. Countries with non-mising langauge-content interaction are listed in Table 8.

Table 6

Estimated Effects of Variables on Adoption

Variable	N	Average Annual Change 1998 - 2004	Implied Increase in Adoption for Adopting Countries[1]	Increase in Internet Usage by Adopting Countries[2]
Adopting Countries				
Internet Users	146	2.284		
Per-Capita GDP	126	0.416	0.32	14.1%
Fraction w/ Phone Lines	35	0.001	0.00	0.1%
Fraction School Enrollment[3]	87	0.005	0.03	1.4%
Civil Liberties Index	134	0.051	0.02	0.7%
Cost of Telephone Call	80	-0.513	0.10	4.6%
Content-Producing Countries				
Per-Capita GDP	38	0.405	0.31	13.7%
Fraction w/ Phone Lines	13	0.006	0.02	0.8%
Fraction School Enrollment[3]	27	0.011	0.07	3.0%
Civil Liberties Index	40	0.054	0.02	0.8%
Cost of Telephone Call	20	-0.429	0.09	3.8%
Content (millions of Hosts)	39	0.857	0.10 [4]	4.3% [4]

[1] Marginal effect evaluated at the means of all other independent variables.
[2] Relative to the average annual increase in Internet users in adopting countries.
[3] Increase from 1999 - 2004.
[4] Assumes all content is "relevant" as defined in the text.
Content-producing countries are identified in Table 2 and adopting countries in Table 8.

Table 7

Effect of Content on Internet Adoption for Sample Countries Meeting Threshold Criteria 1998 - 2004 (N = 1,082)

	> 0% Threshold	>= 25% Threshold	>= 50% Threshold	>= 75% Threshold
Per-Capita GDP	0.7731 ***	0.7788 ***	0.7712 ***	0.7966 ***
	(0.0670)	(0.0659)	(0.0676)	(0.0669)
Fraction w/ Phone Lines	2.8210	2.9087	3.1316	2.8206
	(3.1422)	(3.1381)	(3.1445)	(3.0953)
Literacy Rate	-6.0545 ***	-5.788 ***	-5.9029 ***	-5.7771 ***
	(1.5298)	(1.5220)	(1.5356)	(1.5220)
Gini Coefficient	-4.0443	-3.7453	-3.9834	-4.2543 *
	(2.6012)	(2.5799)	(2.6084)	(2.5832)
Normalized Internet Price (1998)	-12.3528	-81.5323	-59.1639	-50.5158
	(375.8801)	(378.3368)	(380.4748)	(381.7518)
Normalized Internet Price (2000)	713.6187	598.6366	632.6763	644.2572
	(555.5750)	(549.7171)	(556.8402)	(559.2102)
Normalized Internet Price (2001)	12.2793	10.7086	11.1932	12.0481
	(11.0013)	(11.0140)	(11.0732)	(10.7540)
Population Density	0.5686 ***	0.5509 **	0.5639 ***	0.5691 ***
	(0.2171)	(0.2172)	(0.2149)	(0.2207)
Age Below 20	-13.4641	-11.5104	-10.1405	-6.3769
	(20.3899)	(20.5111)	(20.5016)	(20.5478)
Age Between 20 and 65	-0.1214	0.9313	1.8698	5.4073
	(16.7895)	(16.7424)	(16.8528)	(16.9070)
Age Above 65	6.1366	8.3259	11.4918	19.0471
	(30.8459)	(31.1071)	(31.0269)	(31.2219)
Fraction Urban Population	1.5707	2.2299	1.8168	0.5578
	(1.8925)	(2.0455)	(2.0130)	(1.9558)
Fraction School Enrollment	6.4917 ***	6.3969 ***	6.3411 ***	6.0043 ***
	(1.9605)	(1.9641)	(1.9500)	(1.9307)
Civil Liberties Index	0.3538 **	0.3321 *	0.3422 **	0.3514 **
	(0.1727)	(0.1719)	(0.1708)	(0.1705)
Cost of Telephone Call	-0.2087 *	-0.2082	-0.2056	-0.1945
	(0.1268)	(0.1271)	(0.1286)	(0.1307)
Language-Content Interaction	0.1179 ***	0.1257 ***	0.1185 ***	0.1668 ***
	(0.0276)	(0.0269)	(0.0268)	(0.0158)
Language-Content Below Threshold	2.7141 *	1.3157 *	0.8096	-0.6004
	(1.5322)	(0.7797)	(0.7418)	(0.7182)
R-Squared	0.6789	0.6785	0.6773	0.6845
Observations Above Threshold	681	188	102	76

Robust standard errors in parentheses. * = 10% significance, ** = 5% significance, *** = 1% significance.
Year dummies and dummy variables for missing values of all variables included in all regressions. Countries with non-mising langauge-content interaction are listed in Table 8.

Table 8

Countries/Territories' Language-Content Interaction Included in Analysis for Each Threshold Level

Africa[1]	The Americas[1]	Asia[1]	Europe[1]	The Pacific[1]
Angola*	Antigua and Barbadu*	Armenia*	Albania	Fiji*
Benin*	Aruba*	Azerbaijan*	Andorra**	French Polynesia*
Botswana*	Bahamas*	Bahrain***	Belarus*	Guam*
Burkina Faso*	Barbados*	Bhutan	Belgium**	Kiribati*
Burundi*	Belize**	Brunei Darussalam*	Bosnia and Herzegovina	Marshall Islands*
Cameroon*	Bermuda****	Cambodia*	Bulgaria*	Micronesia*
Cape Verde Islands*	Bolivia**	Cyprus*	Croatia	New Caledonia**
Central African Republic*	Costa Rica****	Georgia*	Czech Republic*	New Zealand****
Chad*	Dominica*	Indonesia*	Denmark*	Papua New Guinea*
Comoros*	Dominican Republic****	Iran*	Estonia**	Samoa*
Congo*	El Salvador****	Israel**	Finland*	Solomon Islands*
Cote dIvoire*	French Guiana*	Jordan****	Greece*	Tonga*
Democratic Republic of the Congo	Greenland	Kazakhstan**	Hungary*	Vanuatu*
Djibouti**	Grenada*	Kuwait**	Iceland	
Equitorial Guinea*	Guadeloupe*	Kyrgyzstan**	Ireland***	
Eritrea*	Guatemala**	Laos	Italy*	
Gabon*	Guyana*	Lebanon****	Latvia**	
Ghana	Haiti*	Malaysia*	Lithuania*	
Guinea-Bissau*	Honduras****	Maldives*	Luxembourg**	
Lesotho*	Jamaica*	Mongolia*	Macedonia	
Liberia*	Martinique*	Nepal*	Malta*	
Libya****	Netherlands Antilles*	Oman***	Moldova*	
Madagascar*	Nicaragua****	Pakistan*	Netherlands*	
Mali*	Panama****	Palestinian West Bank and Gaza***	Norway*	
Mauritania****	Paraguay*	Philippines*	Poland*	
Mauritius*	Puerto Rico****	Qatar*	Romania*	
Mozambique*	Saint Kitts & Nevis*	Singapore***	Slovakia*	
Namibia*	Saint Lucia*	South Korea*	Slovenia*	
Reunion*	Saint Vincent & the Grenadines*	Sri Lanka*	Sweden*	
Rwanda*	Surimane*	Tajikistan*	Switzerland*	
Sao Tome e Principe*	Trinidad & Tobago*	Thailand*	Ukraine*	
Seychelles*	Uruguay****	Turkey*		
Sierra Leone*	U. S. Virgin Islands*	Turkmenistan*		
Swaziland*		United Arab Emirates**		
Togo*		Uzbekistan*		
Uganda*		Viet Nam*		
Zambia*				
Zimbabwe*				
# Countries Baseline 38	33	36	31	13
# Countries 0% Threshold 36	32	34	26	13
# Countries 25% Threshold 3	12	11	6	2
# Countries 50% Threshold 2	9	5	1	1
# Countries 75% Threshold 2	9	2	0	1
Ethnologue # Countries 57	51	50	45	25

[1] Classifications according to Ethnologue (2005).

No asterik indicates included only in baseline analysis. * indicates included in baseline and 0% threshold analysis. ** indicates included in baseline, 0% threshold and 25% threshold analyses. *** indicates included in basedline, 0% threshold, 25% threshold, and 50% threshold analyses. **** indicates included in basedline, 0% threshold, 25% threshold, 50% threshold, and 7%% threshold analyses. Regressions also inlcude the following countries/territories with missing language information: Afghanistan, Faroe Islands, Falkland Islands, Hong Kong, Liechtenstein, Macao, Mayotte, Monaco, Myanmar, San Marino, Serbia and Montenegro, and Tuvalu.

Table 9

First-Differences Estimates for All Sample Countries
1998 - 2004 (N = 914)

	> 0% Threshold	>= 25% Threshold	>= 50% Threshold	>= 75% Threshold
Per-Capita GDP	1.7519 ***	1.7483 ***	1.7425 ***	1.7621 ***
	(0.3715)	(0.3715)	(0.3709)	(0.3684)
Fraction w/ Phone Lines	5.1059	5.5189	5.7395	5.7691
	(14.1650)	(13.9471)	(13.9222)	(13.9147)
Fraction School Enrollment	-9.1278 **	-9.2172 **	-9.1281 **	-9.0758 **
	(4.0632)	(4.0801)	(4.0826)	(4.0845)
Civil Liberties Index	-0.0243	-0.0162	-0.0077	-0.0154
	(0.2304)	(0.2336)	(0.2341)	(0.2336)
Cost of Telephone Call	0.0364	0.0380	0.0406	0.0379
	(0.0687)	(0.0690)	(0.0690)	(0.0690)
Language-Content Interaction	0.0470	0.0536	0.0638 *	0.0874 **
	(0.0339)	(0.0348)	(0.0353)	(0.0416)
Language-Content Below Threshold	0.1685	0.0248	0.3207	0.2722
	(0.4578)	(0.2417)	(0.2530)	(0.2779)
R-Squared	0.1825	0.1829	0.1837	0.1847
Observations Above Threshold	519	151	82	61

Robust standard errors in parentheses. * = 10% significance, ** = 5% significance, *** = 1% significance. Year dummies and dummy variables for missing values of all variables included in all regressions. Countries with non-mising langauge-content interaction are listed in Table 8.

Table 10

Effect of Content on Internet Adoption for All Sample Countries
1998 - 2004 (N = 1,082)

	Geographic Density	Linguistic Concentration
Per-Capita GDP	0.7795 ***	0.7673 ***
	(0.0657)	(0.0686)
Fraction w/ Phone Lines	2.0709	3.1973
	(3.1091)	(3.1929)
Literacy Rate	-6.1929 ***	-6.2635 ***
	(1.5116)	(1.4970)
Gini Coefficient	-4.4132 *	-2.6314
	(2.5895)	(2.6967)
Normalized Internet Price (1998)	-54.9229	12.4161
	(382.7916)	(368.6464)
Normalized Internet Price (2000)	657.3793	775.4573
	(566.3965)	(545.7531)
Normalized Internet Price (2001)	10.8806	15.5642
	(10.8637)	(11.1269)
Population Density	0.5018 **	0.3917 *
	(0.2276)	(0.2315)
Age Below 20	-14.9127	-21.7188
	(19.3636)	(20.7680)
Age Between 20 and 65	-1.5367	-3.3591
	(16.2051)	(17.0198)
Age Above 65	4.8404	-9.0799
	(29.0850)	(31.2741)
Fraction Urban Population	1.0497	1.5993
	(1.8279)	(1.7835)
Fraction School Enrollment	6.9233 ***	6.4674 ***
	(1.9117)	(1.9805)
Civil Liberties Index	0.4130 **	0.3423 *
	(0.1723)	(0.1764)
Cost of Telephone Call	-0.2117 *	-0.1885
	(0.1278)	(0.1278)
Language-Content Interaction	0.0241	1.6058 ***
	(0.0527)	(0.4069)
Population Density*Language Content Interaction	0.1497 ***	
	(0.0497)	
Language Herfindahl Index		0.2628
		(1.3994)
Language Herfindahl Index*Language Content Interaction		-1.4983 ***
		(0.4180)
R-Squared	0.6835	0.6882

For Columns 1 and 2, robust standard errors in parentheses. * = 10% significance, ** = 5% significance, *** = 1%significance. Standard errors in Column 3 are unadjusted [This will be done in a future version of the paper.] Year dummies and dummy variables for missing values of all variables included in all regressions. Countries with non-mising language--content interaction are listed in Table 8.

Figure 1