Making the 'Next Billion' Demand Access*

The Local-Content Effect of `google.co.za` in Setswana

Bastiaan Quast†

Abstract

Recent attempts to connect the current 'next billion' to the Internet in places such as sub-Saharan Africa have not met expectations. In places where Internet infrastructure has come online and prices have gone down, the expected consequent increase in uptake was not observed. I develop a framework that incorporates language in the the two-sided markets framework, viewing differences as transaction costs. As a result of the cross-side network effects, it is difficult to isolate the causal effect of one on the other. The exogenous introduction of the Setswana language interface on the South African Google Search website was a spillover of the development of that interface for the Botswanan Google website. This exogenous improvement in the accessibility of Setswana-language content has resulted in a substantial increase in the number of native Setswana speakers coming online and owning personal computers. This is turn has also led to increased usage of the Setswana language online. This adoption appears to also lead to improvements in employment.

1 Introduction

Internet uptake is a two-sided market, with users on one side and content creators on the other side. Positive cross-side network effects mean that increases in content leads to increases in user adoption and visa versa. This market exists separately for each language, however for many indigenous languages this virtuous circle fails to start properly, keeping usage and content levels low.

With this study I seek to answer the question whether an increase in local language content, does indeed lead to an increase in uptake of Internet usage among native speakers of this language.

Because of the cross-side network effects in a two-sided market, any observed changes are inherently endogenous. I remedy this problem by using an exogenous shock in accessibility of Setswana language content in South Africa, namely the introduction of the Google Search interface in Setswana. I find that this leads to a strong increase in both the proportion of households reporting to have spent on Internet access in the last 30 days, as well as individuals owning a

*http://qua.st/internet-access

†http://qua.st/ | bquast@gmail.com | bastiaan.quast@graduateinstitute.ch
Maison de la paix, Geneva, Switzerland

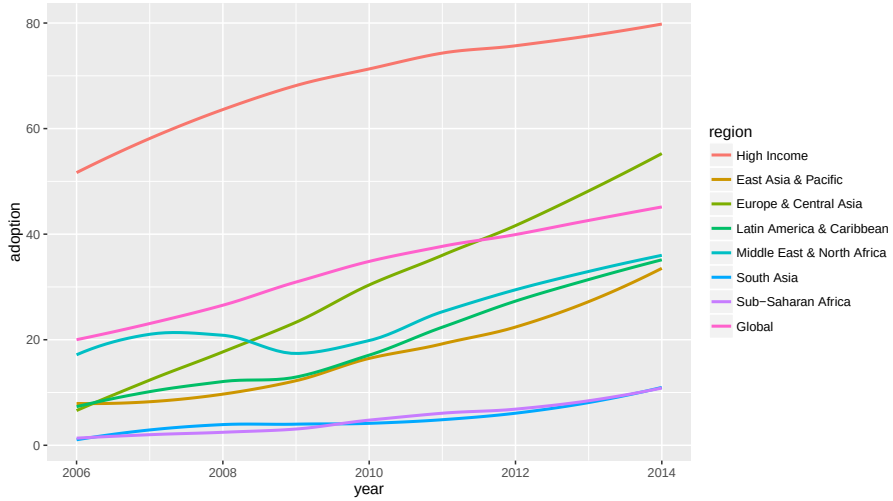
computer. This in turn has led to a large increase in the usage of the Setswana language online (in Google search queries). There also appears to be a strong improvement in employment status among individuals who spend on Internet or own a computer after the introduction of the interface. Suggesting that the expanded demographic of Setswana Internet users is benefiting from increased Internet adoption in terms of employability.

The term 'Connecting the Next Billion' was introduced in The Economist's 2006 'End of Year Report' (Standage, 2006), discussing the infrastructural requirements for connecting the second billion individuals to the Internet. Since then, close to 2 billion people are estimated to have been connected to the Internet, up from the just over one billion at the time of writing (Sanou, 2015). However, it seems increasingly unlikely that the current 'Next Billion' will be connected as easily as the previous ones.

In the period 2010-2014 the average annual growth of Internet bandwidth in sub-Saharan Africa was over fifty percent. This increased bandwidth also causes downward pressure on the cost of Internet access, which brought the sub-Saharan average cost of a 500MB prepaid Internet bundle down to around $10, increasingly putting it within range of the emerging middle classes. Yet, despite increased range and improved affordability, sub-Saharan Africa is showing stagnation in the growth of Internet connected individuals.

As is shown in figure Figure 1, growth of Internet usage in Sub-Saharan Africa is rapidly decreasing. Unlike in other regions in the figure, this observed stagnation is not a consequence of near market saturation, as adoption levels are still relatively low.

Figure 1: Internet Adoption



A crucial factor in Internet adoption by native speakers of a language is the interplay with content creators using that language. This dynamic is know as a two-sided market, which is characterised as having two different sides, which exhibit positive cross-side network effects (Parker and Van Alstyne, 2000a,b, 2005).

1.1 Two-Sided Markets

In the case of Internet adoption in a certain language, these sides are on the one hand the content creators, such as news websites and on the other hand the content consumers, or Internet users. Ideally, adoption should follow a virtuous circle, whereby the content offering encourages more users to come online, which in turn incentivises more content creation and so forth (Rochet and Tirole, 2003, 2006). Unfortunately this virtuous circle sometimes fails to properly start for certain languages, this is especially important as it typically concerns communities who's linguistic characterisation can already hamper economic growth (Arcand and Grin, 2013). Herein also lies the difficulty with finding empirical evidence supporting these dynamics, since the process of adoption by users and content creators is inherently endogenous. With this paper I seek to empirically address the question if increased accessibility of content does indeed lead to an increase in Internet adoption.

By considering that these markets are shaped by the protocols they agree upon and by also viewing language as such a protocol, it emerges that the two-sided market for Internet exists separately for each language. We can model the differences in protocols (and language) as transaction costs, if the linguistic barrier is significant enough, the transaction is not conducted and agents leave the market.

A two-sided market can be defined as an economic platform with two distinct sets of agents (in this case Internet users and content creators), that provide each other with network effects, meaning that they provide the other side with greater value for entering the market.

The model of a two-sided market is a specific case of a market with simple network effects, such as the market for telephones. In the market for telephones, all users act as both creators and consumers of content, by speaking and listening respectively. The more other people have telephones, the greater the benefit derived from owning one is, since it can connect to more people (simple network effects).

The two-sided market is a specific case, since it restricts a set of agents to one role (e.g. transmitting) and the other agents to a complementary role (e.g. receiving). The network effects here manifest as a sort of seesaw, more transmitters means greater utility is derived from being a receiver, and more receivers in turn means that greater utility is derived from being a transmitter, etcetera. These are called cross-side network effects and a market that exhibits them is called a Two-Sided Market (2SM).

Protocol Stack / Convention Ensemble

Although often discussed as such, these markets and platforms do not existing in isolation, but in fact build upon each other in a value chain (much like other markets).

The Internet itself is a set of agents (computers) that are connected to each other via a protocol called the Internet Protocol (IP). Viewed jointly, this can be seen as a platform with simple network effects, since additional computers in the network communicate bidirectionally. Therefore, more computers increase the utility of owning a computer, as it can communicate with more other computers and their users. In this market, there is no single intermediary, since users can

connect through any ISP and will still be able to connect to any other computer on the Internet.

Table 1: Protocol Stack

Social Media (simple network effects, proprietary)	Spam (one-way network effects)
Websites (cross-side network effects)	Email (simple network effects)
TCP/IP (simple network effects)	

This platform is built upon by using additional protocols so that eventually communication is conducted using a combination of protocols called the protocol stack (in economic terms we might call this a "convention ensemble"). For instance, emails are sent through protocols such as the Simple Mail Transfer Protocol (SMTP), which builds upon the Internet Protocol. Of course, proprietary variants of such additional platforms with simple network effects also exists. For instance Twitter or WhatsApp behave similar to email in many ways, but in both cases, there is only one single intermediary that users can use to connect to the network.

These additional protocols can change the nature of the network effects, e.g. from simple network effects to the cross-side network effects of a two-sided market. For instance, the WWW (websites) uses the HTTP protocol to transmit files with HTML content. In simplified view, the HTTP protocol is designed to transmit websites from servers to users. In this case, we thus distinguish between agents that send content (servers/websites) and agents that receive content (users/browsers), the two sides benefit from cross-side network effects, since more users means a greater audience to websites and more websites means more content to consume for users.

However, these additional protocols need not change the nature of the market. For instance, agents that use email for communication will both send and receive emails, making it a platform with simple network effects, just like the Internet itself at the IP level. Particularly, since users can chose from many email providers and still communicate with all other users.

General View of Protocols

We can now use this approach of analysing the nature of the network effects at every additional level of protocols to examine the role of language. We return to the example of the WWW, which is a two-sided market based on the fact that the HTTP basically works to transmit websites from servers to users.

These technical protocols enable the communication between the various computers connected to the Internet in various ways. However, essentially these computers merely fulfil a role in enabling human communication. As such other protocols of that human-to-human communication also play a role, chief among those is the language of writing (or speaking).

By moving beyond the technical-only interpretation of protocols, and including language, we see that the additional protocols again lead to the creation of distinct platforms. It is thus the case that the two-sided market of the WWW is actually composed of many two-sided markets in different languages (or non-linguistic communication).

Transaction Costs

We can interpret these differences in protocol as a form of transaction costs (Coase, 1937). Transaction costs have been broadly defined by Steven N. S. Cheung as: "any costs that are not conceivable in a "Robinson Crusoe economy". We can interpret this as costs that stem from the presence of institutions or conventions.

For instance, consider the way in which people answer a fixed-line telephone. Since the phone is shared by the household, the person calling does not know who will answer the phone, it is therefore customary to state your name when you answer the telephone. With the rise of mobile phones, people who did not have mobile phones or used them less expected people to answer them the same way you answer a fixed-line phone, by stating your name, but people who did use mobile phones quickly realised that it could only logically be expected that they would be the one answering the phone, eliminating the need to state your name and use just "hello" instead. This was a partial breakdown of protocol. We can interpret this as transaction costs, if the person calling was sufficiently confused by not being told who they are speaking to, he/she might hang up the phone. Alternatively, they might feel offended and irritant, making the conversation less effective or possibly cutting it short.

Similar things apply to technical protocols. For instance, consider the introduction of something like a new HTML tag, this tag will be interpretable by new browsers that support these tags, but not by older browsers, causing the older browsers to misrender part of the page. Depending on the gravity, this might make the page harder to view, or simply impossible. Similarly, old tags are sometimes deprecated (e.g. IE6 specific tags), which means that they can no longer be interpreted by newer browsers (e.g. IE10), giving similar results. The transactions costs might thus be sufficient to stop communication, or simply hinder it. If the hindering continues, it might be that the users consumers a smaller amount of content.

Language can also be interpreted as a protocol or convention for communication and we can therefore also interpret linguistic differences as a transaction cost.

Intuitively we might think of language as a discrete variable, perhaps making transactions costs a poor way to approximate this. However, there are many cases in which linguistic differences are a definite barrier, but not necessarily a breaking point. For instance, a Portuguese speaker might be able to read a Spanish website if Portuguese is not available, but they would always have less effort in reading Portuguese (and thereby consume more). This transaction cost would increase if we consider for instance an Italian speaker reading Spanish. Moving to a French speaker reading Spanish, which is really quite distinct, but nevertheless some similarity in terms of vocabulary. Moving to a different language family, such as German, the differences probably become insurmountable. For a Japanese speaker not a single word would be meaningful. This is especially the case, as the Japanese speaker in question would not understand the latin script.

If we assume that the Internet's value only stems from communicating with other people, then we might define a utility function like this:

$$u_i() = -p + \sum_{j=1}^{n} d_{ij} l_{ij} v_{ij} c_j \tag{1}$$

Whereby p represents the cost of connecting to the Internet and the the other term is the value of all possible connections to all individuals. Here, v_{ij} represents the value of the connection between individual i and j, to individual i. This is multiplied with the l_{ij} term, which is the likelihood that individuals i and$j$ will establish a connection given that they both have Internet connectivity. This is again multiplied by the d_{ij} term, which represents the difficulty of communication between individuals i and j, Finally, this is multiplied with the decision of j to connect to the Internet or not c_j ,where $c_j \in 0, 1$. Of course this decision in turn depends on the utility of connecting to j, where $w_j \in 0, 1$. The term w_j represent whether it is worth it or not to connect:

$$c_j() = w_j(-p + \sum_{k=1}^{n} d_{jk} l_{jk} v_{jk} c_k) \tag{2}$$

Which means if:

$$p < \sum_{k=1}^{n} d_{jk} l_{jk} v_{jk} c_k \tag{3}$$

then $w_j = 1$, otherwise $w_j = 0$.

The move from simple network effects to the cross-side network effects is made when the ease of communication variable d_{ij} stays high, but d_{ji} becomes very low (e.g. HTTP). This means that is would be relatively easy for a certain type of agent (server) to communicate something to another type of agent (Internet user), but that reverse communication is less easy. In the context of language we could consider a Spanish speaker who could communicate something to a Portuguese speaker, if the Portuguese speaker has some passive understanding of Spanish. But that this communication cannot be reversed, since the Portuguese speaker does not have an active understanding or Spanish and the Spanish speaker does not have a passive understanding of Portuguese. Note that a market can also become two-sided if the information being communicated is different (e.g. buy vs. selling a product).

When communication breaks down in both directions, then markets start to exist parallely. For instance, in the above example, if the Portuguese speaker loses his passive understanding of Spanish, then they cease to operate in a common market, rather they each operate in a separate - parallel - market, in this case a Portuguese-language market for communication and a Spanish-language market for communication.

The nature of these parallel markets is again determined by the protocols that they use. For instance, the market of Portuguese-language websites, is a two-sided market, based on the fact that the underlying HTTP protocol functions so.

These transaction costs could possibly explain why the automatic optimal allocation of resources (according to the Coase theorem) does not allocate any to groups with certain languages (even if the resource - i.e. space on the Internet - is virtually unlimited).

2 Data

The National Income Dynamics Study (NIDS) collects data on a representative set of around ten thousand South African households across several time periods. The first wave was gathered in 2008, the second wave in 2010, and the third wave was gathered in 2012 (Southern Africa Labour and Development Research Unit, 2008, 2012, 2013).

In addition to this, in may 2016 a fourth wave of data has also been partially published, which I used to relate the expanded demographic of Setswana speaking Internet users to improvements in employment levels.

The dataset contains an extensive household questionnaire, which contains detailed information on income and expenditure. In particular, it breaks household expenditure down into many forms of food and non-food expenditure, one of which is household expenditure on Internet access in the last 30 days. In addition to this, the household income is calculated and imputed with other income such as home ownership. The individual (adult) questionnaires also contain information on linguistic skills in both English and in the interviewees native language, as well as a series of variables relating to communication technology ownership and utilisation, in particular computer ownership. In table Table 5 an overview of the dependent variables, broken down by wave and native language is presented.

Table 2: Dependent Variable Descriptive Statistics

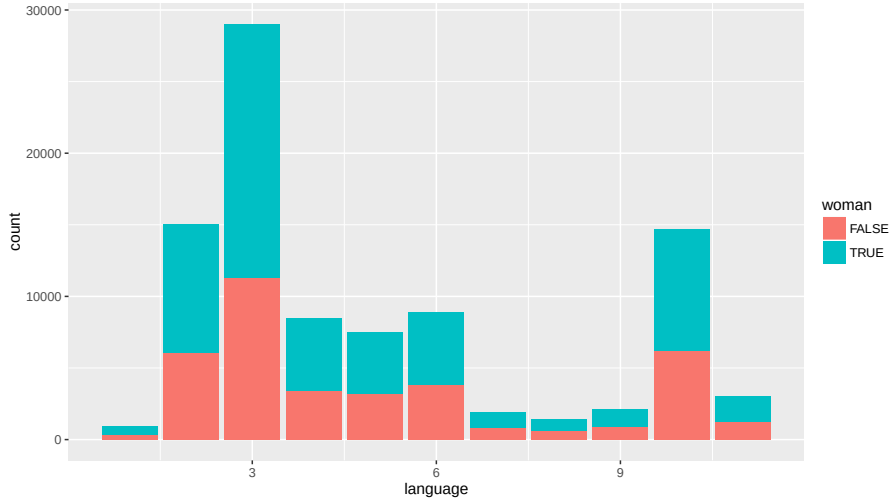
	Wave	Setswana	Other
Computer Ownership	1	0.0351	0.0567
	2	0.0359	0.0417
	3	0.0711	0.0632
Household Internet Expenditure	1	0.0068	0.0162
	2	0.0037	0.0224
	3	0.0106	0.0140

I use both household expenditure on Internet access and computer ownership as dependent variables. Household expenditure includes a variety of way in which this expense can be made, including paying for a fixed-line subscription, a mobile Internet subscription, as well using Internet in an Internet cafe. More than 99% of South Africa is covered by mobile Internet networks, this figure did not change during the time periods used in this study (Union, 2015). Computer ownership is recorded in the individual adult questionnaire, which provides me with more observations, unlike expenditure in the last 30 days, this shows a more long-term investment.

In addition to the variables of interest, I include a number of relevant covariates, such as household income and education levels. Furthermore, I include information on English and native language reading and writing skills. Around 45% of the individuals report being able to read and write English fluently, whereas around 55% report being able to do so in their native language.

In Figure 2 we can see the number of native speakers for each language in the dataset, coloured by gender. The dataset contains a total of 51,612 observations (adult individuals), of which 2,806 are female native Setswana speakers and 2,140 are male native Setswana speakers.

Figure 2: Native Language and Gender



3 Empirical Methodology

With this paper I aim to answer the question whether increased content or accessibility of content leads to an increase in demand. This section begins with a discussion of the identification strategy employed, followed by a explanation of the estimator used to operationalise this.

3.1 Identification Strategy

This paper exploits the introduction of the Setswana interface language to Google Search in South Africa as a spillover of the development of that interface for the Botswanan Google Search website. By comparing the number of native Setswana speakers in South Africa being Internet users, with the number of South Africans with a different native language around the same time, I isolate the effect of this introduction.

The Setswana language interface was first developed for the Botswanan Google Search website (`google.co.bw`). As such, the introduction of Setswana to the South African Google Search (`google.co.za`) was a spillover effect of that development. This allows me to rule out any possible endogeneity issues that might otherwise arise in contexts such as these. For instance, the Afrikaans language is almost solely spoken in South Africa. When we observe that the introduction of the Afrikaans Google Search interface occurs around the same time as a growth in the number of native Afrikaans Internet users, it will be hard to isolate the effect from the introduction from its cause (since an increase in native Afrikaans Internet users would be a good reason to introduce it as an interface language).

Substantial numbers of Setswana speakers exist in Botswana, South Africa, Zimbabwe, and to some extend Namibia. However, the language is most important in Botswana, where it is spoken by approximately 80% of all people, and where it is the only official language other than English. As such, it is

also the place where most linguistic work on the Setswana language takes place. The Setswana Google Search interface was also developed at the University of Botswana by prof. Otlogetswe.

It is worth noting that it is very common to not personally own a computer, therefore 'paying for Internet access' also includes a lot of people who use the Internet in other locations such as Internet cafe's.

In addition to using the propensity to spend on Internet (in the last thirty days), I also use the propensity to own a computer as a dependent variable.

3.2 Regression Specification

As mentioned in the above section, I compare the change in the level of Internet users among native Setswana speakers in South Africa, with that of native speakers of other language in South Africa around the introduction of the Setswana interface to the South-African Google Search, after the second wave of the NIDS. For this I use a Difference-in-Differences estimator (Abadie, 2005; Imbens and Jeffrey M. Wooldridge, 2009) using a native-Setswana speaker dummy variable (`setswana`), interacted with an event dummy variable (`event`). The former is `TRUE` when the native language of the individual (`a_lng`) is `Setswana` and `FALSE` otherwise. The latter is `FALSE` for data collected prior to the introduction of the Setswana interface language (late 2010, here wave 1 and 2) and `TRUE` after this introduction (here wave 3). The model than takes the form as described in equation (4).

$$y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \beta X_{it} + \epsilon_{it} \tag{4}$$

Where α_i represents the individual fixed effects, λ_t represent the time fixed effects, and X_{it} are the time varying covariates. The ϵ_{it} is the error term. Finally the term of interest is D_{it} which represents the treatment effect.

The `h_nfnet` variable is recorded at a household level, as such a standard error correction needs to be applied (White, 1980), the model with standard error corrections is reported in the appendix.

Lastly, the dependent variables are both logical or binary variables, as such, normally a model such as logit should be used. However, since I am using Difference-in-Differences, this model would be undefined (Jeffrey M Wooldridge, 2010), I therefore use a standard linear model.

4 Results

In the base model, I use an interaction of the `event` dummy and `setswana` dummy in order to isolate the effect on the explanandum, a dummy variable describing household expenditure on Internet in the last thirty days or not (`Internet_expenditure`, household non-food Internet). The results of this estimation are presented in table Table 3.

I find that the interaction term of the event dummy (`event`) and the native Setswana speaker dummy (`setswana`) is positive and highly significant, with a p-value around `0.0018`. Both the individual dummy variables (`event` and `setswana`) yield significant negative parameter estimates.

In addition to this, the covariates included in the estimation are also highly significant. The highest education level of the individual (`best_edu`) and the

household income (hhincome) are both positive and significant. The parameter estimate of woman here is negative but not at all significant, this is unsurprising as I use Internet expenditure at a household level. Most women live in a household which includes men and visa versa, suggesting that this effect cannot be isolated in this estimation. I further investigate this issue in a separate estimation discussed below. The variables describing linguistic skills in reading and writing in both English and the native language do yield many significant results, though lower levels of English writing skill seems to be correlated with a lower propensity to use the Internet (a_edlitwrten for levels 2 and 3, but not the very lowest: 4).

In an alternative formulation, I include the native language variable as a categorical variable (language), interacted with the event dummy. In this estimation I only find significantly positive results for Setswana and Venda (as small language from the region bordering Zimbabwe), and a significantly negative effect for the language Afrikaans.

When using the propensity of adults (own_computer) to own a computer is used as an explanandum, I find similar results. This is of particular relevance, as the explanandum here (own_computer) differs from the base model's explanandum in two ways. Firstly, it does not include expenditure on Internet in ways such as Internet cafes, but focusses on actual ownership, signalling a more long-term investment and interest. Secondly, the Internet_expenditure variable is at a household level, whereas the own_computer variable is at the level of an individual adult. The results from this estimation are included in table Table 3. This form of the estimation yields similar results to those estimated in the base model. Firstly I find that the variable of interest, the interaction term between the event and the language dummy (event * setswana) is positive at 0.024 and highly significant, with a p-value smaller than 0.001. This means that the event increased the propensity to own a computer by 2.4%. The individual dummy variables (event and setswana) again are significant and negative with the former's p-value smaller than 0.01 and the latter's smaller than 0.001. In terms of the linguistic skill, I find that the lower levels of English reading as well as English writing are correlated with lower propensities of computer ownership. Similar to Internet expenditure model, household income (hhincome) and highest level of education (best_edu) are both positive and highly significant (p-value: ~0). However, unlike in the household Internet expenditure model, the gender of the individual here is highly significant, specifically, parameter estimate of woman is negative and highly significant (p-value: ~0). As mentioned above, this variable is difficult to interpret when using a household-level variable as an explanandum, however, here, the computer ownership variable is at an individual level, which makes the coefficient more easily interpretable.

The below figure present a graphical illustration of the above mentioned result. As we can see, initially, the propensity to own a computer for setswana speakers (blue line) was similar to the average of speakers of other languages (red line), however after the introduction of the Setswana language search interface, between waves 2 and 3, we observe a sharp increase in this propensity for setswana speakers in wave 3, whereas average of other languages remains relatively unchanged.

Table 3: Internet Access and Computer Ownership

	Internet	(P > |t|)	Computer	(P > |t|)
event * setswana	**0.012**	0.00	**0.024**	0.00
event	**-0.012**	0.00	-0.054	0.01
setswana	**-0.014**	0.00	**-0.015**	0.00
income	**0.000**	0.00	**0.000**	0.00
woman	-0.001	0.23	**-0.023**	0.00
education	**0.001**	0.00	**0.006**	0.00
Observations	47665		46464	

Figure 3: Computer Ownership Setswana

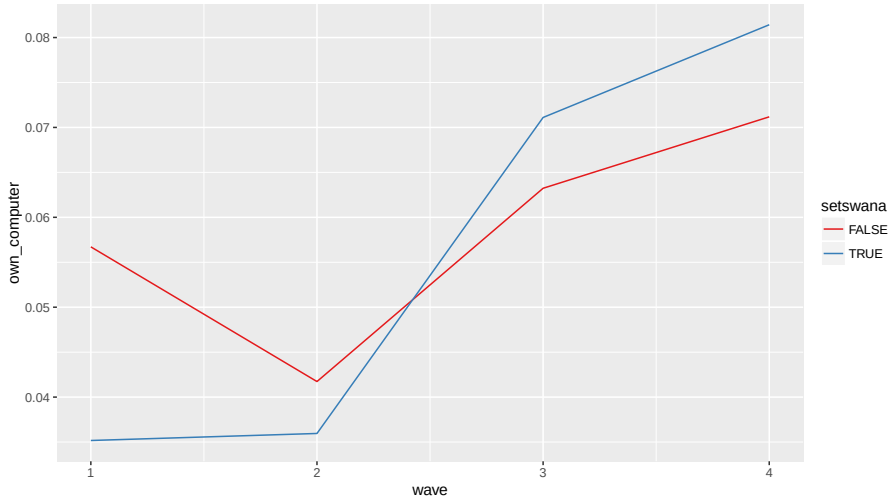
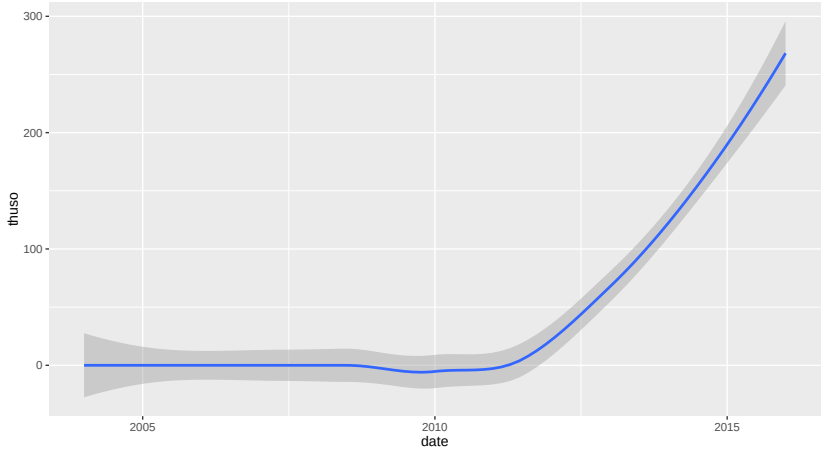


Figure 4, illustrates how usage of the Setswana word 'thuso', meaning 'help' in search queries, was zero prior to the introduction of the Setswana search interface and became widespread thereafter. Therefore, in addition to the increased Internet usage by Setswana speakers, we can also observe an increased usage of the Setswana language itself. This can lead to greater amounts of Setswana language content being found and engaged with, which in turn incentivised content creators to provide more Setswana language content. This could help break the vicious circle of low levels of content and few users.

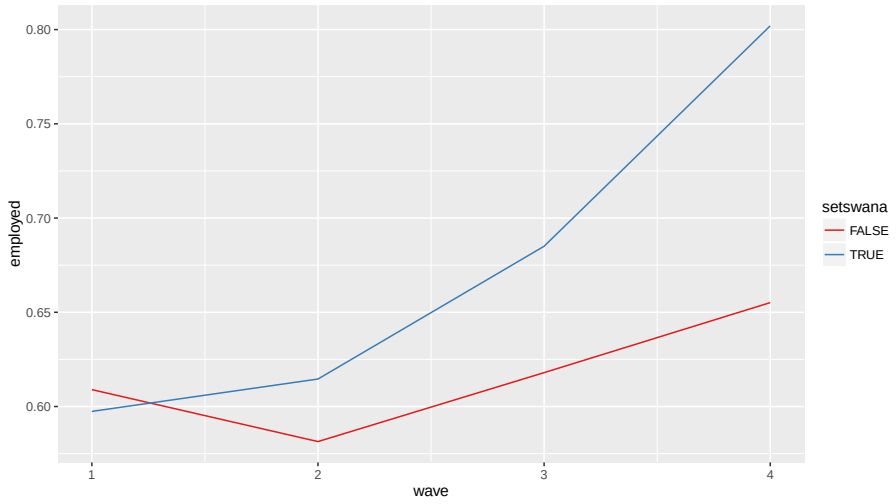
Figure 4: Usage of Setswana Words on Google.co.za



As mentioned in section §section 2 I use the partially released 4th wave of the dataset, which covers the year 2014 in order to relate the increased Internet usage with employment levels. I do this by subsetting data to include one the one hand, only the individuals that owned a computer in after the event in wave 3 and on the other hand the individuals in households that reported spending on internet access. The idea is that the owners/users in wave 3 are an expanded demographic for Setswana speakers but remain relatively unchanged. With this, we can see the evolution of the employment level for this expanded demographic.

In the below figure, I plot the employment status of individuals that own a computer in the first wave after the introduction of the Setswana-language interface (wave 3). The figure shows that there is a sharp uptick in the proportion of employed individuals among Setswana speakers.

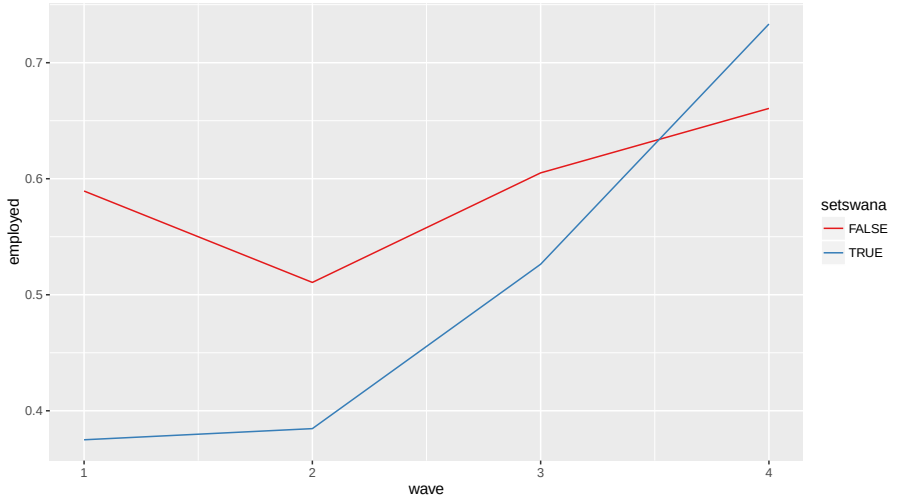
Figure 5: Employment for Individuals who Own a Computer in Wave 3



Similarly, the individuals that lived in a household that spent on Internet

access in the last 30 days also saw a strong increase in the proportion of employed individuals, overtaking the proportion for the rest of the population.

Figure 6: Employment for Individuals with Internet Expenditure in Wave 3



5 Conclusions and Limitations

In conclusion, despite recent advances in the reach, speed, and affordability of Internet connectivity in sub-Saharan Africa, actual uptake has been stagnant. Internet adoption is a two-sided market, as a result, one aspect of this is that users benefit from cross-side network effects from content. Unfortunately, in certain languages, a sufficient amount of content is not always created, as a result, Internet adoption among native speakers of these languages can lag behind.

This paper demonstrates that this failure can in part be attributed to the dynamics of a two-sided market, whereby the vicious circle of few users and little content perpetuates a situation of low levels of adoption. Due to the endogenous nature of two-sided markets, there are few methods of isolating a causal effect.

I exploit the introduction of the Setswana-language interface on `google.co.za`, as a spillover of the development of this interface for `google.co.bw`, I find that it leads to a substantial increase in Internet usage and computer ownership among native Setswana speakers.

Furthermore, when comparing the Setswana speakers that own a computer or spend on Internet access after the event, with non Setswana speakers, we see a marked increase in the proportion of employed individuals over that among the rest of the population. This increase is even stronger in wave 4 of the data, suggesting that it takes some time for the effect to fully materialise and that it is persistent.

This increase of Internet usage among the Setswana speaking population as a result of the newly introduced interface language on `google.co.za`, suggests that there is a serious lack in the availability of local content in many indigenous African languages, which serves as an impediment to further Internet adoption.

This suggests that the effect is unlikely to be ephemeral in nature, since computer ownership constitutes a more long-term investment in Internet access.

As I discussed in the results section, the increase in computer ownership for Setswana speakers between wave 2 and wave 3, was 115%, compared to 70% for the rest of the population, or a 50% greater increase. The effect on Internet expenditure in the household, which includes expenditure in Internet cafes etc. is even greater. The proportion of Setswana households that spent on Internet in the last 30 days increased by 217%, whereas for the rest of the population it fell be 22%.

References

Abadie, Alberto

2005 "Semiparametric difference-in-differences estimators", *The Review of Economic Studies*, 72, 1, pp. 1–19.

Arcand, Jean-Louis and Francois Grin

2013 "Language in Economic Development: Is English Special and is Linguistic Fragmentation Bad?", in *English and development: Policy, pedagogy and globalization*, ed. by Elizabeth J. Erling, Multilingual Matters, chap. 11, pp. 243–266.

Coase, Ronald H

1937 "The nature of the firm", *economica*, 4, 16, pp. 386–405.

Croissant, Yves

2013 "pglm: Panel Generalized Linear Model", *R package version 0.1-2*, `http://CRAN.%20R-project.org/package=%20pglm`.

Croissant, Yves, Giovanni Millo et al.

2008 "Panel data econometrics in R: The plm package", *Journal of Statistical Software*, 27, 2, pp. 1–43.

Git Team

2016 *Git: Software Code Manager*, 137 Montague ST STE 380, Brooklyn, NY 11201-3548, `http://www.git-scm.org/`.

Hoekwater, Taco, Hartmut Henkel and Hans Hagen

2016 *LuaTeX*, `http://www.luatex.org/`.

Imbens, Guido W. and Jeffrey M. Wooldridge

2009 "Recent Developments in the Econometrics of Program Evaluation", *Journal of Economic Literature*, 47, 1 (Mar. 2009), pp. 5–86, DOI: `10.1257/jel.47.1.5`, `http://www.aeaweb.org/articles/?doi=10.1257/jel.47.1.5`.

Knuth, Donald Ervin

1984 "Literate programming", *The Computer Journal*, 27, 2, pp. 97–111.

Lamport, Leslie

1985 *I1 (\ LaTeX)—A Document*, pub-AW, vol. 410.

Leisch, Friedrich

2002 "Sweave: Dynamic generation of statistical reports using literate data analysis", in *Compstat*, Springer, pp. 575–580.

LyX Team

2016 *LyX*, Free Software Foundation, Inc., 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301, USA, `http://www.lyx.org/`.

Parker, Geoffrey G and Marshall W Van Alstyne

2000a "Information complements, substitutes, and strategic product design", in *Proceedings of the twenty first international conference on Information systems*, Association for Information Systems, pp. 13–15.

2000b "Internetwork externalities and free information goods", in *Proceedings of the 2nd ACM Conference on Electronic Commerce*, ACM, pp. 107–116.

2005 "Two-sided network effects: A theory of information product design", *Management science*, 51, 10, pp. 1494–1504.

R Core Team

2016 *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, `http://www.R-project.org/`.

Rochet, Jean-Charles and Jean Tirole

2003 "Platform competition in two-sided markets", *Journal of the European Economic Association*, 1, 4, pp. 990–1029.

2006 "Two-sided markets: a progress report", *The RAND journal of economics*, 37, 3, pp. 645–667.

Sanou, Brahima

2015 "The World in 2015: ICT facts and figures", *International Telecommunications Union*.

Southern Africa Labour and Development Research Unit

2008 *National Income Dynamics Study, Wave 1*, version 5.3, `http://www.nids.uct.ac.za/home/`.

2012 *National Income Dynamics Study, Wave 2*, version 2.3, `http://www.nids.uct.ac.za/home/`.

2013 *National Income Dynamics Study, Wave 3*, version 1.3, `http://www.nids.uct.ac.za/home/`.

Standage, Tom

2006 "Connecting the next billion", *The Economist-The World in 2006*, p. 117, `http://www.economist.com/node/5134746`.

Union, International Telecommunication

2015 *The World in 2011: ICT Facts and Figures*, ITU.

Venables, William N and Brain D. Ripley

2013 *Modern applied statistics with S-PLUS*, Springer Science & Business Media.

White, Halbert

1980 "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", *Econometrica: Journal of the Econometric Society*, pp. 817–838.

Wooldridge, Jeffrey M

2010 *Econometric analysis of cross section and panel data*, MIT press.

Xie, Yihui

2015 *Dynamic Documents with R and knitr*, Chapman and Hall/CRC, vol. 29, ISBN: 978-1498716963, http://yihui.name/knitr/.

Zeileis, Achim

2004 "Econometric Computing with HC and HAC Covariance Matrix Estimators", *Journal of Statistical Software*, 11, 10, pp. 1–17, http://www.jstatsoft.org/v11/i10/.

2006 "Object-Oriented Computation of Sandwich Estimators", *Journal of Statistical Software*, 16, 9, pp. 1–16, http://www.jstatsoft.org/v16/i09/.

Zeileis, Achim and Torsten Hothorn

2002 "Diagnostic Checking in Regression Relationships", *R News*, 2, 3, pp. 7–10, http://CRAN.R-project.org/doc/Rnews/.

6 Clustering

Table 4: Clustering

```
library(plm)     # panel linear model estimation
library(lmtest) # Standard Error corrections
library(broom)  # output formatting using tidy()

# specify panel model
plm4_3 <- formula(as.numeric(h_nfnet)  ~ interface_intro*setswana  +
                                         factor(a_edlitrden)  +
                                         factor(a_edlitwrten) +
                                         factor(a_edlitrdhm)  +
                                         factor(a_edlitwrthm) +
                                         a_woman              +
                                         hhincome             +
                                         best_edu              )
# estimate
plm4_3e <- plm(plm4_3, data=pNIDS, model='within')

## Error in row.names(data):  object 'pNIDS' not found

# correct errors
tidy( coeftest(plm4_3e, vcov=vcovHC(plm4_3e,
                                    type="HC0",
                                    cluster="group")) )

## Error in coeftest(plm4_3e, vcov = vcovHC(plm4_3e, type =
"HC0", cluster = "group")):  object 'plm4_3e' not found
```

7 Dependent Variable Breakdown

Table 5: Descriptive statistics on Ownership and Expenditure

a_lng	wave	a_owncom	h_nfnet
IsiNdebele	1	0.0331126	0.0000000
IsiNdebele	2	0.0270270	0.0284091
IsiNdebele	3	0.0333333	0.0000000
IsiNdebele	4	0.0532319	NaN
IsiXhosa	1	0.0112269	0.0012043
IsiXhosa	2	0.0186335	0.0054517
IsiXhosa	3	0.0345622	0.0019763
IsiXhosa	4	0.0372470	NaN
IsiZulu	1	0.0132693	0.0013730
IsiZulu	2	0.0127882	0.0086473
IsiZulu	3	0.0251126	0.0045006
IsiZulu	4	0.0383356	NaN
Sepedi	1	0.0265273	0.0048309
Sepedi	2	0.0226818	0.0021994
Sepedi	3	0.0718697	0.0050477
Sepedi	4	0.0805596	NaN
Sesotho	1	0.0366044	0.0062598
Sesotho	2	0.0457010	0.0213640
Sesotho	3	0.0949535	0.0113032
Sesotho	4	0.1025924	NaN
Setswana	1	0.0351724	0.0068681
Setswana	2	0.0359537	0.0037783
Setswana	3	0.0711086	0.0106264
Setswana	4	0.0814226	NaN
SiSwati	1	0.0441640	0.0000000
SiSwati	2	0.0612813	0.0091185
SiSwati	3	0.0458221	0.0134771
SiSwati	4	0.0762332	NaN
Tshivenda	1	0.0334928	0.0000000
Tshivenda	2	0.0000000	0.5441176
Tshivenda	3	0.0225080	0.0000000
Tshivenda	4	0.0808625	NaN
IsiTsonga	1	0.0235294	0.0000000
IsiTsonga	2	0.0118483	0.0932836
IsiTsonga	3	0.0397196	0.0023529
IsiTsonga	4	0.0776892	NaN
Afrikaans	1	0.1345441	0.0465116
Afrikaans	2	0.0904233	0.0424528
Afrikaans	3	0.1107348	0.0399729
Afrikaans	4	0.1096479	NaN
English	1	0.2969374	0.1016043
English	2	0.3234127	0.1070707
English	3	0.3156934	0.1023766
English	4	0.3233333	NaN

8 Covariate Descriptive Statistics

Figure 7: Household Income

```
ggplot(adulthh, aes(x=hhincome, fill=a_lng )) +
    stat_bin(bins=50)
```

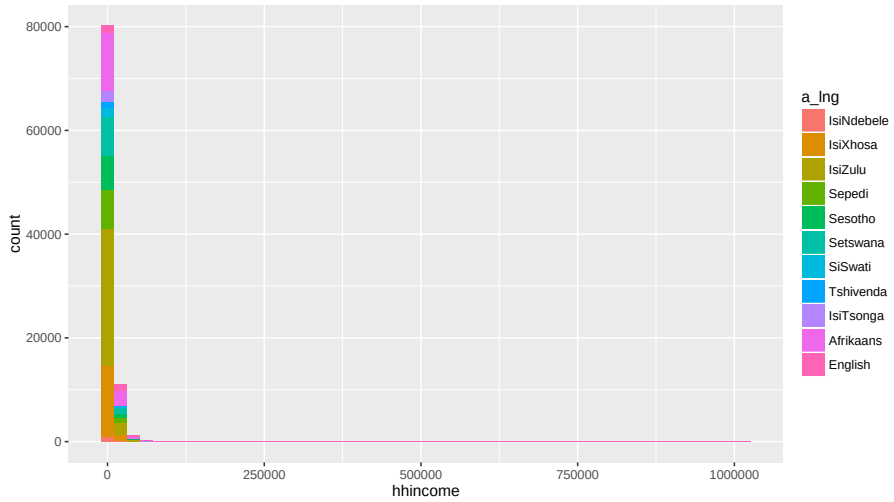
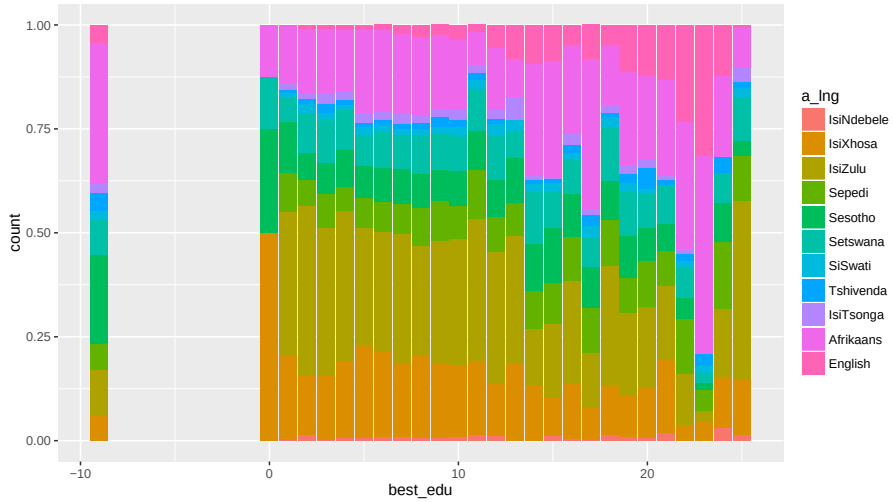


Figure 8: Years of Education

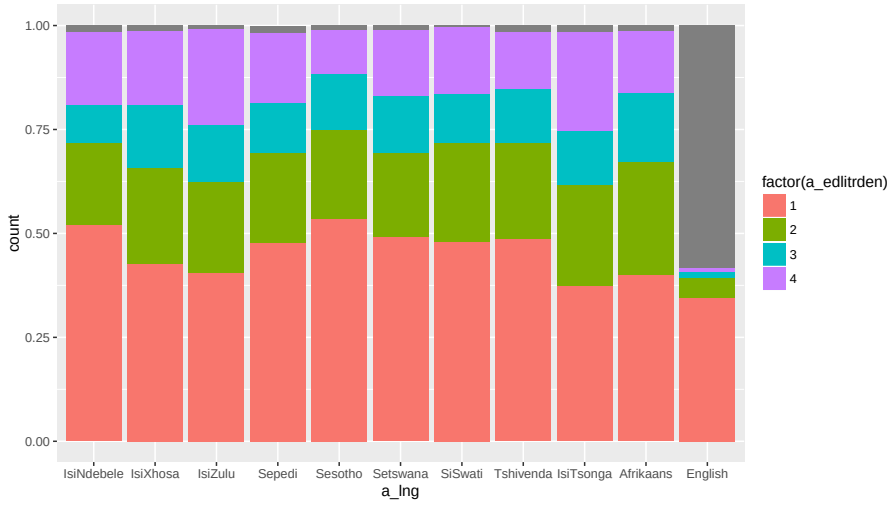
```
ggplot(adulthh, aes(x=best_edu, fill=a_lng )) +
    geom_bar(position = 'fill')
```



The Figure 9 describes the skill of individuals in reading the English language, where 1 the best and 4 is the worst, grey values are `NA`.

Figure 9: English Language Reading Skills

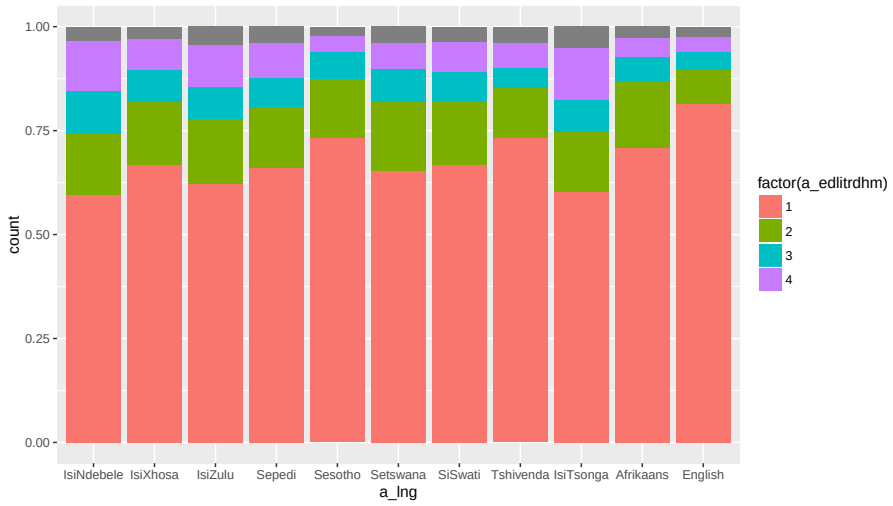
```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrden)) ) +
        geom_bar(position = 'fill')
```



The Figure 10 does the same but with regards to the native language.

Figure 10: Native Language Reading Skills

```
ggplot(adulthh, aes(x = a_lng, fill = factor(a_edlitrdhm)) ) +
        geom_bar(position = 'fill')
```



9 Original Estimates

Table 6: Computer Ownership

```
lm(a_owncom ~ interface_intro*setswana_logical +
               factor(a_edlitrden)                          +
               factor(a_edlitwrten)                         +
               factor(a_edlitrdhm)                          +
               factor(a_edlitwrthm)                         +
               a_woman                                      +
               hhincome                                     +
               best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0095949	0.0025592	3.7492016	0.0001775
interface_introTRUE	-0.0061590	0.0016430	-3.7486711	0.0001779
setswana_logicalTRUE	-0.0155134	0.0043191	-3.5918390	0.0003285
factor(a_edlitrden)2	-0.0323438	0.0053645	-6.0292402	0.0000000
factor(a_edlitrden)3	-0.0341564	0.0070115	-4.8714783	0.0000011
factor(a_edlitrden)4	-0.0392838	0.0089335	-4.3973416	0.0000110
factor(a_edlitwrten)2	-0.0185701	0.0053510	-3.4703684	0.0005200
factor(a_edlitwrten)3	-0.0202804	0.0069315	-2.9258377	0.0034362
factor(a_edlitwrten)4	-0.0204843	0.0087822	-2.3324736	0.0196780
factor(a_edlitrdhm)2	0.0035074	0.0048950	0.7165378	0.4736612
factor(a_edlitrdhm)3	0.0002144	0.0067936	0.0315626	0.9748209
factor(a_edlitrdhm)4	-0.0293401	0.0089210	-3.2888737	0.0010063
factor(a_edlitwrthm)2	-0.0011334	0.0049194	-0.2303878	0.8177909
factor(a_edlitwrthm)3	-0.0054732	0.0067905	-0.8060182	0.4202345
factor(a_edlitwrthm)4	-0.0398754	0.0089154	-4.4726177	0.0000077
a_womanTRUE	-0.0269718	0.0014992	-17.9909045	0.0000000
hhincome	0.0000054	0.0000001	71.7878843	0.0000000
best_edu	0.0061744	0.0001527	40.4384106	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0235410	0.0052777	4.4604463	0.0000082

Table 7: Living in Household that Spent on Internet (last 30 days)

```
lm(h_nfnet ~  interface_intro*setswana_logical +
              factor(a_edlitrden)                        +
              factor(a_edlitwrten)                       +
              factor(a_edlitrdhm)                        +
              factor(a_edlitwrthm)                       +
              a_woman                                    +
              hhincome                                   +
              best_edu)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0016483	0.0015285	-1.0783754	0.2808703
interface_introTRUE	-0.0118218	0.0009715	-12.1690167	0.0000000
setswana_logicalTRUE	-0.0134334	0.0023200	-5.7902814	0.0000000
factor(a_edlitrden)2	0.0027653	0.0033372	0.8286286	0.4073177
factor(a_edlitrden)3	-0.0000533	0.0043178	-0.0123394	0.9901549
factor(a_edlitrden)4	-0.0027134	0.0055723	-0.4869402	0.6263024
factor(a_edlitwrten)2	-0.0091843	0.0033309	-2.7572930	0.0058298
factor(a_edlitwrten)3	-0.0087232	0.0042760	-2.0400443	0.0413500
factor(a_edlitwrten)4	-0.0058617	0.0054752	-1.0706002	0.2843533
factor(a_edlitrdhm)2	-0.0007155	0.0030629	-0.2336010	0.8152955
factor(a_edlitrdhm)3	-0.0008317	0.0042061	-0.1977297	0.8432572
factor(a_edlitrdhm)4	-0.0056993	0.0055568	-1.0256456	0.3050625
factor(a_edlitwrthm)2	-0.0011172	0.0030899	-0.3615735	0.7176720
factor(a_edlitwrthm)3	-0.0012645	0.0042158	-0.2999539	0.7642133
factor(a_edlitwrthm)4	-0.0093993	0.0055594	-1.6907221	0.0908947
a_womanTRUE	-0.0006753	0.0009301	-0.7259811	0.4678530
hhincome	0.0000027	0.0000000	56.7219590	0.0000000
best_edu	0.0012953	0.0000936	13.8375596	0.0000000
interface_introTRUE:setswana_logicalTRUE	0.0120078	0.0030875	3.8892010	0.0001007

10 Software

The estimation is primarily performed using R (R Core Team, 2016), specifically using `lm()` and `glm()` functions included in the `stats` package (Venables and Ripley, 2013). Additionally, I make the of the `plm()` and `pglm()` functions which are available in packages by the same names (Croissant, 2013; Croissant, Millo et al., 2008). Standard error corrections are computed using the `lmtest` and `sandwich` packages Zeileis (2004, 2006); Zeileis and Hothorn (2002).

In order to make the result as easily reproducible as possible, this research and writing in the article has been done exclusively using open-source software such as R (R Core Team, 2016). This document is written and LyX (LyX Team, 2016) in the LaTeX(Lamport, 1985) language and compiled using the LuaTeX implementation(Hoekwater et al., 2016). The integration of R code and output in the document is performed using a process call literate programming Knuth (1984) using the knitr implementation Xie (2015) of the Sweave framework (Leisch, 2002).

All changes are logged using the version control system Git (Git Team, 2016) and publicly available on GitHub at https://github.com/bquast/Making-Next-Billion-Demand-Access/[1]

[1]The repository can by cloned to a local computer by entering in following command in a terminal (with Git installed):
`git clone https://github.com/bquast/Making-Next-Billion-Demand-Access.git`