London School of Economics and Political Science

Essays on the Role of the Internet in Development and Political Change

Luke Ian Miner

A thesis submitted to the Department of Economics of the London School of Economics for the degree of Doctor of Philosophy, London

July 2012

Declaration

I certify that the thesis I have presented for examination for the PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 39,299 words.

Statement of conjoint work

I confirm that Chapter 3 was jointly co-authored with Dr. Larcinese and I contributed more than 60% of this work.

Abstract

This thesis contains three independent chapters aimed at increasing our understanding of the effects of Internet diffusion on politics and development. The first chapter proposes a novel methodology for measuring Internet penetration. Using IP geolocation data, a new measure of Internet access is created, which counts the number of IP addresses per person in a region. This is the first measure of Internet penetration that is comparable not only across countries but across sub-regions of countries such as states or even electoral districts. The second chapter applies this measure to test whether Internet diffusion can weaken incumbent power in a semi-authoritarian regime. Using Malaysia as a test case, I find that the Internet is responsible for a six point swing away from the incumbent party in the 2008 elections. In the third chapter, co-authored with Valentino Larcinese, we look at the effects of the Internet on U.S. presidential elections. In accordance with anecdotal evidence, we find that increased Internet penetration leads to an increase in small donations to the Democratic Party and a swing towards the Democratic presidential candidate.

Acknowledgments

This work would not have been possible were it not for the assistance and advice of a wide array of people. First I would like to thank my PhD supervisors at the LSE, Robin Burgess and Gerard Padro-i-Miquel, for the invaluable feedback that they provided over the course of my thesis. Their advice was key in helping me manage the roadblocks that occurred along the way and their criticisms vastly improved both my work and my thinking. I am also extremely grateful to Greg Fischer. Without his advice and encouragement at the beginning of my thesis, this project would not have been possible. I would like to thank Henry Overman and Steve Gibbons. Their input into the geo-spatial aspects of my thesis was invaluable.

During the 2009-2010 year I was a guest of the University of Tokyo and I would like to thank the faculty, in particular Yasuyuki Sawada, for their helpful advice.

My colleagues have provided inestimable help over the course this project: as sounding boards to generate ideas, as participants of seminars where I presented, as teachers of particularly difficult points of theory. I'd like to thank Michael Best, Anne Brockmeyer, Konrad Burchardi, Guilhem Cassan, Francisco Costa, John De Quidt, Jason Garred, Tara Mitchell, Ashwini Natraj, Miriam Sinn, Dan Stein, and Oliver Vanden Eynde. I'm especially grateful to Dimitri Szerman for being such a wonderful teacher of Stata, and to John Barrdear for his great advice and support.

This work has benefited from the comments of many people. The second chapter has benefited from the comments of members of the EC501 Development and Growth Seminar Series: Tim Besley, Maitreesh Ghatak, Oriana Bandiera, and Torsten Persson. The third chapter received invaluable criticism from the PSPE Doctoral Workshop: Brenda Van-Coppenolle, Rafael Hortala-Vallve, and James Snyder.

Finally I would like to thank my wife for her infinite patience during this process and for her patient reading of each chapter for spelling and grammar mistakes.

Contents

Abstract ii

Aknowledgments iv

Table of Contents vi

Preface 4

1 Measuring Internet Diffusion from IP Addresses 7

1.1 Introduction . . . 7
1.2 Data . . . 10
1.2.1 Data Sources . . . 10
1.2.2 Evaluating Data Sources . . . 12
1.3 Method . . . 15
1.3.1 Accounting for Error from Dynamic IP Allocation . . . 16
1.3.2 Accounting for Geographical Measurement Error . . . 16
1.4 Results . . . 17
1.4.1 Measurement error . . . 19
1.5 Robustness . . . 20
1.6 Conclusion . . . 21
1.7 Tables and Figures . . . 22
1.8 Data appendix . . . 33

2 The Unintended Consequences of Internet Diffusion 35

2.1 Introduction . . . 35
2.2 Theoretical Framework . . . 39
2.2.1 Basic Model . . . 39
2.2.2 Results when all districts are identical . . . 41

2.2.3 Extension: Internet penetration and population size vary across districts . . . 43
2.3 Background . . . 45
2.3.1 Political regime . . . 45
2.3.2 The Internet and politics . . . 47
2.3.3 Internet placement . . . 49
2.4 Data . . . 52
2.4.1 Political Data . . . 52
2.4.2 Demographics and geography . . . 53
2.4.3 Internet . . . 54
2.5 Empirical Analysis . . . 55
2.5.1 Basic Correlations: OLS Estimates . . . 55
2.5.2 Identification Strategy . . . 58
2.5.3 First Stage . . . 60
2.5.4 Instrumental Variable Results . . . 61
2.5.5 Validity of the Exclusion Restriction . . . 62
2.5.6 Additional robustness checks . . . 64
2.6 Additional results . . . 65
2.6.1 Turnover . . . 65
2.6.2 Turnout . . . 66
2.6.3 Predicted results in absence of Internet . . . 67
2.7 Conclusion . . . 68
2.8 Tables . . . 70
2.9 Figures . . . 82
2.10 Appendix A: Constructing a measure of Internet penetration . . . 91
2.11 Appendix B: Derivation of bias . . . 96
2.12 Appendix C: Proofs of propositions . . . 98
2.12.1 Proof of Proposition 1 . . . 98
2.12.2 Proof of Proposition 2 . . . 99
2.12.3 Proof of Proposition 3 . . . 99
2.13 Appendix D: Data appendix . . . 100
2.14 Appendix E: Appendix on Election Irregularities . . . 103

3 The Political Impact of the Internet on US Presidential Elections 105
3.1 Introduction . . . 105

3.2 Background . . . 107
3.2.1 Personal Use of Internet . . . 108
3.2.2 Campaign Use of Internet . . . 109
3.2.3 Expected Outcomes . . . 110
3.3 Data . . . 110
3.3.1 Census Data . . . 110
3.3.2 Political Data . . . 111
3.3.3 Right-of-Way Data . . . 111
3.3.4 Internet Data . . . 113
3.4 Method and Results . . . 114
3.4.1 Basic Results . . . 115
3.4.2 Endogeneity Concern . . . 116
3.4.3 First Stage and Reduced Form . . . 117
3.4.4 IV Results . . . 118
3.4.5 Robustness Checks . . . 119
3.5 Conclusion . . . 120
3.6 Tables and Figures . . . 121
3.7 Data Appendix . . . 132

Bibliography **135**

List of Tables

Correlations between providers and high-speed lines per capita 132

OLS estimates of democrat vote share on ISP growth 132

OLS estimates of vote share, turnout, and donations on ISP growth 132

OLS estimates of all outcomes on ISP growth with lagged dependent variable 1996-2008 132

First stage relationship between providers and index of ROW laws 1996-2008 132

Reduced form relationship between democrat share and index of ROW laws 1996-2008 132

IV estimates of democrat share on Internet access 1996-2008 132

IV estimates of other outcomes on Internet access 1996-2008 132

Reduced form estimates of democrat share on ROW index for previous years 132

Placebo OLS estimates of change in democrat vote share on ISP growth . . 132

Data Appendix: The Political Impact of the Internet on US Presidential Elections 133

List of Figures

1.1 Geographic Placement of GeoIP Data 22
1.2 Peninsular Malaysia's Road and Railway Network 23
1.3 Population Distribution 24
1.4 Legislative District vs. Census District Boundaries 25
1.5 Fraction of Sample Not Assigned to Geographic Location 27
1.6 Interpolation of 25 Square Kilometer Buffer Zone 28
1.7 Interpolation: Inverse Distance Weighting Interpolation 29
1.8 Linear IPperPerson Against Percent with Household Internet 31
1.9 Log IPperPerson Against Percent with Household Internet 31

2.1 Political Boundaries 82
2.2 Peninsular Malaysia Land Use and Elevation 83
2.3 Peninsular Malaysia's Road and Railway Network 84
2.4 Population Distribution 85
2.5 Legislative District vs. Census District Boundaries 86
2.6 Inverse Distance Weighting Interpolation 87
2.7 Backbone Location 88
2.8 Skewness of Households with Internet 2004 89
2.9 Relationship Between Log Households with Internet 2004 and Fraction of Total Eligible Voters 89
2.10 Relationship Between Change in BN Vote Share and IP per Voter Growth 90

3.1 Counties in Sample 121
3.2 Correlation Between Providers and High Speed Lines per Capita . . . 122
3.3 Predicted Broadband Adoption by Zip Provider Count. *Note.* Solid line represents predicted values, and dotted lines represent upper and lower bounds of 95% confidence interval. **Source: Kolko (2010).** . 123

Preface

Since its inception in the 1980s the Internet has grown from an obscure project of the U.S. Department of Defense to a medium used by billions to communicate, find information, trade opinions, view content, and connect. Unlike previous technological innovations of the 20th century like radio and television, the Internet is less beholden to national boundaries and much more difficult to regulate. As a result, in much of the world, content is less controlled on the Internet than it is on other mediums. For example, it is possible to access political content that, in some countries, could never be printed in newspapers. Content of all kinds can be downloaded, bringing libraries worth of information to users previously starved of content, some of it free but much of it proprietary. The Internet also brings new kinds of engagement: the ability to meet new people who share similar interests or simply comment on the same posts; instantaneous updates about friends who live on the other side of the globe; advertisements tailored to individual browsing patterns.

With the phenomenal growth of the Internet has come equally phenomenal claims about its influence on our lives. The Internet has been credited with bringing great economic benefit, with many regions attempting to duplicate the successes of Silicon Valley and Bangalore. The Internet is also seen to foment political change, whether it be Obama's victory in the 2008 U.S. presidential elections, protests in Russia, or the "Facebook" revolution in Egypt. But to date most evidence is anecdotal. There is very little agreement as to whether the Internet actually has as large of an effect as some have claimed, and much less of an idea of what kind of effect that it might have. This thesis aims to broaden our knowledge of this understudied area, providing some of the first evidence on the effects of Internet based media on the political economy both of industrialized and developing countries.

Chapter 1, addresses one of the main obstacles to empirical research on the Internet, a lack of reliable data. Today there is no publicly available, official data on Internet usage at the sub-national level. This chapter aims to fill this gap, proposing a new

method for measuring Internet usage based on IP geolocation data, a dataset which matches IP addresses with physical locations. Using Malaysia as a test case, I create a measure for Internet penetration at the state electoral district level. I show that this is a strong measure, with a high degree of correlation with official survey data on fraction of households with Internet subscriptions.

Chapter 2, tests whether the introduction of the internet can undermine incumbent power in a semi-authoritarian regime. I use Malaysia as a test case, where the incumbent coalition lost its 40-year monopoly on power in 2008. I draw upon the methodology from Chapter 1 to construct a measure of Internet penetration for the 2004 to 2008 period in Malaysia. Using an instrumental variable approach to account for endogenous internet placement, I find that areas with higher internet penetration experience higher voter turnout and higher candidate turnover, with the internet accounting for one-third of the 11% swing against the incumbent party in 2008. In fact, the results suggest that, in the absence of the internet, the opposition would not have achieved its historic upset in the 2008 elections.

Finally, in chapter 3, co-authored with Dr. Valentino Larcinese, we analyze the Internet's effect on the 2008 U.S. presidential elections. According to anecdotal evidence, the internet is said to have played a key role: the Obama campaign's online fundraising arm brought in a record $500 billion in small individual donations; and the campaign's heavy use of social media purportedly contributed to the highest rate of youth turnout since voting was extended to 18-year-olds. We test these assertions exploiting geographic discontinuities along state borders with different right-of-way laws, which determine the cost of building new infrastructure. We find that areas with higher internet growth are more likely to swing to the Democratic presidential nominee and are more likely to provide small donations to the Democratic Party.

1 Measuring Internet Diffusion from IP Addresses

1.1 Introduction

The Internet has been in existence for well over twenty years, and yet we know surprisingly little about it, lacking reliable, comparable sub-national data on Internet topology, infrastructure, supply, and usage. This a problem evident even in the United States, the birthplace of the Internet, which didn't draw up a map of broadband availability until 2009.

Reliable data on the extent and usage of the Internet is of critical importance. First, it would facilitate decision-making by policy-makers. All around the world, governments have engaged in a myriad of costly programs to expand Internet capacity under the belief that this will promote growth.[1] Consistent data would make it easier to assess the relative efficacy of these programs. More importantly, it could shed light on what economic benefits the Internet brings and to whom. Second, a reliable measure of the Internet would help academics across multiple disciplines gain a better idea of the social and politial impact of this near-ubiquitous technology. The Internet has been credited with a wide variety of phenomena: revolutions in the Middle-East,[2] increased political polarization,[3] the life of political campaigns[4] and

[1] Malaysia, for example, created the Multimedia Super Corridor, which includes a new city, Cyberjaya, geared towards ICT companies and their employees.

[2] See en.wikipedia.org/wiki/Facebook_Revolution.

[3] See comments by Eric Schmidt of Google at bits.blogs.nytimes.com/2010/09/17/googles-chief-on-the-web-and-political-polarization/

[4] According to Arianna Huffington Obama would not have won the 2008 elections without the help of the Internet. See techcrunch.com/2008/11/07/the-internet-as-a-force-in-politics-obama-would-not-have-won-without-the-internet/.

also their death,[5] a powerful tool for monitoring corruption,[6] even increased sexual violence (Bhuller, Havnes, Leuven, and Mogstad, 2011). However, research thus far has been hampered by a lack of a consistent measure of the Internet or how it is used.

In this paper, I propose a novel method of measuring Internet penetration that can be applied to almost any country. I use data from the MaxMind IP geolocation service, which maps IP addresses to physical locations. I propose four measures to take account of the fact that IP addresses can change geographical location over time. Because the data is only available at the city level, I assess an additional three methods for smoothing the data such that it produce estimates for all areas in a country.

I test these measures using a government survey on the fraction of households with Internet connections. I find that the strongest measure (1) takes the sum of the number of times IP addresses are assigned to each location over a year long period and (2) uses Inverse Distance Weighting (IDW) Interpolation to smooth point data into a measure for each location in the country. The optimal measure is strong, with a .63 correlation with survey data.

A network of networks, the Internet is difficult to measure due to its diffuse nature. Supra-national with no central governing authority, the Internet is a patchwork of interlocking networks varying in size, type of ownership, and speed, spanning whole continents, and tied together by a web of underwater cabling hundreds of thousands of kilometers in length. If assembling information from so many disparate sources weren't enough of a challenge, the commercial sensitivity of information on usage and infrastructure makes Internet Service Providers (ISPs) loathe to release this data.

Due to the Internet's complexity, there are many different aspects that can be measured. First there is the physical infrastructure: the mass of copper, fiber, routers, switches, towers, data centers that keep the Internet running. Next there is the mass of information flowing through the network, the nodes through which it travels, the way the data is routed and the paths taken.[7] Finally there is the question of avail-

[5] Senator George Allen's loss to Democrat challenger James Webb has been attributed to the video of Allen uttering a racial slur, which was uploaded to YouTube. See www.surveyusa.com/client/PollReport.aspx?g=a99a9b7d-89aa-4e5f-9a0e-35d657ae1db3.

[6] See www.ipaidabribe.com

[7] In the field of Web Science, organizations like the Cooperative Association for Internet Data Analysis (CAIDA) actively collect data on the topology of the Internet.

ability to consumers and business and to what extent this availability translates into usage.

Some countries and private companies[8] have started collecting survey and census data on Internet usage. However, it is difficult to quantify Internet usage even in this case. Many early surveys only count home Internet subscriptions, ignoring the many people who access the Internet from diverse locations such as the workplace, Internet cafes, and mobile phones.

Due to a lack of official survey data, researchers have turned to a number of alternate methods. Kolko (2010) uses the number of high speed ISPs registered in a zip code as a proxy for broadband availability, finding a significant, non-linear relationship between the number of high speed ISPs registered in a U.S. zip code and broadband adoption. Hu and Prieger (2008) use DSL availability at the central office/wire-center level as reported by Ameritech at the time of a merger with SBC in 1999. However, in both cases this data is only available for the United States. This is the first paper to employ a measure that can be used in most countries in the world, and the first to employ IP geolocation data as a proxy for Internet usage.

Much work has been performed by computer scientists to map out the network topology of the Internet. Danesh, Trajkovic, Rubin, and Smith (1999) provide a survey of some of the earlier attempts relying on traceroute utility, which returns all of the nodes that a packet passes through on the way to a destination. Faloutsos, Faloutsos, and Faloutsos (1999) consider how networks and ISPs connect at the more aggregate Autonomous System (AS) level. Heidemann, Pradkin, Govindan, Papadopoulos, Bartlett, and Bannister (2008) perform an Internet census, probing every visible host on the Internet. Although these studies have greatly improved our understanding of the network topology of the Internet—namely, the flow of data between computers, routers, and networks—they do not provide much information on the physical location of Internet users.

Currently, the Cooperative Association for Internet Data Analysis (CAIDA) has combined these and other techniques to get the most complete picture of the Internet. Fomenkov and Claffy (2011) outlines the various tools at their disposal. Under the auspices of CAIDA, Huffaker, Fomenkov, and Claffy (2011) analyzes the accuracy of IP geolocation databases in operation today. They find that the database used in this paper, MaxMind GeoIP, is one of the better performers, especially for

[8]Such as Forrester.

Asian addresses.

This paper proceeds as follows. In section 3.3, I introduce the concept of the IP address before explicating the data used in this analysis. I then move on to section 1.3, where I present my methodology for measuring Internet usage with IP geolocation data. Section 1.4 shows the result, and section 1.5 checks robustness. In section 3.5, I conclude.

1.2 Data

1.2.1 Data Sources

In order to properly explain the data sources, it is necessary to give a brief definition of IP addresses. An Internet Protocol address (IP address) is a number that is assigned to every device that connects to the Internet, ranging from servers to personal computers to mobile phones. There are two types of IP addresses: IPv4, the only type of IP address up until 1995, and IPv6, which was developed to address the problem of the depletion of available IPv4 addresses. This paper focuses solely on IPv4 addresses, as IPv6 addresses were few and unused during the period of analysis. In addition to there being two types of IP addresses, there are also two ways of assigning IP addresses: statically and dynamically. Dynamic IP addresses are reassigned to multiple devices over time, whereas static IP addresses remain attached to the same device. The majority of IP addresses are dynamically assigned. As we will see below, this leads to significant geographic movement in the dataset, introducing measurement error.

GeoIP database The MaxMind GeoIP City database forms the core of my Internet measure. GeoIP City is a service that matches IP addresses to geographic locations, allowing web services to tailor advertisements based on visitor location and to detect fraud. The GeoIP City database comprises monthly data from 2005 to the present and covers virtually all IP addresses in the world.[9] For each IP address, the GeoIP City database provides the name and location of the nearest large city on a monthly basis.

[9]MaxMind does not cover IPv6 addresses. However, IPv6 adoption was infinitesimal in Malaysia at the time.

MaxMind's technique for matching IP addresses to physical locations is proprietary, thus it is not possible to give a full account of their methods. But their technique likely relies on a mixture of the following three approaches.[10] The first approach involves delay-based methods, where information on the delay[11] from a collection of IP addresses with known geographic location is used to triangulate the location of IP addresses without known locations. The second approach relies on database-driven methods: a map of all static IP addresses is amassed using information from public and private datasets. The final approach involves topology-driven methods: areas that are close to each other in terms of the Internet's topology will also be close to each other physically.

APNIC database The second source of data comes from the Asia Pacific Information Centre (APNIC). APNIC is one of five regional Internet registries in charge of distributing and managing IP addresses. They release historical data going back to 2001, updated monthly, which gives a complete list of which IP addresses were allocated to what companies, and on what date.[12] As we shall see below, this information is valuable because large chunks of IP addresses are often assigned to companies, who use them for a number of purposes unrelated to connecting users to the Internet.

Census measure I use official Internet measures from the Population and Housing Census 2000 and the Household Basic Amenities and Income Survey (HBAIS) for 2004. Both datasets provide the fraction of households with Internet subscriptions at the *mukim* (census district) level. As explained in Section 1.2.1, I use ArcGIS to aggregate the HBAIC data to the legislative district level, which introduces some measurement error.

Demographics and geography I have complete geospatial data for Malaysia. Figure 1.1 illustrates clutter data (which classifies all land as either urban, semi-urban, plantation, jungle, inland water, or open) and elevation data (which allows for the calculation of land-gradients). Figure 1.2 shows the locations of all major roads,

[10] For a more thorough description of these techniques, see Huffaker, Fomenkov, and Claffy (2011)
[11] Delay can be thought of as the amount of time it takes a data packet to travel to the host computer performing the analysis.
[12] See http://ftp.apnic.net/apnic/stats/apnic/.

highways, and railways in Malaysia. Finally, figure 1.3 represents data from the LandScan service, which estimates population distribution at one square kilometer resolution through a combination of census data and satellite imagery.[13]

State legislature districts, on average, are 21% urban and 50% farmland (rural), with the remaining 29% classified as jungle. Although jungle covers large swaths of the country, the fairly extensive road network spans more than 80,000 kilometers of roads as of 2007.

I have constructed a dataset of controls using the Population and Housing Census of Malaysia for 2000; Malaysia's Household Basic Amenities and Income Survey for 2004; and geographically disaggregated measures of GDP per capita 2005, generated by the consultancy Booz & Company. Unless otherwise stated, this data is available at the level of Malaysia's 927 census districts, called *mukim*.

Figure 1.4 shows mukim boundaries alongside state legislative district boundaries. As can be seen, mukim level data does not match up perfectly with state legislative districts. To address this discrepancy, I use the LandScan population data to assign a weight to each one kilometer cell within each mukim. State electoral district values are generated from the weighted sum of these one square kilometer cells.

Malaysia is a multi-ethnic society. Ethnic Chinese, the wealthiest group in Malaysia, comprise 26% of the population. Indians currently comprise roughly 8% of the population. The remainder (65% percent) is largely Malay, except for several ethnic groups on Borneo and a few small tribes.

1.2.2 Evaluating Data Sources

Dynamic IP Addresses and Measurement Error The accuracy of data and level of geographic disaggregation varies from country to country. A major reason for this is that most IP addresses are dynamically assigned. When an IP address is dynamically assigned a device may keep the same IP address for days or even weeks, but the number will eventually change. Often the device will simply be reassigned to a random IP address from the same general area, but in some cases the new IP address could have previously been assigned to a device in a completely different region.

[13]See <http://www.ornl.gov/sci/landscan/> for details on the construction of this dataset.

Table 1.1, shows the extent to which this is a factor in the dataset. In this table, I take the average of the GeoIP data for the twelve months of 2005. The Internet survey that I will later use to evaluate my measure was conducted in 2004. In order to deal with this issue, I drop all IP addresses from the sample that were assigned after the date of the survey.[14] I accomplish this task by augmenting the sample with the data from APNIC, which provides the date that the IP address was assigned and the company that managed the IP address.[15] The first row of Panel A shows that 27% of IP addresses were assigned after the date of the survey.

Another issue with the dataset is that a substantial number of IP addresses are never assigned to a geographic region. One reason for this is that a substantial fraction of IP addresses are not used by consumers to access the Internet. For example, Petronas, Malaysia's national oil and gas company, controls half a million IP addresses, but only employs 40,000 people. Webhosting platforms, like Blogger, also control a large amount of IP addresses, which are used for webpages rather than to allow a device to connect to the Internet. We can see this phenomenon in Panel A: the second row shows that out of the IP addresses assigned up to 2004, almost a quarter of them have not been matched to any location over the entire period; the third row shows that if we limit the sample to ISPs that actually serve consumers, only 10% of the addresses are never matched to a geographic location.

Since IP addresses are dynamic, it is possible for them to be unmatched one month and then matched the next. The first column of Panel B shows the fraction of IPs that are unmatched to a geographic location and for how many months. 39% of IP addresses are unmatched to a geographic location for at least one month, but this drops off heavily to 6% unassigned for two months, and less than .2% for three months. Figure 1.5 shows the fraction of IP addresses that are unmatched by month. As can be seen, the fraction of IP addresses unmatched is between 30% and 40% during the first half of the year, but drops to below 10% during the second half. Thus another aspect of the dataset is that the geographic precision is increasing with time.

Discussions with engineers at Malaysia's ISPs suggest that there is a large bias in the data towards large cities, in particular Kuala Lumpur. We can see evidence for this assertion in the fourth row of Panel A, which shows that 23% of IP addresses are

[14] Results are similar if these IPs are included.

[15] Not all regional registries provide this information. However, MaxMind provides some of the same information in its other products.

matched with Kuala Lumpur and remain fixed to Kuala Lumpur during the entire twelve month period. Moreover, of the IP addresses that shift location, the second column of panel B indicates that an additional 25% of IP addresses were assigned to Kuala Lumpur once, and 7% were assigned twice. Since Kuala Lumpur only makes up 6% of the population, it is unlikely that this discrepancy can be entirely explained by increased Internet uptake in Kuala Lumpur.

Panel C shows how much IP addresses change location in practice. In the first row, we see that the average IP address changes locations around 2.4 times during the sample. The average amount of time an IP spends attached to any location is around 6 months. Breaking this number down somewhat, in the third row I find that the average time spent at the location that is held the longest by an IP address is eight months. This implies that on average IP addresses spend the majority of their time allocated to one location.

As IP addresses are always being added to a country's IP space, there's a possibility of bias due to some IP addresses only appearing later in the year. The row of Panel C implies that this is less of a worry as it implies that the average IP address appears in 11.6 months. The problem of constant assignment over the year is more likely reflected in the numbers of IP addresses either missing or assigned to Kuala Lumpur. These IP addresses are likely to recently have been acquired by an ISP, but not yet allocated to a more permanent location.

One last possible source of error comes from the diminishing number of IPv4 numbers. Because there is a scarcity of IPv4 numbers and in 2004 IPv6 was hardly used, IPv4 numbers were rationed. One way to address this problem is to place an entire network behind a single IP address. However, IPv4 depletion was a long way off in 2004 and there is little evidence that this type of behavior was occurring on a large scale.

Geographic Matching and Measurement Error Figure 1.1 shows how the 647 cities matched to IP addresses in Malaysia are distributed spatially. As can be seen, the cities are concentrated on the coasts, where the majority of the population resides. Importantly, although the data provides the coordinates of the city center, it does not give any information on the boundaries of the city. Typically a city will encompass a greater region around it, including suburbs and smaller towns. This spatial indeterminacy is a potential source of error dependent upon what level of

geographic disaggregation a measure is sought.

1.3 Method

In the following section I construct a measure of Internet penetration, which I call $IPperVoter$ to reflect the fact that the measure is at the level of the state legislature district level. Before proceeding any further, it is helpful to quickly summarize the challenges facing this measure due to data limitations.

The two primary sources of measurement error are as follows:

1. *Change in IP block location over time:* For reasons discussed in Section 1.2.1, the locations to which IP addresses are assigned change over time. The monthly data is very noisy with smaller cities disappearing and reappearing from month to month. Moreover, larger cities tend to be overcounted, especially Kuala Lumpur.

2. *Geographical Measurement error:* The dataset only provides the coordinates for the city center; it does not specify the city's limits. Thus, an IP address corresponding to a computer in a small town outside a city (and in a different electoral district) may be incorrectly attributed to the city, introducing further bias toward large cities. Furthermore, since only point data is available, the boundaries between cities are unclear. This can most easily be seen in figure 1.1, which presents GeoIP city center data alongside legislative district boundaries. As can be seen, the point locations often appear on the border between two districts, complicating the task of divvying up IP addresses between adjacent districts.

In Section 1.3.1, I will show how I account for the error introduced by the first of these problems. Section 1.3.2 will show how the second of these problems is addressed. Finally, in Section 1.4 the various methods proposed will be tested against official survey data and the optimal measure will be chosen.

1.3.1 Accounting for Error from Dynamic IP Allocation

In response to the problem of IP locations changing over time, I test four methods for assigning IP addresses to cities based on data from a year-long period.

1. *$IPfix$:* I limit the sample to IP addresses that never change location over the twelve month period and divide by twelve. This is the most conservative measure, yielding the advantage that the IP addresses are almost certainly assigned to the correct location. The disadvantage is that the majority of the sample is lost with most remaining IPs assigned to Kuala Lumpur.
2. *$IPsum$:* I sum the number of IP addresses assigned to each city over a twelve-month period. This is the second most conservative measure, making equal use of all of the information in the dataset. However, it likely leads to over-counting Kuala Lumpur. For example, there are many cases in which ten or eleven out of twelve observations occur in a single city, with the remainder going to Kuala Lumpur.
3. *$IPmax$:* I calculate the number of IP addresses assigned to each city for each month. $IPmax$ is the value from the month with the most IP addresses. This measure is meant to account for under-counting of smaller cities. The assumption is that, once an area obtains Internet access, it is unlikely to subsequently have access physically dismantled. If a location does not appear in subsequent months, this is due to measurement error rather than a subsequent loss of Internet connectivity.
4. *$IPavg$:* Again I calculate the number of IP addresses assigned to each city for each month. I then take the average across months when the city appears in the data. This approach is similar to $IPmax$, but aims to correct for possible over-counting of small cities by $IPMmax$.

1.3.2 Accounting for Geographical Measurement Error

I propose three methods for dealing with the second challenge, geographical measurement error.

1. *No Smoothing:* As a baseline, no smoothing of the data is performed. Regions containing city points get all IP addresses assigned to them. This is especially problematic for large cities that are broken up into a number of electoral

districts. The district that happens to contain the geographic center of the city will get all the IP addresses and the rest of the districts will get no IP addresses.

2. *Smooth into a 25 Square Kilometer Area:* The IP addresses are allocated to a 25 square-kilometer, circular buffer area around the center of the city as in figure 1.6. This addresses some of the issue of IP addresses for an entire city being assigned to a district, but still doesn't deal with the problem of suburbs of the city.

3. *IDW Interpolation:* I smooth the city point-data into a surface covering the entire area of peninsular Malaysia, using inverse distance weighting (IDW) interpolation in ArcGIS. IDW interpolation assigns an IP measure to every point in Malaysia: the value at each interpolated point is a weighted sum of the values in the N known points, where closer points get higher weighting. Figure 1.7 shows an example of IDW interpolation: darker areas have higher numbers of IP addresses.[16]

1.4 Results

The approach outlined above produces twelve alternate measures of Internet penetration, four for each of the three geographical interpolation methods. Table 1.2 tests the accuracy of these measures showing how well they correlate with the number of Internet subscriptions per household, which was collected as part of the 2004 Malaysian Household Basic Amenities and Income Survey.

Specification (1) includes all state legislative districts for peninsular Malaysia and for Kuala Lumpur's Parliamentary districts.[17] It is immediately evident that out of the three types of geographical interpolation, IDW has the highest level of correlation, performing nearly twice as well as the best performer among the other two categories. Looking at figure 1.6, we can see that both the point estimate method and the buffer method leave a large number of districts without any IP addresses, even those adjacent to large cities. The IDW method does a better job of smoothing this information geographically, leading to a stronger measure.

[16]The process can be altered so that values are not calculated for areas over the sea. Since areas over the sea are not included in any calculations, this does not alter the results.

[17]Kuala Lumpur cannot vote in state legislative elections because it was ejected from the state of Selangor in 1974 and made into a Federal Territory.

Of the four methods for accounting for error due to dynamic IP allocation, the worst performer in specification (1) is the fixed estimator. This is the estimator that only counts IP addresses that are not dynamically reassigned to a new location over the twelve month period. The value is close to zero for the point estimate and buffer interpolation techniques, and is negative in the IDW case. This suggests that dynamically assigned IP addresses form too large a fraction of the IP base to be ignored.

Of the remaining estimators, IDW IPmax estimate performs the best with a .46 correlation coefficient. Recall that the IPmax estimator is the value from the month during which a location was assigned the maximum amount of IP addresses. This was meant to correct for smaller cities being undercounted, as they frequently disappear from the data for a few months at a time. This suggests that there is indeed a significant bias towards large cities in the data.

As argued above, the large city with the most potential for introducing bias is Kuala Lumpur. From table 1.1, nearly half of all IP addresses have been assigned to Kuala Lumpur at least once during the twelve month period even though the city only accounts for 7% of the population. In order to addresses this concern, in specification (2) I drop the largest city in Malaysia from the sample, Kuala Lumpur. The removal of Kuala Lumpur greatly increases every correlation coefficient except for IDW IPfix. This suggests that Kuala Lumpur was indeed heavily biasing the measure.

Once Kuala Lumpur has been removed, the top performer is the IDW IPsum estimator with .62 correlation coefficient as compared to .53 for IPmax. The IPsum estimator is simply the summation of the number of IP addresses assigned to a location over the twelve-month period. The danger with the IPmax estimator is that it overcounts small areas that are only present in the data for a few months. The IPsum estimator, however, takes this into account by aggregating all available information over the twelve-month period.

Given the drastic effects of dropping Kuala Lumpur from the sample, another concern is that the correlations are simply being driven by outliers. Figure 1.8 shows a linear graph of IDW IPsum estimator against the census data to assess to what extent outliers may be affecting the results. The figure suggests several large outliers as well as a large skew in the distribution to the left. In order to address this concern, I take the log of all the measures, which will draw in the outlying results.

Figure 1.9 shows the effect: there is a significant positive slope. Column (3) of table 1.2 gives the result of the log specification: almost all correlation coefficient increase. Again, IDW IPsum is the top performer by large margin.

In summary, the optimal measure of Internet penetration is IDW IPsum for the following reasons:

1. IDW interpolation to account for spatial measurement error.
2. Summation of the number of IP addresses assigned to a location over a twelve month interval to deal with measurement error from dynamically assigned IP addresses.
3. Take log to deal with outliers and skew in distribution.
4. Drop outliers that are heavily biasing results. In the case of Malaysia, Kuala Lumpur was introducing a huge amount of error: dropping the twelve observations corresponding to Kuala Lumpur increased the correlation of the top performing estimator by almost a third.

From here on, I will refer to the IDW IPsum measure in logs as IPperPerson.

1.4.1 Measurement error

The IDW IPsum estimator at .62 correlation, indicates a strong correlation, but still leaves a large, unexplained difference between the two measures.

There are several explanations for why this difference occurs. First, since the HBAIS data only counts households with Internet subscriptions, it most likely underestimates the percentage of households with access to the Internet by omitting people who access the Internet at work or in Internet cafes. IPperPerson should capture IP addresses tied to work and to Internet cafes. However, since IPperPerson contains no metric for intensity of usage, it would not take into account that hundreds of individuals might use the same IP address on any given day.

The other most likely reason for imperfect correlation is measurement error. This error could take the form of bias toward major cities, first because of the nature of the GeoIP database, which only counts city centers, and second because of the IDW interpolation technique, which treats the centers of major cities as peaks, with connectivity decreasing as we move outward. Thus one possibility is that IPperPerson is proxying for urbanization. I address this concern in the next section.

1.5 Robustness

In this section I test the extent to which IPperPerson is a good proxy for Internet penetration. Column (1) of table 1.3 presents a regression of fraction of households with Internet subscriptions on IPperPerson. As expected, the coefficient is large and highly significant.

A first concern is that the relationship between IPperPerson and Internet penetration is non-linear. However, specification (2) suggests that a quadratic relationship does not fit the data. This can also be seen in figure 1.9, where the quadratic fit is hardly different from the linear fit.

A second concern is that IP per person is picking up the effects of urbanization: all the geographic locations are large towns, and large urban areas are more likely to have large amounts of IP addresses. I test for this possibility in column (3) where I include controls for population density, road density, distance from the centroid of the district to the nearest major road, percent of the district that is urban vs. rural vs. jungle, and the ruggedness of the district, which I calculate as the standard deviation of the slope. The results suggest that IPperPerson is not simply proxying for urbanization: the coefficient on IPperPerson diminishes somewhat from column (1), but remains large and highly significant. As expected, population density explains some variation in Internet access. Distance to the major roads also has much explanatory power. This is likely because much of Malaysia's fiber-optic infrastructure runs along its major roads.

A third concern is that IPperPerson is proxying for wealth. Wealthier areas could have more companies using IP addresses for commercial purposes unrelated to the provision of Internet access (e.g. web hosting). As we can see in specification (4), the result is hardly changed by the inclusion of GDP per capita as a control.

A final concern with IPperPerson is that it is simply picking up overall engagement with electronic equipment such as phones, televisions, radios and computers. In column (5) we see that the coefficient of interest still retains a large amount of its explanatory power with the inclusion of these controls.

More importantly, this explanatory power is maintained even when a fraction of the population with computers and fixed lines are included. I interpret this to be the case because not all people with computers and fixed lines have Internet connections. Still, computer ownership has substantial explanatory power across specifications.

Could this be a better proxy than IPperPerson? The problem is that information on computer ownership is just as hard to find as information on Internet usage.

1.6 Conclusion

This paper shows how the number of IP addresses per person is a good proxy for Internet penetration. In order to assemble a strong measure a number of steps had to be taken. First, in order to deal with dynamically allocated IP addresses, the number of IP addresses assigned to a particular region should be aggregated over a year period. Second, because only the coordinates of the middle of a city are given, IDW interpolation helps to smooth the data spatially so that analysis can be performed at a regional level that doesn't correspond to the city level.

The measure was then tested to see whether it might be proxying for urbanization, wealth, or technological development and was found robust to all specifications.

There are many advantages to using the number of IP addresses per person as a measure of Internet penetration. First, it is the first measure that provides a consistent measure of Internet use at a sub-country level. Second the data has been collected for almost every country in the world using the same methodology, making it comparable across countries. For some countries, the data is far more comprehensive. In the U.S., for example, IP data is available at the zip code level, obviating the need for IDW interpolation. Similar levels of detail hold for wealthy European and East Asian countries.

This data can be used to understand a wide variety of phenomena. Miner (2012), for example, uses IP data to evaluate the effects of the Internet on elections in Malaysia. This IP measure can also be used to evaluate the impact on factors like GDP growth, firm productivity, and labor productivity, all of which are prime concerns for policy makers.

1.7 Tables and Figures

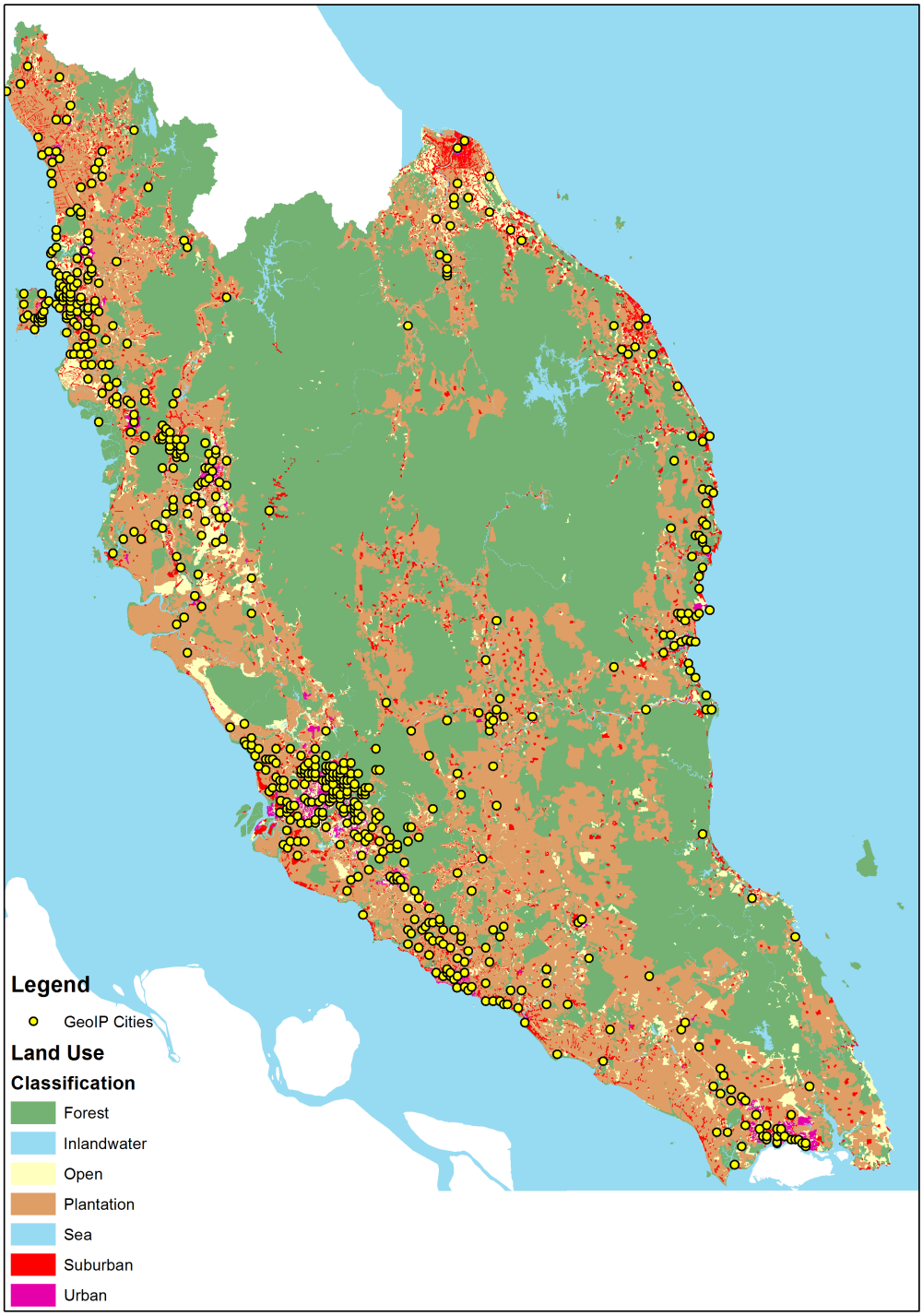

Figure 1.1: Geographic Placement of GeoIP Data

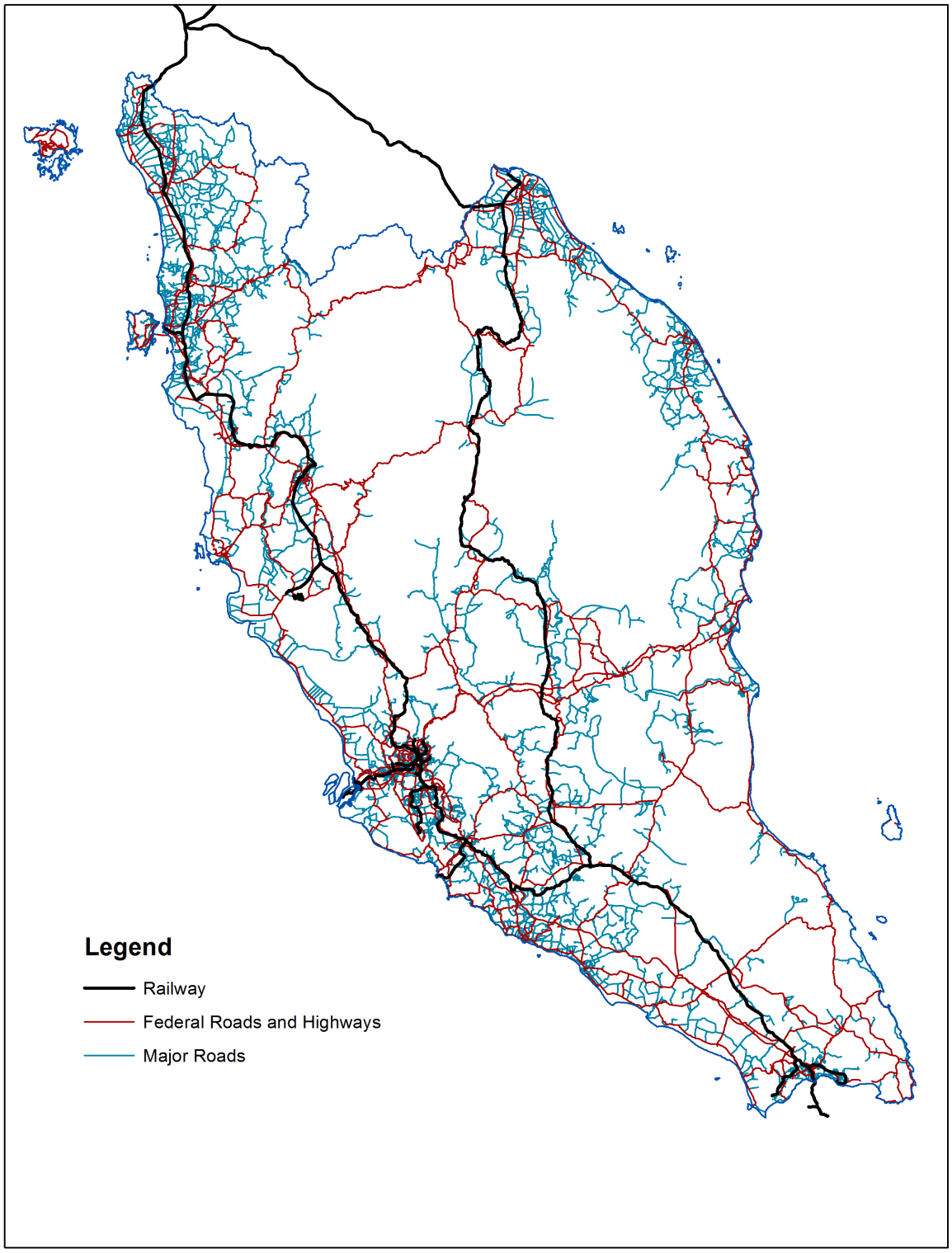

Figure 1.2: Peninsular Malaysia's Road and Railway Network

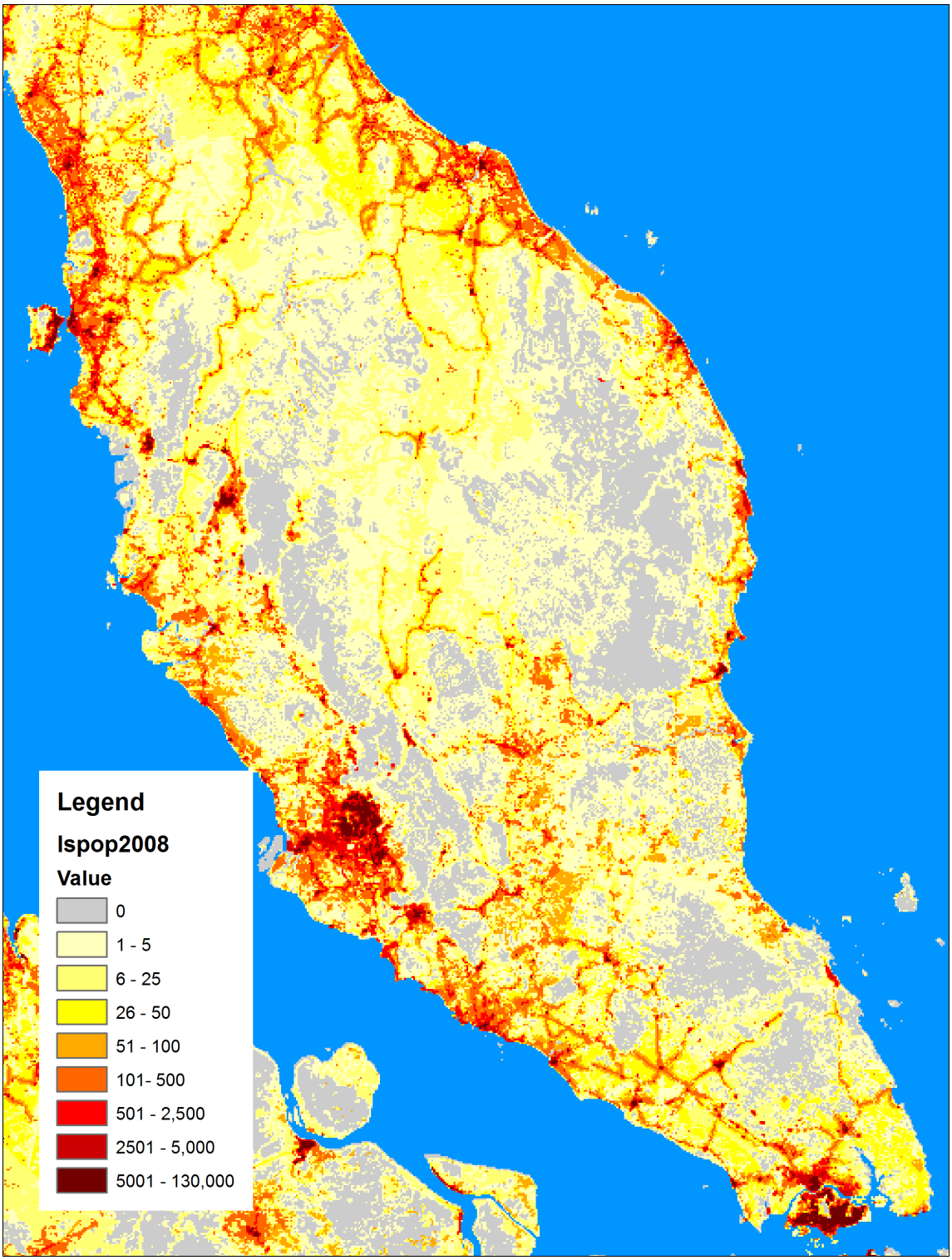

Figure 1.3: Population Distribution

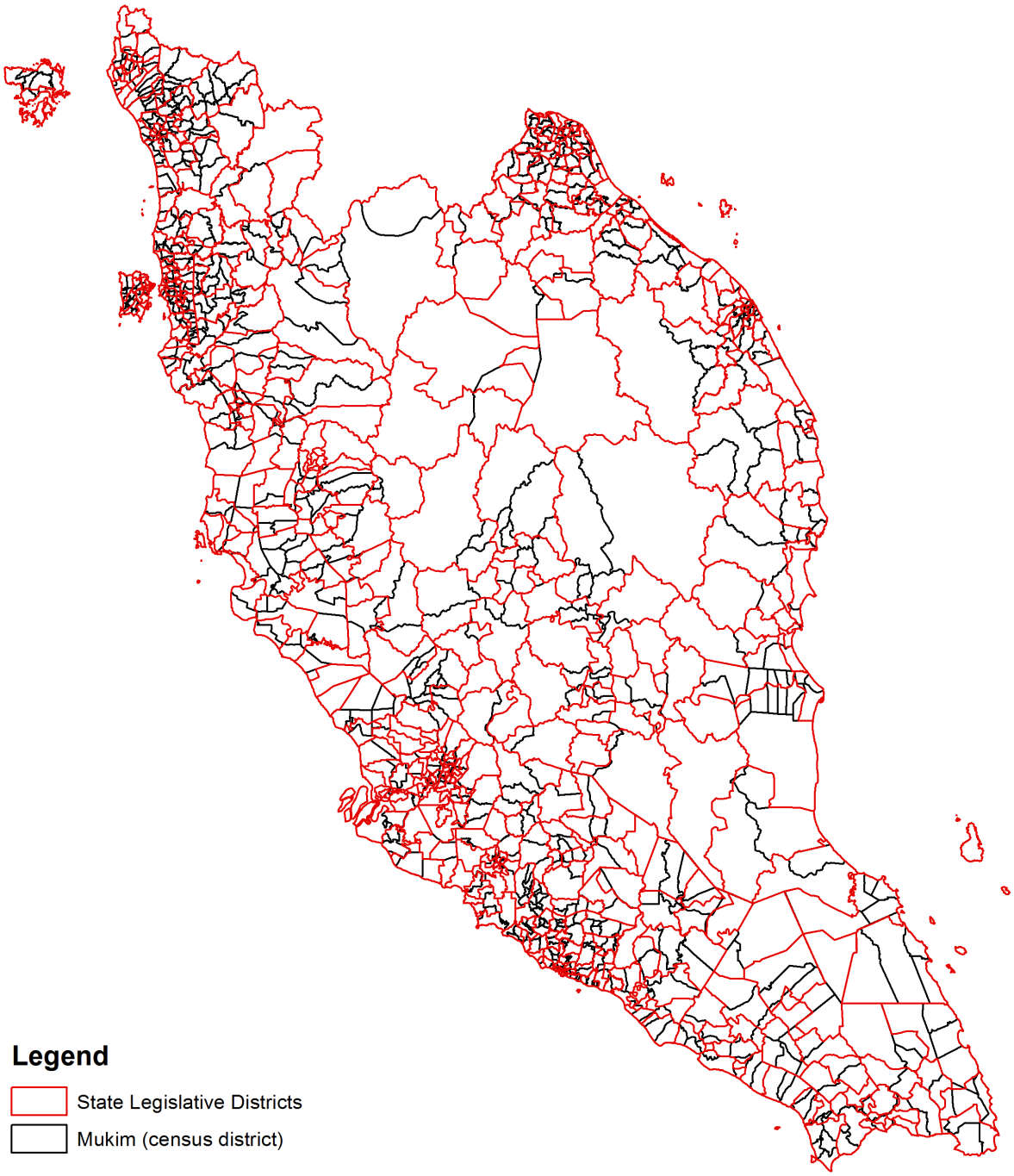

Figure 1.4: Legislative District vs. Census District Boundaries

TABLE 1.1

Measurement error in IP assignment

Panel A: IP Addresses Always Misassigned			
Assigned after 2004 (Pct)	27.19		
Never assigned location (Pct)	23.69		
Consumer ISP never assigned location	9.64		
Assigned to Kuala Lumpur (Pct)	23.06		
Panel B: IP addresses occasionally misassigned			
	No Location	Kuala Lumpur	
Assigned one month (Pct)	38.54	25.14	
Reassigned two months (Pct)	5.53	7.76	
Reassigned three months (Pct)	0.17	0.55	
Panel C: Reallocation of IP addresses over time			
	Mean	Standard Deviation	N
Number of times IP changes location	2.404648	1.354786	1476673
Mean months IP in same place	6.345744	3.720512	1476673
Maximum months IP in same place	8.446008	2.998054	1476673
Minimum months IP in same place	4.951242	4.649864	1476673
Months IP appears in data	11.62275	1.786861	1476673

Notes. The table presents summary statistics on IP address movement within the dataset over a one year period. Panel A looks at IP addresses that are never assigned to a geographic location. Panel B looks at IP addresses that occasionally are not assigned to a geographic location. Panel C shows how IP addresses that are assigned to geographic locations change location over time.

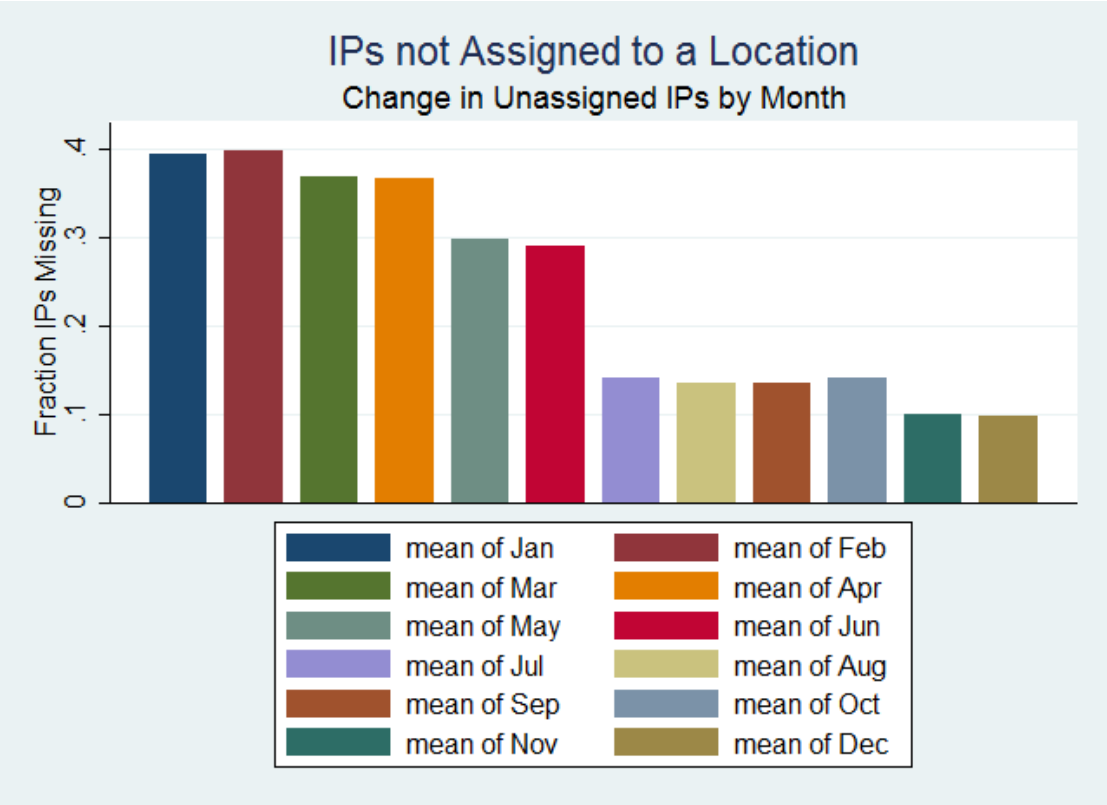

Figure 1.5: Fraction of Sample Not Assigned to Geographic Location

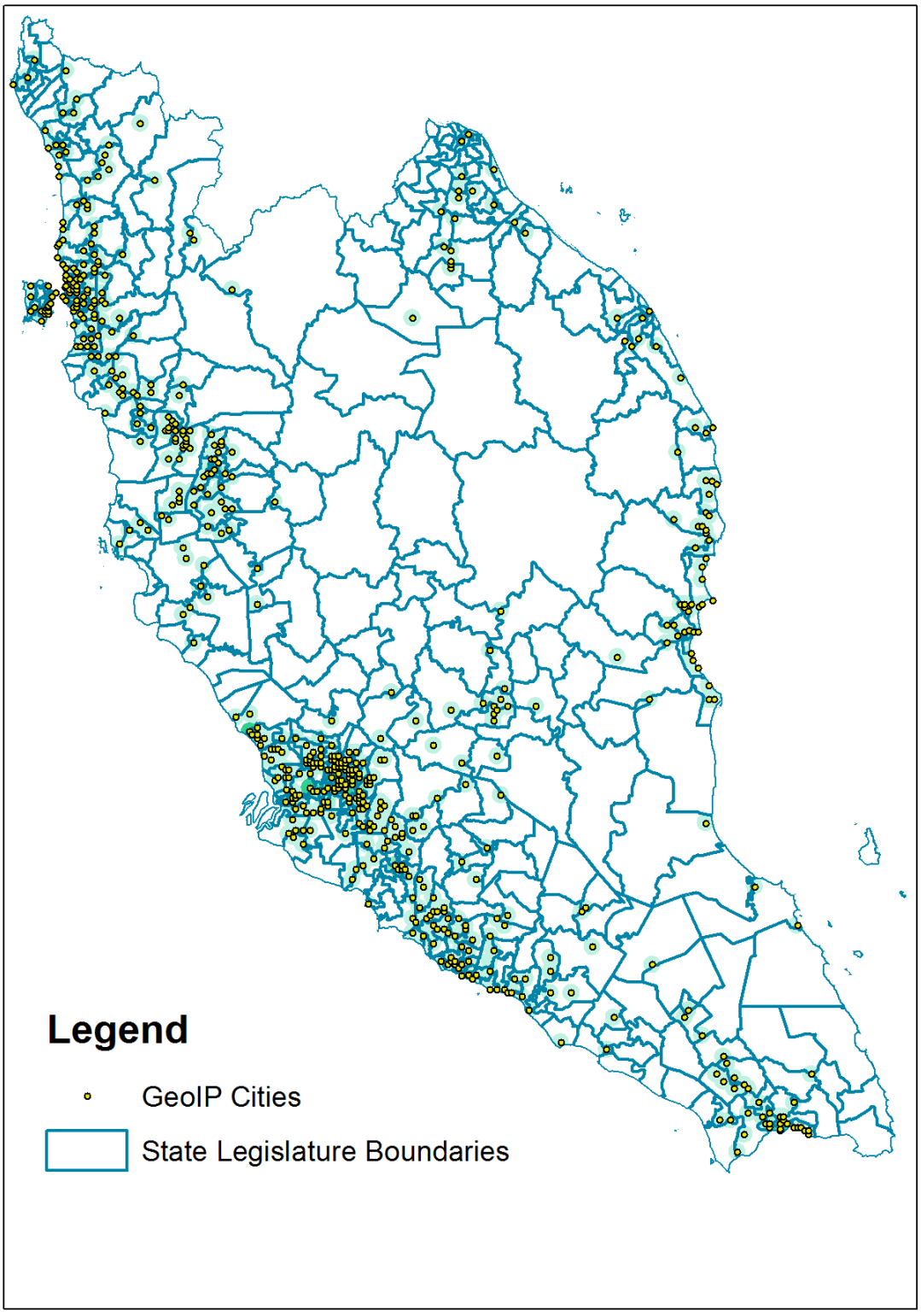

Figure 1.6: Interpolation of 25 Square Kilometer Buffer Zone

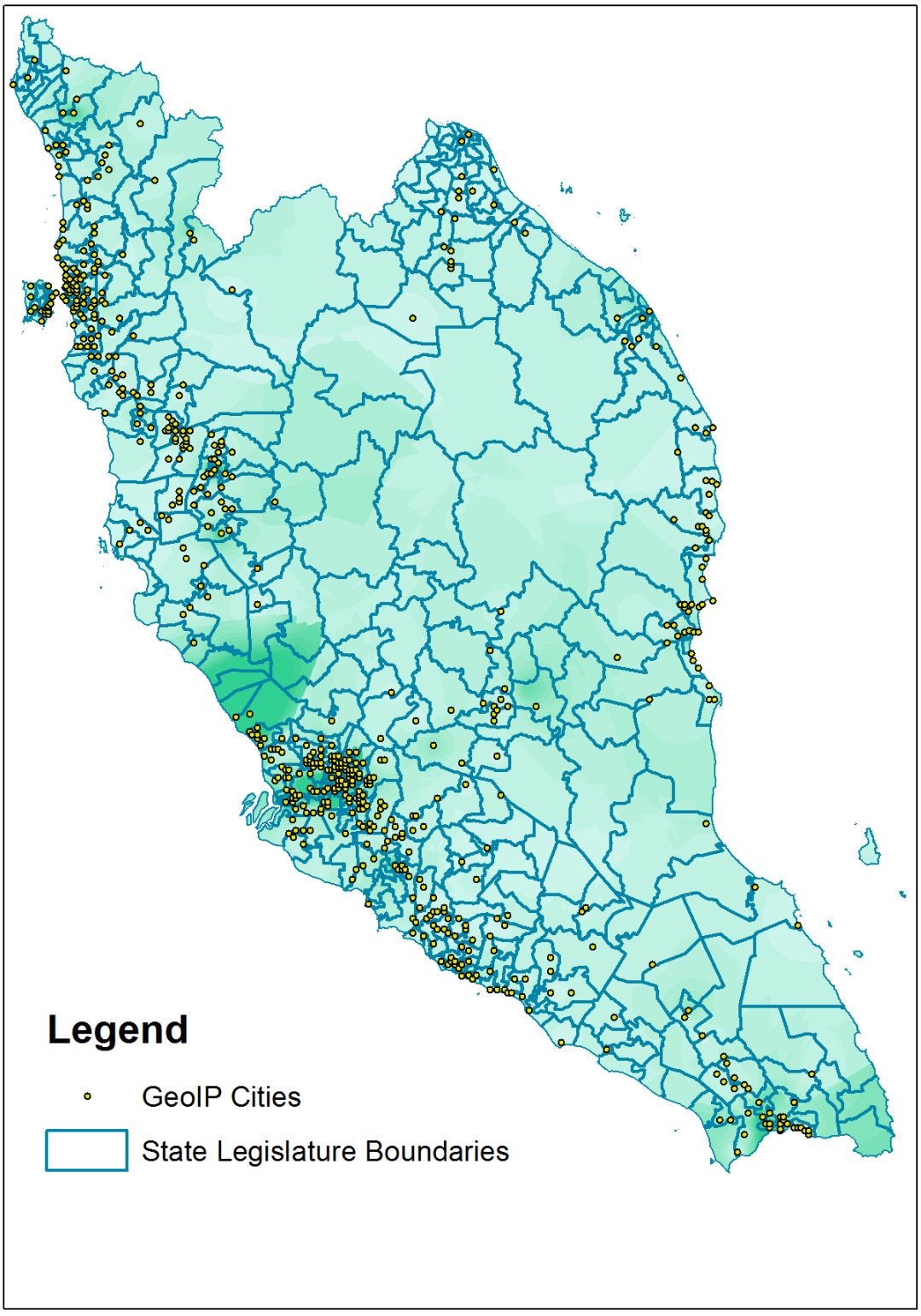

Figure 1.7: Interpolation: Inverse Distance Weighting Interpolation

TABLE 1.2
Evaluation of Internet penetration measures

	% Households with Internet 2004		
	(1)	(2)	(3)
Point estimate IPsum	.0918617	**.22665753**	**.29286847**
Point estimate IPavg	.09090891	.15557172	.19382615
Point estimate IPmax	**.15075733**	.22006066	.26132963
Point estimate IPfix	.05274528	.10717812	.11291494
Buffer estimate IPsum	.18926857	**.21265246**	**.39235348**
Buffer estimate IPavg	.17435006	.12764648	.17987505
Buffer estimate IPmax	**.24318167**	.20515007	.28516337
Buffer estimate IPfix	.12409332	.12403433	.12790432
IDW estimate IPsum	.36036258	**.62124655**	**.63251108**
IDW estimate IPavg	.31745242	.4920734	.49031368
IDW estimate IPmax	**.46292961**	.53265917	.55639539
IDW estimate IPfix	-.05517231	-.13518059	-.14431686
Kuala Lumpur	Y	N	N
N	460	445	445

Notes. Correlation between percentage households with Internet subscription 2004 and self-constructed Internet penetration measures. Percentage Households with Internet access in 2004 was derived from Household Basic Amenities Survey 2004. See Section 1.3 for details on the construction and source of variables.

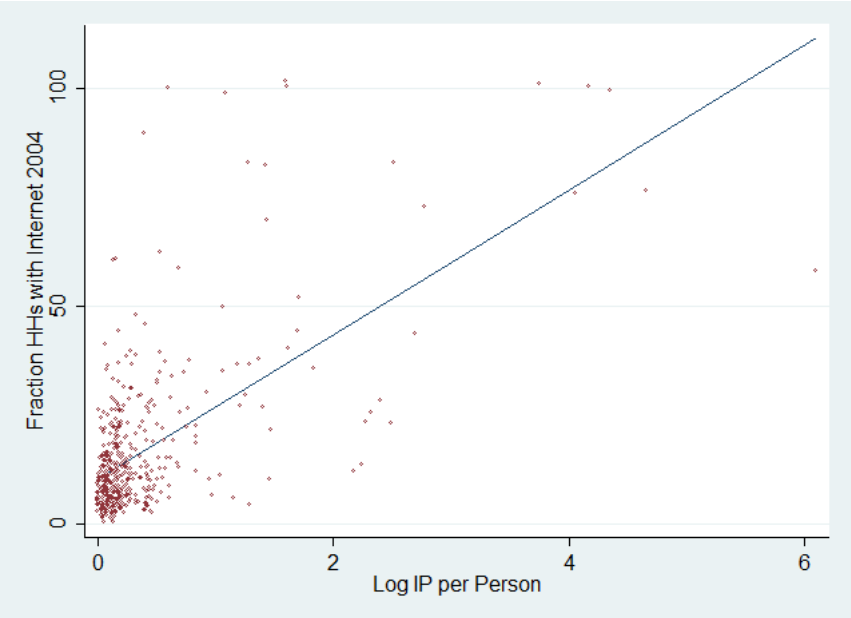

Figure 1.8: Linear IPperPerson Against Percent with Household Internet

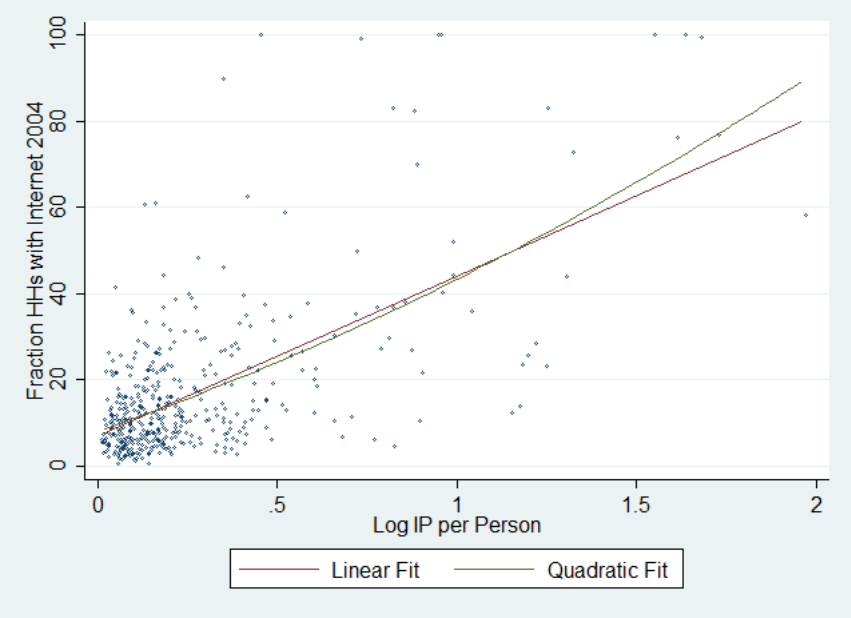

Figure 1.9: Log IPperPerson Against Percent with Household Internet

TABLE 1.3

Performance of IP per person with other covariates

	% Households with Internet 2004				
	(1)	(2)	(3)	(4)	(5)
Log IP per person	**30.944*** **	**21.625** **	**17.287*** **	**16.933*** **	**11.362*** **
	(3.289)	(8.597)	(3.722)	(3.775)	(3.652)
Log IP per person SQ		6.775			
		(6.095)			
Population density			4.862***	4.545***	1.333**
			(0.659)	(0.707)	(0.631)
% Urban			4.566	4.688	1.599
			(6.183)	(6.207)	(4.946)
% Rural			-2.311	-2.026	-2.250
			(3.403)	(3.439)	(2.947)
Road density			3.243	3.233	2.241
			(2.467)	(2.447)	(1.706)
Distance to major road			6.859***	6.627***	4.199**
			(1.805)	(1.835)	(1.667)
Ruggedness			0.907***	0.896***	0.310
			(0.288)	(0.285)	(0.241)
Log GDP PC				2.868	-1.223
				(2.077)	(1.801)
Computers					0.646***
					(0.065)
Telephones (fixed)					0.155***
					(0.058)
Mobile phones					-0.207***
					(0.039)
Television					-0.147
					(0.111)
Radio					0.264**
					(0.122)
N	433	433	433	433	433
R^2	.56	.563	.691	.692	.786

Notes. The table reports the results of regressing percentage of households with an Internet subscription on IPper-Person. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

1.8 Data appendix

Data Descriptions and Sources

Variable	**Description**	**Source**
Original Variables		
Eligible voters	Number of people eligible to vote in a district.	Election Commission (1986, 1990, 1995, 1999, 2004, 2008)
Generated Variables		
GDP	Measure of average GDP per capita at the mukim (census district) level, generated by the consultancy Booz & Company. Aggregated up to the state legislature district level using ArcGIS and LandScan as outlined in Section 1.2.1.	Booz & Company (2005)
% Household with Internet 2004	Fraction of households with an Internet subscription at the mukim level. Aggregated up to the state legislature district level using ArcGIS and LandScan as outlined in Section 1.2.1.	Household Basic Amenities and Income Survey (2004), Population and Housing Census (2000)
Ruggedness	The standard deviation of the average steepness of land in a district. Calculated from digital elevation satellite imagery using ArcGIS first to calculate the slope at each point and then to derive the average and standard deviation across a district.	Pitney Bowes (2008)
% Urban rural	Percentage of a district that is classified as urban and rural farm using satellite imagery. The omitted category is jungle. Used ArcGIS to calculate percentage at district level.	Pitney Bowes (2008)
Road density	Kilometers of road in a district divided by total area of the district. Calculated using ArcGIS.	Pitney Bowes (2008)

Population density	From Oak Ridge National Laboratory LandScan product, which uses census data in conjunction with satellite information to estimate population at the 1 km resolution. I use ArcGIS to aggregate up to the district level.	LandScan, Oak Ridge National Laboratory (2008)
Km to roads	Distance from the centroid of a district to the closest major road and closest federal road as of 2008. The road data is from Pitney Bowes. ArcGIS was used to calculate the centroid of each district and then derive the distance measure.	Pitney Bowes (2008)
IPperPerson	The number of IP addresses per eligible voter. Constructed with ArcGIS from IP geolocation data from MaxMind in conjunction with records from APNIC. See Section 1.3 for details.	MaxMind (2004-2008), APNIC (2004-2008)

2 The Unintended Consequences of Internet Diffusion: Evidence from Malaysia

2.1 Introduction

The Internet has grown enormously over the past two decades: from its DARPA roots in the U.S. Department of Defense, to today where it is a near-ubiquitous method of communication and information exchange in developed countries and—thanks to the rise of mobile telephony—a rapidly expanding technology in developing countries. Since its inception, much has been made of the Internet's potential as a democratizing force that frees information from the control of governments, implodes the distance between users around the world, and provides access to new viewpoints. Indeed, the Internet's ability to provide unfiltered access to information has caused consternation among many governments. This response has been notable in China, which has invested billions in keeping the Internet under tight rein. Social media has also been identified as a driving factor behind protests the world over, such as the recent revolutions across the Middle East. Despite a wealth of anecdotal evidence, however, little quantitative work has been conducted to test the ability of the Internet to foster democratization.

Malaysia serves as a particularly compelling test case in this regard. First, the ruling coalition, the Barisan Nasional (BN), enjoyed veto-proof control over all branches of government from 1969 to 2008. Although Malaysia holds regular democratic elections, the BN maintained power through strict controls on the judiciary, the police, and, importantly, the mass media.

Second, the BN's hold on power was so secure that it initiated an aggressive information and communications technology (ICT) led development strategy, based

on an uncensored Internet. The government has invested heavily in the ICT sector since 1996 as a means to promote growth and enjoys a very high rate of Internet penetration—60% as of 2008. At the same time, to attract foreign direct investment (FDI), the government pledged not to censor the Internet.

Third, since the Internet is uncensored, it has become home to a vibrant opposition blogosphere and a number of popular, independent news sites. In March 2008 the BN lost its two-thirds majority in parliament for the first time since 1969, as well as control of 5 out of 13 states. In the aftermath, commentators argued that the Internet played a leading role in this outcome by providing access to alternative viewpoints. In the rush to promote an information economy, the government overlooked the consequences with regard to political control. This paper tests whether Internet penetration influenced voting behavior in Malaysia, focusing on the 2004 and 2008 elections.

I develop a simple model to understand the mechanism by which the Internet influences election results. The model, building on Besley and Prat (2006), presents a retrospective voting framework in which the incumbent decides whether to buy off the media. This model shows how the Internet can influence electoral outcomes in regimes that use mass media control to ensure reelection. The key insight is that the Internet can weaken the ruling party's hold on power by undermining its ability to suppress negative information on candidate type. The model's main prediction is that areas with relatively higher Internet connectivity will experience lower vote shares for the incumbent party and higher turnover.

An important contribution of this paper is a novel measure of Internet penetration, which can be applied to almost any country. For most countries there are no geographically disaggregated measures of change in Internet access across time. To address this problem, I use a dataset that maps all of the IP addresses in Malaysia to approximate geographical locations. I aggregate the data up to the yearly period to deal with changes in assignment location across months. Next I use inverse-distance weighting interpolation to convert the data from the city level to the state legislature district level. Finally, I normalize by the number of eligible voters in a district to create the final measure. I find that this measure performs well when tested against census data from 2004.

To address problems of endogenous Internet placement and confounding political trends, I instrument for Internet growth. I calculate the shortest distance from each

electoral district to the backbones of Malaysia's main Internet Service Providers (ISPs). An increase in distance to the backbone leads to higher costs of supplying Internet connectivity (e.g., digging new trenches and laying cabling). This provides exogenous variation in Internet supply across districts. I exploit differences across ISPs in terms of geographical constraints on the placement of their backbones to argue that distance to the backbone is unlikely to affect voting outcomes directly, conditional on covariates. The identifying assumption is that conditional on baseline district characteristics (ethnic distribution, GDP per capita, population density) distance to the backbone does not affect change in vote share independently of growth in Internet access.

Based on the identifying assumption, I show a large, causal effect of Internet growth on election results: that the Internet can explain about one-third of the 11% drop in support for the BN from the 2004 to the 2008 election.

I run a number of checks on my identifying assumption. I show that distance to the backbone is uncorrelated with election swings in the pre-Internet period. By controlling for distance to roads and road density, I find that these instruments do not proxy for distance to roads. Lastly, I show that the instruments are not capturing the effect of distance to railroads: when I drop the backbone that runs along the railways, the result is unchanged.

Next I examine the effect of Internet growth on the turnover of politicians in the incumbent party. I show that Internet growth led to increased turnover, if baseline Internet access is sufficiently high.

Finally, I test to see whether the Internet affected voter turnout. I find that a one standard deviation increase in Internet growth corresponds to a 1.5% increase in participation.

To the best of my knowledge, this paper is the first to measure the Internet's effects on elections. It relates most directly to a large and growing strain of political economics literature on the complex relationship between media, government, and voters. From an empirical standpoint, Besley and Burgess (2002) provide evidence of a positive effect of the mass media on government responsiveness to natural disasters. Reinikka and Svensson (2004) show how a newspaper-based information campaign on education spending in Uganda improved education outcomes. Snyder and Strömberg (2010) exploit variation between newspaper markets and congressional districts to identify positive effects of newspapers both on voters' knowledge

of their representative and on federal spending in their district. In terms of broadcast mediums, Stromberg (2004) finds access to the radio significantly affected public spending during the New Deal. DellaVigna and Kaplan (2007) find that the emergence of the conservative Fox News Channel had a large impact on the 2000 U.S. Presidential elections.

While much has been written about the traditional media, very little empirical work examines the effect of Internet media. Gentzkow and Shapiro (2011) look at the effects of the growth of new media on ideological segregation in the U.S.A. I believe this paper is the first to provide causally interpretable evidence of the effect of Internet media on election outcomes and the first paper to concentrate on a developing country.

In terms of theory, this paper relates to Mullainathan and Shleifer (2005), which finds that increased competition in the media market could lead to increased bias as newspapers slant their news toward their readerships' priors, and to Baron (2006), which ties increased in bias to journalists' career concerns. This paper draws heavily from Besley and Prat (2006), which presents a theoretical framework for government capture of the media and shows how increased competition in the media market can yield better information on candidate quality and increased turnover. I extend their model by differentiating between traditional and web-based media. In modeling the effect of Internet-based media, this paper relates to Edmond (2011), which presents a model of media and regime change that distinguishes between print and broadcast media and online social media.

More broadly, this paper relates to literature on the effects of information technology on development. Jensen (2007) looks at the effects of the introduction of mobile phones on fish markets in Kerala; Goyal (2010) similarly analyzes the effects of Internet kiosks on crop prices in Madhya Pradesh; and Jack and Suri (2011) explore the impact of mobile payment on informal risk sharing.

Finally, this paper's empirical strategy pertains to literature exploiting geographic variation for identification. Geography has been used to identify the effects of dams (Duflo and Pande, 2007); electrification (Dinkelman, 2011; Barham, Lipscomb, and Mobarak, 2011); social capital (Olken, 2009; Paluck, 2009); ethnic violence (Yanagizawa-Drott, 2012); and the long-term effects of slavery (Nunn, 2008; Nunn and Puga, 2012).

I start in section 2.2 by outlining a general theoretical framework to help understand

the main mechanism at play. Section 2.3, shows how the model pertains to Malaysia, providing background on politics, media, and the Internet. In section 2.4, I describe my data sources, before outlining my method for constructing a measure of Internet penetration. Section 2.5 presents my empirical strategy results. I start by exploring the strong correlation that exists in the data and then move on to the issue of causality. In section 2.6, I examine additional outcomes and then conclude in section 2.7.

2.2 Theoretical Framework

In this section I develop a simple model for understanding how the Internet influences voting outcomes when all conventional sources of information are government-controlled. The model is based on Besley and Prat (2006), extended to account for differences between traditional media outlets (e.g., TV, radio, print) and web-based media outlets. The model shows how increased Internet access can lead to strengthened government control over traditional media outlets, but less control over the general flow of information. This weaker control over information in turn will lead to diminishing vote share for the incumbent and increased turnover. The model goes on to show how this equilibrium can hold even when more than half of voters have access to (uncensored) Internet-based media.

2.2.1 Basic Model

I use a two-period retrospective voting model. In the first period, an incumbent party is exogenously in power and is of two possible types $\theta \in \{g, b\}$ ("good" and "bad" respectively). A good party will deliver a benefit of one to voters; whereas, a bad party delivers zero benefit. The incumbent party's type is realized at the beginning of time and is good with probability γ.

Voters do not observe their payoffs when deciding to reelect the incumbent party and must rely on media reporting for information on type. Voters are distributed in a continuum of districts, which can vary along two dimensions. First, districts can differ in terms of the fraction of the population with Internet access ϕ_j in district j. Second, districts can differ in terms of population size. ψ_j represents

relative population size and is the fraction of total population in each district j. The incumbent party must win in a majority of districts to retain power.

The media comprises two firms: a mainstream media firm of type M, which we can think of as encompassing print, television, and radio; and a web-based firm of type W, which encompasses all Internet-based news sources without offline counterparts.[1] This could range from web-based news sites to blogs and Twitter feeds.

A key distinction between the mainstream firm and the web-based firm is that the mainstream firm can reach all voters; whereas, the web-based firm's potential audience is limited to the fraction of voters with Internet access. The mainstream firm may also have an online presence, but I assume they are more vulnerable to government capture due to their offline core (I provide a justification for this assumption in section 2.3.2).

If the incumbent party is good, neither firm observes a signal on quality. If the incumbent party is bad, firms receive a signal s that $\theta = b$ with probability $q \in [0, 1]$. Firms then can report either bad news or no news. As in Besley and Prat (2006), I assume that only verifiable information can be reported; it is not possible to fabricate bad news. I also assume that both firms receive the same information.

The media receives two types of payoffs: revenue from consumers and revenue from government payoffs. If a media firm reports bad news, the total audience-related revenue available in the market is a. If no news is reported, this revenue is normalized to zero. Viewers prefer informative news. If only the mainstream firm reports bad news it captures the entire market, earning revenue a and all voters are informed. If only the web-based firm reports bad news, its payoff is limited by Internet coverage $a\Phi$ and only fraction Φ of voters are informed. If both media firms report bad news, the mainstream news gets all offline consumers and splits the Internet-capable audience equally with Internet firms. All voters are informed and payoffs for mainstream and web firms are $a(1 - \frac{\Phi}{2})$ and $a\frac{\Phi}{2}$ respectively.

The incumbent party receives a payoff r for staying in office and can manipulate the news before the vote in the second period. This is modeled as a bargaining game between media outlets and the incumbent party. The party can offer non-negative transfers t_M (for the mainstream firm) and t_W (for the web firm). If a media firm accepts the transfer, they agree to suppress their signal and report no news. However, their payoffs for accepting are limited by transaction costs $\tau_i \in [1, \infty)$.

[1] This model can be extended to multiple firms of each type, without modifying the result.

For any given t_i, a firm receives only $\frac{t_i}{\tau_i}$. The incumbent's payoff is $r - t_M - t_W$ if reelected and $-t_M - t_W$ if not.

As noted in Besley and Prat (2006), transfers can take a number of forms, from all-out bribery to special perks for firms owned by the same company that controls the media outlet. The transaction costs will differ depending on factors such as legal institutions and the ownership type. They speculate that transaction costs should be lowest for state-owned firms and higher for independently owned media. In the context of Malaysia, history shows that capture is possible for the mainstream media, but that the web media is too costly to capture.[2] For simplicity, I model the prohibitive cost of capturing the web media with the assumption that the transaction costs for the web are infinite: $\tau_W = \infty$.[3]

The model's timing is as follows. We begin with an incumbent party in office whose type is realized with probability γ. If the party is of good type, all media outlets observe no signal. If the party is a bad type, all media firms observe the signal $s = b$ with probability q and $s = \emptyset$ otherwise. The incumbent party observes the signal that the media firms receive and chooses transfers t_i. Each media outlet observes its transfer and decides whether to accept. If it accepts, it reports $\tilde{s}_i = \emptyset$ and receives $\frac{t_i}{\tau_i}$. If it rejects, it reports $\tilde{s}_i = b$. Finally, voting is sincere. Each voter observes media reports and updates the posterior probability that the incumbent is good $\hat{\gamma}$. She votes for the incumbent if $\hat{\gamma} > \gamma$ and for the challenger if $\hat{\gamma} < \gamma$.[4]

2.2.2 Results when all districts are identical

I focus on a perfect Bayesian equilibrium, restricted to pure strategies in which voters always vote sincerely for the candidate they prefer. I start with the special case in which all districts are of equal size with identical Internet penetration rates. First, note that a bad incumbent party will never choose to capture either media outlet if $\Phi \geq \frac{1}{2}$, since by assumption τ_W is too high to capture the web media and web access is so widespread that a majority of voters will discover the party's type and vote them out.

[2] I provide evidence for this in section 2.3.2.

[3] τ_W doesn't need to be infinite, but rather high enough such that it is never profitable for the incumbent to capture the web: $\tau_W > \frac{r}{a\Phi}$.

[4] As in Besley and Prat (2006) sincere voting is assumed for analytical simplicity.

Proposition 1 *Assuming the Internet is too costly to capture ($\tau_W = \infty$) and all districts are of equal size with identical Internet penetration rates, equilibrium in the game overall is of two kinds:*

1. *A bad incumbent party will capture the mainstream media if $\Phi < \frac{1}{2}$ and $r \geq \tau_M a\left(1 - \frac{\Phi}{2}\right)$, but will not capture the web media. The party will win in all jurisdictions.*
2. *Otherwise, a bad incumbent party will not capture either outlet and will be discovered with probability q.*

Proof: see appendix.

To see this, note that since the web firm will never accept, the most that the mainstream firm can earn by deviating and not accepting the transfer is $a(1 - \frac{\Phi}{2})$ as outlined above. Thus, the mainstream firm will accept only if $t_M \geq \tau_M a\left(1 - \frac{\Phi}{2}\right)$ and, by extension, the bad incumbent will choose to make a transfer only if $r \geq \tau_M a\left(1 - \frac{\Phi}{2}\right)$. A notable implication is that the cost of capturing the mainstream media is actually decreasing with Φ as long as $\Phi < \frac{1}{2}$, reflecting the fact that the incumbent need capture only a majority of the market rather than the entire market.

If the mainstream firm accepts, its expected audience-related revenue is zero, instead of $qa\left(1 - \frac{\Phi}{2}\right)$ in the case in which neither firm is captured. In contrast, the web-based media's market share is $qa\Phi$ instead of $qa\frac{\Phi}{2}$. The mainstream media loses viewers to the web-based media.

Turning to voters, if the mainstream outlet is captured, viewers without Internet access will receive no information on candidate quality and will reelect the incumbent. If the fraction of voters without Internet is at least one-half, the incumbent will be reelected with certainty and turnover will be zero. If the mainstream firm is not captured, a bad incumbent will be discovered with probability q and the signal will be reported. Turnover, then, is simply the probability that the incumbent is bad and that the media receives a signal to this effect: $q(1 - \gamma)$

This yields the following implications.

Proposition 2 *Assuming Internet is costly to capture and districts are identical:*

1. *If $\Phi < \frac{1}{2}$ an increase in Internet access leads to*
   a) *Lower voting shares for the incumbent*

b) *No change in turnover*

c) *Loss of market share by mainstream media outlets*

2. *If* $\Phi \geq \frac{1}{2}$ *an increase in Internet access leads to*

a) *Lower voting shares for the incumbent*

b) *Turnover increases from* 0 *to* $q(1-\gamma)$ *across all districts.*

Proof: see appendix.

2.2.3 Extension: Internet penetration and population size vary across districts

In this section I relax the assumption that Internet penetration ϕ and population size ψ are identical across districts. I show the sufficient distributional assumptions needed for the incumbent party to capture the mainstream media and win a majority of seats even when Internet penetration across the country as a whole is greater than 50%.

First, consider the case in which only Internet penetration ϕ is allowed to vary and let its distribution be $f(\phi)$. If $f(\phi)$ is rightly skewed around $\phi = \frac{1}{2}$, it is possible to have a greater mass of districts with $\phi < \frac{1}{2}$, but a long right tale leading to $\Phi > \frac{1}{2}$. Under these conditions, government capture could be sustained with $\Phi > \frac{1}{2}$, since a majority of districts have Internet penetration rates lower than 50%. The maximum value of $\Phi > \frac{1}{2}$ for which media capture is still optimal, $\overline{\Phi}$, is increasing in the rightward skew of $f(\phi)$. [5]

If ψ (fraction of overall population per district) is allowed to vary as well, an equilibrium can be sustained where $\Phi > \frac{1}{2}$, the mainstream media is captured and there is no skew in $f(\phi)$. Suppose there is no skew and let $\bar{\phi} = \frac{1}{N}\sum \phi_i$ be the average measure of Internet penetration across districts. In the case in which population is identical across districts this value is the same as Φ (fraction of the country's population with Internet), but once population varies across districts it is possible for a gap to emerge between these two values $\delta = \Phi - \bar{\phi}$. δ can be thought of as the bias on the electoral effect of the Internet due to differences in population across

[5] Negative skew would lead to the opposite result: mainstream media capture could not be sustained even in a situation where nation-wide penetration rates are less than 50%.

districts. The sign and size of this gap depends on the joint distribution of ϕ and ψ. If Internet penetration and population size are positively correlated, δ will be positive and increasing in the covariance between ϕ and ψ.

I call the media "capturable" if any of the above conditions is met, such that the incumbent can win the election by paying off the mainstream media. This leads to the following results:

Proposition 3 *Suppose Internet penetration, ϕ, and population, ψ, vary by district, and $\Phi \geq \frac{1}{2}$. The equilibrium in the game is of two kinds:*

1. *A bad incumbent party will capture the mainstream media if: Internet penetration is rightly skewed around $\phi = \frac{1}{2}$ and/or ϕ and ψ are positively correlated such that $\phi < \frac{1}{2}$ for a majority of districts but $\Phi \geq \frac{1}{2}$. The party will win in all jurisdictions where $\phi < \frac{1}{2}$ and will retain a majority. The incumbent party will be discovered with probability q in districts where $\phi \geq \frac{1}{2}$.*
2. *Otherwise, the incumbent party will not capture either outlet and will be discovered with probability q.*

Proof: see appendix.

In section 2.4.3, I show that the distribution of Internet in Malaysia is rightward skewed and Internet access is higher in districts with higher fractions of the total population. This suggests the incumbent party can win elections by capturing the media even when Internet access is greater than 50% across the country as a whole.

To summarize, the model's main empirical prediction is that an increase in Internet access will cause a decrease in the incumbent party's vote share, in the presence of media capture. Intuitively, the Internet allows voters to circumvent media controls, and thus enables them to receive negative signals on candidate quality. The incumbent party's vote share will shrink as an increasing fraction of the population gains access to negative signals. I test this prediction in detail in section 2.5.

A secondary implication of the model is that Internet growth will yield higher turnover in districts where access is above the 50% threshold, but have no effect in districts with low levels of Internet access. Section 2.6 finds evidence corroborating this prediction.

2.3 Background

In this section I relate Malaysia to the theoretical framework above. I start by outlining Malaysia's political regime, move on to describe the state of the country's media sector, and finish with a discussion of its Internet.

2.3.1 Political regime

The model above provides a useful framework for thinking about the Internet's effect in Malaysia. Classified variously as "partly free"[6], a "flawed democracy"[7], and a "pseudo-democracy"[8], Malaysia's political regime combines democratic and autocratic elements.

Malaysia is a federation of thirteen states with a parliamentary system of governance. Elections are first-past-the-post and occur for both the national parliament and each state legislature. Since independence, Malaysia has been ruled by the same coalition in various guises, the Barisan Nasional (BN, or the Alliance prior to 1969). Though the BN includes parties representing minorities, most notably the Malaysian Chinese Association (MCA) and the Malaysian Indian Congress (MIC), it is effectively run by the United Malays National Organization (UMNO). The UMNO represents Malay and other "native" ethnic groups, known collectively as Bumiputera (meaning sons of the soil).

As in the model in section 2.2, the incumbent coalition has captured the print and broadcast media. Media ownership is concentrated in a handful of conglomerates that are controlled by the government, constituent members of the BN, and closely connected businessmen. For example, UMNO founded and controls the Utusan Group, which includes the *Utusan Melayu*, the oldest and most widely distributed Malay daily. Also, Media Prima, the largest media conglomerate in Malaysia, owns the largest English daily, two of the largest Malay dailies, four television channels, and three radio stations, and is itself controlled by business proxies of UMNO.[9] Although the traditional media show some variation in their level of bias, in general

[6]See Freedom House: <www.freedomhouse.org>
[7]See Economist Intelligence Unit: <www.eiu.com>
[8]See Case (2001)
[9]See Abbott (2011) for a more detailed discussion of the ownership structure of the media.

they tend to under-report on opposition candidates and downplay scandals.[10]

In addition, strict legal restrictions on media outlets prevent the emergence of any mainstream outlet that is overly critical of the government. First, media firms can only operate with a permit and face tightly controlled distribution. Opposition parties are denied permits to publish newspapers, even though constituent members of the BN control multiple media outlets. In the past, publications with critical views of the BN have quickly lost their operating permits and have either shut down or changed ownership.[11] Second, laws such as the Sedition Act, the Control of Imported Publications Act, and the Official Secrets Act allow the government to censor material with impunity.[12] Finally, the Internal Security Act, enacted in 1960 to fight a Communist insurgency, allows detention without trial for up to two years and can be renewed indefinitely.[13]

As in the model, population sizes vary greatly across districts. Apart from media control, the BN has effectively used redistricting as a means to maintain power. This trend, evident across the country at both the state and the parliamentary levels, tends to grossly over-represent rural areas at the expense of cities. Many rural areas act as "vote banks" for the BN, where allegations of vote-buying abound.[14] The period between 1986 and 2008 alone saw three redelineation exercises, with parliamentary seats rising from 177 to 222 and state legislature seats (excluding Sabah and Sarawak) rising from 351 to 455.

Malaysia has had opposition parties since independence, but only in recent years have they posed a real threat to the BN's hegemony. The rise of a viable opposition can be traced to the late-1990s, when a split between then-Prime Minister Mahathir Mohamad and Deputy Prime Minister Anwar Ibrahim led to Anwar's sacking and subsequent imprisonment under charges of sodomy and corruption. Anwar, once ejected from government, founded an opposition movement called *Reformasi*. After his imprisonment, the movement coalesced into a political party and, following a name change and a merger, is now known as the PKR (People's Justice Party). The other members of the opposition are the Democratic Action Party (DAP), a secular

[10] See Centre for Independent Journalism Malaysia (2008) or <www.malaysiakini.com/news/168567> for specific examples.

[11] See Kua (1990) for examples.

[12] For example, the Home Ministry censored an article in the July 16th, 2011 issue of the Economist on an electoral reform rally.

[13] As of September 2011, an announcement was made to reform these laws. See section 2.3.2 for details.

[14] See Pepinsky (2007)

party backed mainly by Malaysian Chinese, and the Pan-Malaysian Islamic Party (PAS), an Islamist party supported largely by Malays in the north of peninsular Malaysia. The PKR, DAP, and PAS contested the 1999 and 2004 elections as part of the Barisan Alternatif (BA), but disbanded the coalition after dismal losses in 2004.

Their fortunes changed dramatically in 2008. The parties wrested control of 5 out of 13 state houses and deprived the BN of its two-thirds majority in parliament. As we shall see below, there is good reason to believe that the Internet played a role in this outcome.

2.3.2 The Internet and politics

The second essential parallel to the model lies in Malaysia's treatment of the Internet. In stark contrast to print, radio, and television, the Internet has never experienced significant censorship.

By 2008, Malaysia had a very high rate of Internet access. Internet users comprised 56% of the population, compared to 24% in Thailand and 75% in Japan.[15] Malaysia's high Internet penetration rate stems from Mahathir's decision, in 1996, to invest heavily in ICT infrastructure as a way to foster a knowledge-based economy. At the forefront of this effort was the creation of the Multimedia Super Corridor (MSC), a high-tech zone south of Kuala Lumpur. This project entailed large-scale investment in constructing a brand-new "high tech" city called Cyberjaya and two new universities. It formed part of the MSC's primary goal: to attract multinational companies through tax breaks and first-class infrastructure. Most important, to make the location more attractive to FDI, the government signed an Internet "bill of rights", pledging not to censor the Internet.[16]

As a result, the Internet is the only platform available for alternate view points and has become an important source for independent news and opposition news and views. As early as 1999, members of the pro-Anwar Reformasi movement used blogs and newsgroups to spread their message, and there is some evidence that the Internet influenced the 1999 general elections.[17] Although Reformasi sites ebbed in subsequent years, the opposition continued to dominate the web as opposition

[15]See World Bank Development Indicators: <data.worldbank.org>
[16]See MSC Malaysia Bill of Guarantees at <www.mscmalaysia.my.
[17]See Zinnbauer (2003)

lawmakers joined citizen bloggers to try to reach a wider audience. The *Harakah Daily*, a PAS-owned news portal, represents the most ambitious online effort by an opposition party; it effectively allows the PAS to distribute a newspaper without obtaining a permit. In addition, a number of independent online news sites, the most famous being Malaysiakini, tend to favor the opposition. Finally, services such as YouTube, and increasingly Twitter, have made it easy to spread the word about political scandals and protest movements.

In terms of the model, which requires that information be verifiable, the Internet's most salient feature is its ability to provide information about scandals that previously would have been suppressed. The best example, the V.K. Lingam video uploaded to YouTube in late-2007, showed high-level officials engaged in judicial fixing for the Supreme Court. The video received millions of hits in a matter of days and erupted into one of the defining issues of the 2008 elections. In line with the model's predictions of effects on the media market, evidence suggests that the popularity of the captured media declined. According to a 2010 survey by the Merdeka Center, an independent polling organization, only 40% of Malaysians trust the mainstream media, down 20% from a similar poll conducted two years earlier.[18]

An important assumption in the model is that the Internet is too costly to capture. This assumption holds true in Malaysia for a number of reasons. From a purely economic standpoint, censorship scares away the FDI needed for Malaysia's ICT-based growth strategy: as user data migrates from the desktop to servers in the cloud, multinationals are loath to expose themselves to governments that try to limit how they can use this data.[19]

The Internet is also physically more difficult to regulate than other forms of media. In Malaysia, most opposition content is hosted on international platforms such as Google (Blogger and YouTube), Twitter, and Facebook. As Google's exit from China and all platforms' response to the Arab Spring convey, these platforms tend to dislike engaging in self-censorship. As a result, even if independent news sites were shut down, users could still access content hosted abroad and add to it by posting anonymously. Efforts could be made to censor the Internet, as in China's "Great Firewall", but they would be prohibitively expensive and ultimately ineffectual. In

[18] See <www.merdeka.org>

[19] For example, Blackberry maker RIM, has fought against attempts by various governments to access their encrypted network. See <http://online.wsj.com/article/SB10001424052748704017904575409093226146722.html>

terms of expense, such censorship would require substantial investment in not only physical capital but also human capital.[20] Internet censoring would also prove futile; it is always possible for users to leak information,[21] browse anonymously,[22] or bypass firewalls—even ones as sophisticated as China's.[23]

There are signs that the inherent difficulty in controlling the Internet is starting to have an effect on government policy. In a speech delivered in August 2011, the current prime minister of Malaysia, Najib Razak, stated: "In today's borderless, interconnected world, censoring newspapers and magazines is increasingly outdated, ineffective and unjustifiable."[24] In September 2011, he went on to announce plans to reform Malaysia's media laws and repeal the Internal Security Act, which allows for detention without trial. It remains unclear if this announcement will translate into meaningful reforms.

2.3.3 Internet placement

Official sources state that the primary motives for building ICT infrastructure were, first, to help Malaysia attain the status of a developed nation[25] and, second, to help promote a Bumiputera business class.[26] In practice, only the first of these goals seems to have played a serious role, with geographical costs being equally important.

The current state of Malaysia's ICT infrastructure originated in efforts to liberalize the telecommunications sector in the early-1990s. The government listed the public telecom, Telekom Malaysia (TM), on the local stock exchange (and retained majority ownership); it issued licenses for private telecoms; and it established a new, independent regulator. The liberalization process was poorly planned, with a large number of licenses issued in a very short period to well-connected businessmen. Hungry for profits, but lacking experience in the ICT sector, these private operators

[20] China has an army of Internet police whose sole job is to peruse forums, blogs, search results, etc. for objectionable content. See <http://www.guardian.co.uk/technology/2005/jun/14/newmedia.china>

[21] Either by posting anonymously to services like YouTube, or to organizations like Wikileaks.

[22] Tor is the best known program for anonymous browsing. See <http://www.torproject.org/>

[23] Ultrasurf, for example, allows users within China to circumvent Internet filtering by routing their connection through proxy servers.

[24] See <http://www.economist.com/node/21526885>

[25] First seen as early as 1991 in Mahathir's Vision 2020 development policy.

[26] This is laid out as an explicit aim in the National Telecommunications Plan, 1994.

went on an uncoordinated infrastructure building spree. Mass bankruptcies ensued in the 1997 Asian financial crisis.[27]

The country emerged from the crisis with a high but uneven level of connectivity: redundant infrastructure in some areas, and a lack of basic telephony services in others. To address infrastructure redundancy, the government encouraged consolidation and infrastructure-sharing. Figure 2.7 gives an idea of the state of Malaysia's Internet infrastructure from 2000 onward, showing the three largest Internet backbones currently in operation:[28]

1. *Telekom Malaysia (TM)*: TM, the state-owned incumbent, has the most coverage and capacity. It accounts for just over half of all private Internet connections between 2004 and 2008.[29] TM also sells capacity to ISPs that lack extensive physical infrastructure. These ISPs complain, however, that the rates that TM charges are too high for them to compete with TM's own services, especially in areas that are not served by other backbones.[30] Interviews suggest a mixed view of government involvement in TM's placement decisions. On the one hand, the government expects TM to perform the bulk of the heavy lifting to bring infrastructure to remote areas. On the other hand, in the wake of costly bankruptcies after the Asian financial crisis, the government has placed increased emphasis on TM turning a profit. Interviews with planning engineers suggest that demand and geography were the primary factors in infrastructure development since 2000.

2. *Time dotCom (Time)*: Time, a private company, has its own ISP geared toward consumers and businesses. Like TM, it sells excess capacity to ISPs that lack physical infrastructure. As shown in the figure, Time covers less area than TM and overlaps almost completely with TM's network. The red points in figure 2.7 are landing stations connected by submarine cabling, which provide network redundancy. There is no evidence of government involvement in Time's placement decisions. If anything, Time went against government wishes by over-investing in redundant infrastructure. As a result, the govern-

[27]For a complete analysis of the liberalization of Malaysia's telecom sector see Salazar (2007) chapter 7.

[28]Malay's fourth major backbone, Fibrecomm, runs along Malaysia's major power lines. However, this could not be included due to a lack of reliable GIS data.

[29]Budde 2009, Malaysia Internet Services.

[30]In fact, the governmental authority in charge of policing the ICT industry has found TM guilty of anti-competitive behavior, but has yet to take any action. See MCMC 2005.

ment had to rescue Time from bankruptcy after the financial crisis (when the current outlines of its present network were already set).

3. *Fiberail*: Fiberail's ownership is split between TM and Malaysia's public railway service. As the name suggests, Fiberail's backbone runs the length of Malaysia's major railways, which were completed in 1931. Given its stake in Fiberail, TM uses Fiberail's network extensively. However, Fiberail positions itself as independent from TM, and sells capacity to ISPs and major corporations. Founded in 1995, Fiberail's initial business activities were restricted geographically to companies with points of presence (access facilities) within a narrow corridor around the railway. In 2006, its license changed such that it could operate throughout the country. In February 2006, Fiberail acquired Petrofibre, a fiber-optic network spanning Malaysia's main gas pipelines.[31] It was impossible to include this additional information on the map, however, as reliable GIS data on pipeline locations are not publicly available.

Annual reports, consultant reports, and interviews concur that cost plays a central role in governing placement, and defining a few key terms will help provide a sense of those costs. ISP backbone refers to the trunk lines, nodes, and routers that form the core of an ISP's network. Linked by bundles of fiber-optic cables, which provide high speed and capacity, backbones are constantly upgraded and occasionally expanded. The backbones form only a part of the network that connects a user to the Internet, however. When a user logs on to the Internet, for example, the signal must first travel along a length of cable (usually copper), which connects the user's location to a local exchange on the edge of the ISP's network. This first step is often called the local loop or the last mile. The signal then travels along a backhaul connection (normally cabling) until it reaches an access point to the ISP's backbone, called a point of presence. Depending on the size of the ISP, the signal may need to pass through several other ISP networks before reaching the Internet. Alternatively, the ISP itself may be directly connected to the Internet via, for instance, an intercontinental submarine cable.

The costs of delivering the Internet to consumers can be divided into several categories. First is the cost of installing the backbone. Geographical and legal factors are the main impediments to backbone placement. In terms of geography, costs include digging trenches so that the fiber-optic cabling can be laid underground.

[31] Budde 2009, Malaysia Telecommunications Infrastructure.

These trenching costs depend on the terrain: it is much more expensive to lay fiber-optic through a jungle than alongside a road. All three backbones therefore follow preexisting routes: roads and highways in the case of TM and Time, and railways in the case of Fiberail. In terms of legal impediments, firms must obtain licenses to run cabling and erect infrastructure. Most land-based trunk cabling runs along federal and not state roads, since it is much less costly and time-consuming to secure a license from the federal government than from state governments. There are substantial differences between state and federal roads. Federal roads tend to be larger than state roads. Whereas the bulk of peninsular Malaysia's federal road system was built by the British before independence in 1957, the state road system continues to grow rapidly.

Once the backbone has been laid, plenty of supplementary costs must be incurred before an ISP can deliver its service to consumers. To serve a new area, an ISP must install a local switch and connect it to the backbone via backhaul cable. This step adds further trenching costs, which increase with distance to the backbone. It also entails the costly and time-consuming process of getting permission from local authorities. Even TM, which owned an extensive telephone network before the advent of the Internet, faces these costs. TM had to upgrade much of its copper wire to carry data signals, and dig up and replace its backhaul cable with fiber-optics to provide the extra capacity and speed needed to delivery Internet.

2.4 Data

2.4.1 Political Data

Malaysia is a federation of ex-British colonies. It is split between peninsular Malaysia, which gained independence in 1957 and houses most of the population, and Sabah and Sarawak, two less developed states on the island of Borneo that joined the federation in 1963. This paper uses election data at the state legislature level for the 1986, 1995, 1999, 2004, and 2008 elections. State elections are held at the same time as elections for the national parliament, with the exceptions of the two states on Borneo, Sabah and Sarawak. Sabah harmonized its state elections with the parliamentary elections only in 2004, and Sarawak continues to hold its state elections on off years. The data includes candidate names, parties, and votes along with turnout,

the number of eligible voters in a district, the number of rejected votes, and the district's ethnic composition. I have manually entered each set of electoral boundaries into ArcGIS to account for changes in district size and number since 1986.

Figure 2.1 shows state and parliamentary electoral district boundaries for the 2004-2008 period. No redistricting occurred in this period. Parliamentary district boundaries perfectly match state legislative district boundaries, with each parliamentary district comprised of two or three state districts.

Table 2.1 provides summary statistics covering the 2004-2008 period for state legislature districts in peninsular Malaysia, excluding Kuala Lumpur. The 2008 election is marked by a large drop in vote share for the BN and a modest increase in turnout. The number of eligible voters varies significantly across districts, with a mean of 18,000 and a standard deviation of around 7,000.

2.4.2 Demographics and geography

I have complete geospatial data for Malaysia. Figure 2.2 illustrates clutter data (which classifies all land as either urban, semi-urban, plantation, jungle, inland water, or open) and elevation data (which allows for the calculation of land-gradients). Figure 2.3 shows the locations of all major roads, highways, and railways in Malaysia. Finally, figure 2.4 represents data from the LandScan service, which estimates population distribution at the one square kilometer resolution through a combination of census data and satellite imagery.[32]

Table 2.1 helps make sense of the geo-spatial data. State legislature districts, on average, are 21% urban and 50% farmland (rural), with the remaining 29% classified as jungle. Although jungle covers large swaths of the country, the fairly extensive road network spans more than 80,000 kilometers of roads as of 2007.

I have constructed a dataset of controls using the Population and Housing Census of Malaysia for 1980, 1991, and 2000; Malaysia's Household Basic Amenities and Income Survey for 2004; and geographically disaggregated measures of GDP per capita 2005, generated by the consultancy Booz & Company. Unless otherwise stated, this data is available at the level of Malaysia's 927 census districts, called *mukim*.

[32]See <http://www.ornl.gov/sci/landscan/> for details on the construction of this dataset.

Figure 2.5 shows mukim boundaries alongside state legislative district boundaries. As can be seen, mukim level data does not match up perfectly with state legislative districts. To address this discrepancy, I use the LandScan population data to assign a weight to each one kilometer cell within each mukim. State electoral district values are generated from the weighted sum of these one square kilometer cells.

Malaysia is a multi-ethnic society. Ethnic Chinese, the wealthiest group in Malaysia, comprise 26% of the population. Brought in by the British as indentured servants to work in the country's rubber and palm plantations, Indians currently comprise roughly 8% of the population. The remainder (65% percent) is largely Malay, except for several ethnic groups on Borneo and a few small tribes.

2.4.3 Internet

I use official Internet measures from the Population and Housing Census 2000 and the Household Basic Amenities and Income Survey (HBAIS) for 2004. Both datasets provide the fraction of households with Internet subscriptions at the mukim (census district) level. As explained in section 2.4.2, I use ArcGIS to aggregate the HBAIC data to the legislative district level, which introduces some measurement error.

In the model in section 2.2.3, I presented two sufficient conditions for media capture when average Internet access is greater than 50%. The first condition is rightward skew in the distribution of Internet connectivity by district. Figure 2.8 shows an approximation of the PDF for Internet subscription per household variable alongside the PDF of a normal distribution. As can be seen the distribution is severely rightward skewed. The second condition is a positive correlation between Internet penetration and the fraction of total population. Figure 2.9 graphs a scatter plot of the log of households with Internet subscriptions in 2004 against the fraction of total eligible voters in a district. Showing a strong positive relationship between these two variables, this graph implies that districts with larger populations also have higher Internet penetration per capita.[33]

Since the census data does not cover the 2008 period, I turn to two extra sources of data on Internet connectivity. The first, the GeoIP City database, is produced by the geo-location company MaxMind. GeoIP City is a service that matches IP

[33]These results hold for alternate measures of Internet penetration for 2004 and 2008 explained in section 2.10.

addresses to geographical locations, allowing web services to tailor advertisements based on visitor location and to detect fraud. The GeoIP City database comprises monthly data from 2004 to the present and covers virtually all IP addresses in the world.[34] For each IP address assigned to Malaysia, the GeoIP City database provides the name and location of the nearest large city on a monthly basis. Figure 2.6 shows the spatial distribution of GeoIP data points for 2008. There are 782 locations that appear in the data for 2004 and 487 for 2008. Although the 2008 data has fewer locations, it has roughly twice as many IP addresses, reflecting the enormous growth in Malaysia's Internet penetration in the 2004-2008 period.

My second data source comes from APNIC (Asia-Pacific Network Information Center), the regional Internet registry responsible for delegating blocks of IP addresses to national Internet registries, ISPs, and large companies in the Asia-Pacific region. As such it has a complete record of all IP blocks allocated to Malaysia along with the recipient of the block (normally an ISP) and the date of allocation.

The GeoIP City database is used in conjunction with the APNIC dataset, which together identify: the initial date of IP assignment to Malaysia, the ISP managing the IP addresses, and the IP blocks location(s) during the 2004-2008 period. I create the measure $IPperVoter$ by aggregating this data up to the electoral district level and then normalizing by the number of eligible voters. This gives the number of IP addresses per voter in each state legislature district and is expressed in logs. I check $IPperVoter$ for 2004 against official government statistics on the percentage of households with Internet subscriptions in 2004, and find a high correlation of .63. Section 2.10 of the appendix provides a full account of the methodology employed and the likely sources of measurement error.

2.5 Empirical Analysis

2.5.1 Basic Correlations: OLS Estimates

I start by examining the basic relationship between Internet penetration and the BN's share of the vote at the state legislature district level. Figure 2.10 plots change in voting share for the BN and growth in $IPperVoter$ during the 2004 to 2008 period.

[34] MaxMind does not cover IPv6 addresses. However, IPv6 adoption was infinitesimal in Malaysia at the time.

As can be seen, there is a strong negative relationship in the raw data, implying that areas with more Internet growth are associated with greater negative swings against the BN.

I explore this relationship in more detail by controlling for other characteristics that might affect changes in BN voter share. Let y_{ist} be BN's vote share for legislative district i in state s at time t and $\triangle IPperVoter_{ist}$ be growth in IP addresses per voter:

$$y_{ist} = \alpha_0 + \alpha_1 t + \alpha_2 IPperVoter_{ist} + \rho_i + \delta_i t + \mu_s + \lambda_s t + \varepsilon_{ist} \tag{2.1}$$

Where ρ_i is the district fixed effect, δ_i is the district trend,μ_s is the state fixed effect, λ_s is the state trend, and ε_{ist} is an idiosyncratic error term. This equation, in turn, can be rewritten in first differences, eliminating ρ_i and μ_s:

$$\triangle y_{ist} = (y_{ist+1} - y_{ist}) = \alpha_1 + \alpha_2 \triangle IPperVoter_{ist} + \lambda_s + (\delta_i + \triangle \varepsilon_{ist}) \tag{2.2}$$

With two periods of data, it is not possible to estimate the legislative district specific trend δ_i. OLS estimation of equation (2.2) will be biased as long as $\delta_i + \triangle \varepsilon_{ist}$ is correlated with $\triangle IPperVoter_{ist}$, which we would expect if Internet is allocated to areas that are trending for or against the BN for unobservable reasons.

As a first pass, I augment equation (2.2) with a vector of legislative district covariates (X_{is}) to control for some factors that might affect δ_i:

$$\triangle y_{ist} = \alpha_1 + \alpha_2 \triangle IPperVoter_{ist} + X_{is}\beta + \lambda_s + (\delta_i + \triangle \varepsilon_{ist}) \tag{2.3}$$

OLS estimates for equation (2.3) for the 2004-2008 period appear in table 2.3. The first column, reporting estimates of equation (2.3) with fixed effects and state trends, indicates a strong negative association between change in BN share and $IPperVoter$ growth. As argued above, ethnicity is a central driver in Malaysian politics, with non-Malays more likely to switch allegiances from 2004 to 2008. Since the Chinese population is wealthier and more urban, it could be that $IPperVoter$ is simply

picking up this trend. In column (2) I control for this possibility by adding ethnicity, and although the magnitude of the effect diminishes, it remains strongly significant. In line with anecdotal evidence, Indians swung heavily against the BN relative to Malays.[35]

Another concern is that Internet access simply proxies for wealth; the opposition party PKR, for example, derives much of its support from wealthier Malays. In column (3) I add a measure of GDP per capita as of 2005, and again the magnitude drops, but the relationship remains very significant.

Finally, it could be that $IPperVoter$ is capturing the effect of urbanization. As mentioned above, rural districts traditionally support the BN. I control for urbanization of a district with the variables population density and the natural log of eligible voters in 2004. Turning to the results of specification (4), the estimated relationship between change in BN share and growth in IP addresses per eligible voter remains unchanged. Meanwhile, there is no evidence of any relationship between population density and voting trends after controlling for district fixed effects and state trends. To give a sense of magnitudes, specification (4) implies that a one standard deviation increase in $IPperVoter$ growth translates to a 1% swing against the BN.

As a further check, I run (2.3) for the 1995-1999 period, when Internet connectivity grew from zero to 15%.[36] As mentioned in section 2.3.2, the Internet was seen to play a decisive role as early as the December 1999 elections. Significantly, demographic composition of the electoral swing differed in the 1999 election. In 2008, Chinese and Indian voters abandoned the BN in favor of the opposition; whereas, in 1999 minority voters stayed with the BN and the Malay electorate instead split. I create a measure of Internet growth from 1995 to 1999, $InternetHH$, which is the natural log of the percentage of households with an Internet subscription in 2000 (Internet penetration was zero in 1995).

Table 2.4 provides the results. Column (1) shows that, in the absence of controls, a significant positive relationship exists between Internet growth and change in vote share between 1995 and 1999. I interpret this as picking up the fact that the bulk of the swing occurred among Malay voters rather than the relatively more connected

[35]The coefficient on percent Chinese is also negative and significant, when percent Malay is the omitted variable.

[36]I cannot run a regression for the 1995-2008 period because of redistricting between 1999 and 2004 and because my measures of Internet penetration are different.

Chinese. Indeed, adding ethnicity controls in specification (2) renders the relationship strongly negative and significant. In specifications (3) and (4), I control for GDP per capita and population density, and the magnitude of the effect increases.

The results from the 1995-1999 period reinforce the initial finding of a negative relationship between Internet growth and BN share of the vote. That this result holds for a completely different measure of Internet growth suggests that the result is not merely artifact of the $IPperVoter$ measure. Moreover, since the relationship holds in the presence of a different demographic shift in the electorate, there is less reason to believe that unobserved state trends are driving the result. Notably, that the relationship is larger in magnitude: a one standard deviation increase in percentage of households with an Internet subscription implies a 2% swing against the BN in 1999 (the total swing against the BN in 1999 was 11%). This is most likely arises because $InternetHH$ is measured with less error than $IPperVoter$.

2.5.2 Identification Strategy

Although the OLS estimates demonstrate a negative relationship between Internet growth and change in BN share, it remains unclear if the relationship is causal. OLS estimation of equation (2.3) will not identify the causal effects of Internet growth if $\delta_i + \triangle\varepsilon_{ist}$ is correlated with $\triangle IPperVoter_{ist}$. If Internet connectivity is allocated more heavily to districts that are trending toward the BN for unobservable reasons (e.g., patronage) then $\hat{\alpha}_{2,OLS}$ would be biased upward toward zero. If anything, however, this would lead me to underestimate the negative relationship. A greater concern is that Internet connectivity was allocated to areas that trended against the BN for unobserved reasons, leading to a negative bias in my results.

To deal with these challenges, I use the distances from the centroid of a state to Malaysia's three largest ISP backbones as instruments (Z_{ij}) that are correlated with growth in Internet penetration, but uncorrelated with district level characteristics that influence voting behavior. As argued in section 2.3.3, cost, which is a major determinant of Internet placement, increases in the distance to the backbone. Since the backbones were being built in the 1995-1999 period, the instruments apply only to the 2004-2008 elections. This produces the following system of equations:

$$\triangle y_{ist} = \alpha_1 + \alpha_2 \triangle IPperVoter_{ist} + X_{is}\beta_2 + \lambda_s + (\delta_i + \triangle \varepsilon_{ist}) \quad (2.4)$$

$$\triangle IPperVoter_{ist} = \pi_0 + Z_{ij}\pi_1 + X_{is}\pi_2 + \gamma_s + \tau_{ist} \quad (2.5)$$

The identification assumption is that, conditional on the baseline district characteristics —ethnic distribution, GDP per capita, population density—distance to the backbone does not affect change in vote share independently of growth in Internet access. So long as the instruments are also uncorrelated with the bias in $IPperVoter$ toward large cities, they will produce consistent estimates even though $IPperVoter$ is measured with error.

The first endogeneity concern is that the backbones for Malaysia's ISPs run through areas more likely to swing against the BN for reasons that the controls do not capture. Since the backbones pass through Malaysia's most populous regions and cities, the instruments could simply be picking up the direct effect of urbanization on voting trends. I supplement my controls for population density (log of eligible voters, log of total area) with variables based on satellite data. With clutter data on land usage, I create additional controls for the percent of the district that is urban vs. rural vs. jungle. Last, following Burchfield, Overman, Puga, and Turner (2006), I control for the effect of physical topography on urbanization, constructing a variable for the standard deviation of the land gradient.[37]

Another concern with my instrumental variables is that they are picking up the direct effect of Malaysia's major roads and railways on district trends (e.g., via increased trade and exposure to outside information). For now, I include a control for road density but will return to this issue in more detail in section 2.5.5.

In sum, I am exploiting exogenous variation in Internet supply due to geographical constraints in backbone placement. I include state fixed effects due to differences across states in terms of Internet connectivity and sociopolitical factors. Thus I am exploiting within state variation.[38]

[37] All regressions have also been run with average land gradient, ruggedness, and the standard deviation of ruggedness with no significant differences. See appendix for details of variable construction.

[38] There are 11 states in the sample and 38.8 legislative districts per state.

2.5.3 First Stage

Table 2.5 shows the first-stage estimates for Internet penetration growth in state legislative districts, using growth in IP addresses per eligible voter as a proxy for growth in Internet access. Column (1) shows estimates of equation (2.5) with minimal controls for ethnicity. The coefficient on distance to Time, which is highly significant and in the expected direction, suggests that growth in Internet access is decreases with distance from Time's backbone.

For Fiberail, both a linear and square term are included. This is meant to capture a non-linear relationship with $IPperVoter$ growth due to restrictions on Fiberail's geographic area of operation until 2006, as mentioned in section 2.3.3.[39] The negative coefficient on the linear term can be interpreted in the same way as the coefficient for distance to Time: Internet growth decreases as distance increases. The positive square term captures the geographical limitation effect: the relationship between Internet growth and distance to Fiberail becomes less negative until it reaches a zero threshold.

Distance to TM's backbone remains insignificant regardless of the specification. In fact, even if I run the same set of regressions only including IP addresses assigned to TM, the results are largely the same. This result is consistent with the idea that the government exerted more influence over TM than its competitors, compelling it to build out infrastructure in areas with low demand.

In specification (2), I control for GDP per capita. The size of the coefficients for the instruments decreases yet remains highly significant. Column (3) shows that including controls for population size and density does not noticeably alter the result. In specification (4), I add geo-spatial controls for urbanization. The coefficients of interest decrease slightly in magnitude but maintain their significance. Internet growth demonstrates a negative relationship with population density, but a positive association with percentage of the district that is urban. The most likely reason for this result is catch-up: ISPs had already brought Internet service to the most densely populated areas by 2004 and thus had the most room to grow in regions that were urban but more sparsely populated.

Column (5), which includes a control for road density, is my preferred, baseline specification. The coefficients of interest are unaffected, helping to mitigate the worry

[39]The linear term by itself is insignificant. There is no evidence of a non-linear relationship for any of the other instruments

that the instruments are simply picking up the direct effect of roads on elections. To give some interpretation of the magnitudes here, for every 10 kilometer increase in distance to Time's backbone, $IPperVoter$ growth decreases by 0.18 of a standard deviation.

Finally, in specification (6), I control for BN share in 2004. As shown, I find no evidence of political interference on Internet roll-out. Several other results suggest that demand, rather than patronage, was the primary determinant of Internet growth. First, there is a strong positive relationship between GDP and Internet growth regardless of the specification. Second, there is no significant correlation between ethnicity and Internet growth, belying the government's stated goal of ICT investment as a way to promote a Bumiputera middle-class.

2.5.4 Instrumental Variable Results

The IV estimates appear in table 2.6. The specifications for (1)-(5) match their first-stage counterparts from table 2.5. The coefficient on $IPperVoter$ is negative, significant, and of roughly the same magnitude throughout. The Hansen test does not reject the null hypothesis that the instruments are uncorrelated with the error term, lending credence to the identification assumption. The strong and stable coefficients on ethnicity confirm the importance of race in the 2004-2008 elections.

The effect's magnitude drops in column (2), suggesting that $IPperVoter$ in (1) was picking up some of the effect of GDP per capita. The result remains large and significant, however, and in (4) the coefficient of interest returns to its previous size once controls for urbanization are also included.

Column (5), the baseline estimate, includes a control for road density. As shown, the coefficient on $IPperVoter$ stays unchanged. GDP per capita loses its significance altogether, suggesting that it was proxying for urbanization and road density.

To get a sense of the change in magnitudes, for specification (5) a standard deviation increase in Internet growth translates to a 3.6% swing against the BN. Putting this shift into context, IP addresses per voter doubled in the 2004-2008 period, while share of the vote for the BN dropped from 63.9% to 52.2%. This implies that Internet growth accounted for about a third of the vote swing.

The magnitude is substantially larger than the OLS estimate, which as I show in the appendix, is likely due to measurement error biasing the OLS estimates toward

zero.

2.5.5 Validity of the Exclusion Restriction

As a reminder, my identification assumption is that, conditional on baseline district characteristics (ethnic distribution, GDP per capita, population density, road density, percent urban vs. rural vs. jungle), distance to the backbone does not affect change in vote share independently of growth in Internet access. Though it is impossible to test this assumption directly, I perform some additional checks to assess its plausibility.

Pre-Internet Trend Tests The most basic concern is that unobservable characteristics of areas close to the backbone make those areas more prone to swing against the incumbent party in general. I check for this possibility by examining the reduced form relationship between distance to the backbone and swings in previous elections.

Since Malaysia regularly redraws electoral district boundaries, it is not possible to run this exercise for the complete set of preceding elections. Fortunately, the 1969-2008 period has only two other elections—1986-1990 and 1995-1999—in which there was a sizable swing against the BN, and on both occasions boundaries were fixed. I control for ethnicity and population density using the 1991 and 2000 censuses. I also include controls for population density, road density, and land usage based on 2008 estimates. A large expansion in state roads occurred during this time, which introduces error into my road density control. Finally, I control for GDP per capita using a 2005 estimate for 2004-2008, 1996-1999, and 1986-1990. Results do not change if the controls measured with error are dropped.

Table 2.7 shows the results of reduced-form regressions for the 1986-1990, 1995-1999, and 2004-2008 periods. Columns (1) and (2) show a negative and insignificant relationship between vote swing and distance to either backbone in the 1986-1990 period. Turning to column (3), the relationship to distance to Time shifts to positive and significant at the 10% level during the 1995-1999 period. This makes sense since both backbones were partway built during this period. However, as can be seen in column (4), distance to Fiberail remains insignificant. In terms of the 2004-2008 period, column (5) indicates a positive relationship with distance to Time that is

significant at the 5% level, and column (6) shows that the linear and square distance to Fiberail variables are jointly significant at the 1% level.

Controlling for alternate channels A second major concern is that there are unobservable characteristics of areas close to the backbone that only switched on in the 2004-2008 period.

Since Time and TM run along Malaysia's major roads, the greatest cause for concern is that some characteristic particular to the distance to the roads (or an omitted variable driving it) affected voting trends through some channel, which switched on only after the 2004 elections. I believe this possibility is unlikely for several reasons.

First, the backbones travel along Malaysia's federal roads, most of which were built before 1980. Thus, the effects would have had to remain dormant for more than twenty years.

Second, since Time and TM only travel along a subset of federal roads, we can control both for distance to federal roads and distance to major roads. Table 2.8 presents the results of equation 2.5 with additional controls for distance to major roads and distance to federal roads. In specifications (2) and (3), I control for distance to major roads and distance to federal roads using distance to Time, distance to Fiberail, distance to Fiberail squared, and distance to TM as IVs. As illustrated, the coefficient on $IPperVoter$ decreases only slightly and maintains its significance regardless of the control. In columns (8) and (9), I run the same set of regressions but use only my road-based IV, distance to Time. In this case, the magnitude stays largely the same. The standard errors increase substantially, but the relationship remains significant at the 5% level.

Finally, I restrict my set of instruments to those based solely on the railway network. In specifications (4), (5), and (6) we see that the coefficient on $IPperVoter$ remains highly significant regardless of the control. Comfortingly, the magnitude stays largely the same as in the case using only road-based instruments.

Since Fiberail travels the length of Malaysia's railroads, it is impossible to include equivalent controls for distance to the railway. However, it is worth noting that the railroad network was completed as early as 1931. Thus, to invalidate the instrument, the effect of proximity to railroads would have to have remained dormant for 75 years and then activate just in time to influence the 2008 elections.

Other issues Another concern is the possibility of heterogeneous effects of Internet access on voting. If the effect of Internet access on voting is more highly negative for areas closer to the backbone, my identification strategy would lead to an overestimation of the effect. An example of this scenario is if areas closer to the Time backbone are better able to exploit Internet technology through better education. Were that the case, however, we would expect to see a markedly different coefficient on $IPperVoter$ when only distance to Fiberail is used as an instrument. This is because many of the districts near to Time's backbone are far from Fiberail's backbone. However, as table 2.8 shows, the coefficient on $IPperVoter$ is largely the same across specifications regardless of the combination of instruments used.

I have run regressions controlling for change in ethnic distribution, eligible voters, and population density between 2004 and 2008. The results are unchanged suggesting that migration is not driving the effect.

2.5.6 Additional robustness checks

Placebo regressions As an additional check, I test whether Internet growth between 2004 and 2008 is higher in areas that were already predisposed to swing against the BN for unobservable reasons. Running OLS on equation (2.3), I use change in BN share for earlier elections as the dependent variable, but keep $IPperVoter$ 2004-2008 as the independent variable and use the same set of controls. As explained in section 2.5.5, this analysis is possible only for two previous elections, 1986-1990 and 1995-1999, with the same limitations to the controls.

The 1986-1990 period is a good test case of whether places that experienced more Internet growth between 2004 and 2008 were already more predisposed to swing against the BN. The year 1990 saw an abortive move toward a multi-party system in Malaysia with the BN suffering its worst setback since 1969.[40] It won only 53% of the vote, but managed to retain its two-thirds majority in parliament thanks to gerrymandering.

Panel A of table 2.9 shows the results of regressing change in BN share from 1986 to 1990 on Internet growth from 2004 to 2008. As indicated, the coefficient on

[40]Two years earlier, divisions in the UMNO, the dominant Malay party within the BN, caused a formal split in the party with a large number of UMNO politicians leaving to form the opposition Malay party *Semangat 46*.

$IPperVoter$ proves small and insignificant regardless of the specification. This suggests no correlation between support for the BN in the 1986-1990 period and Internet growth in the 2004-2008 period.

Next, I run the equivalent regression for the 1995-1999 period, regressing BN share 1995-1999 on $IPperVoter$ 2004-2008. Recall from section 2.5.1 that a robust negative relationship exists between growth in Internet connectivity (as measured by the 2000 census) and BN share. Panel B shows that this result does not hold if 2004-2008 measures are used instead. No sign of a relationship between the 1995-1999 election swing and 2004-2008 Internet growth appears, regardless of controls. This suggests that the areas with the greatest swing in 1995-1999 differ from areas experiencing the greatest relative growth in Internet access in 2004-2008.

2.6 Additional results

In this section, I consider the effect of Internet diffusion on additional electoral outcomes. I start by checking the secondary prediction of the model: that Internet growth leads to higher turnover once Internet penetration reaches a high enough level. Next, I check if the Internet promoted greater turnout. Last, I predict the outcome of the election if had there been no Internet growth over the 2004-2008 period.

2.6.1 Turnover

A secondary prediction of the model is that higher Internet penetration will yield higher turnover in incumbent party seats once Internet access is sufficiently high. I test this prediction by comparing turnover in seats defended by the BN during the 2008 elections when Internet penetration was greater than 50% to turnover in the 1999 elections when Internet penetration was below 20%. In contrast to previous specifications the analysis is at the cross-sectional level. I run probit regressions of a BN victory dummy on the level of Internet penetration while limiting the sample to districts won by the BN in the previous election. Table 2.10 reports the results.

First, I examine the Internet's effect on turnover in the 1999 election. Since Internet penetration across the country as a whole had reached only 15%, the model would not predict a significant effect on turnover of BN candidates. Turning to the data, I

find no turnover in BN-defended seats in the states of Johor and Negeri Sembilan. Since my empirical strategy exploits within state variation, I drop the 68 observations corresponding to these two states. I also drop 7 observations corresponding to the state of Kelantan, where all BN-defended seats fell to the opposition. Specification (1) reports the result of a probit regression for 1999. The effect, positive and insignificant, provides no evidence that low levels of Internet penetration substantially affect voter turnover.

In 2008 Internet penetration for Malaysia as a whole had surpassed 50%, high enough for the model to imply an increase in turnover. Column (2) affirms this prediction, implying that the BN had less chance of retaining a seat in districts with higher Internet penetration. This specification includes the full set of baseline controls plus distance to federal roads. Logit and linear probability specifications yield commensurate results.

To address endogeneity concerns, in columns (3)-(6), I instrument for $IPperVoter$ 2008 using distance to the backbone. The coefficients on $IPperVoter$ 2008 are relatively stable across specifications (3)-(6), but much larger in magnitude than the simple probit case, pointing again to measurement error biasing the result to zero. In column (3), I include all instruments and the effect proves significant at the 10% level. Turning to specification (4), I drop my weakest instrument, distance to TM, and the significance jumps to the 5% level. In column (5), I restrict the instruments to distance to Time and, in column (6), I use only distance to Fiberail and distance to Fiberail squared. Although the point estimates remain similar, the standard errors are much higher, leading to insignificant results. The most likely explanation for the lower significance is that distance to the backbone variables are weak instruments for the level of Internet access as opposed to change in Internet access. The coefficients on the instruments in the first stage are largely insignificant. Additionally, I run an IV regression for the equivalent linear probability model and include the F-statistics from the first stage. As can be seen, the F-statistics are much smaller than in the case when distance to backbone instruments for Internet growth.

2.6.2 Turnout

Although turnout is not modeled, there are both theoretical and empirical reasons to believe that access to better information on politician quality yields increased

turnout (e.g., Banerjee, Kumar, Pande, and Su (2011)). I look at the effect of Internet diffusion on turnout, focusing on the 2004-2008 and the 1995-1999 periods.

Turnout measures in Malaysia are noisy due to electoral irregularities. Allegations of electoral manipulation range from phantom voters (in which deceased individuals still manage to cast ballots) to multiple votes by the same individual to vote-buying.[41] To address this challenge, I include an extra set of regressions that drop districts with serious irregularities.[42] For the 2004-2008 elections, 13 out of 427 districts are dropped. However, a lack of information on specific examples of irregularities in earlier elections makes it impossible for me to do the same for the 1995-1999 period.

Table 2.11 presents the results. In specification (1), I run equation (2.3) using change in turnout as the dependent variable. The relationship between Internet growth and turnout is positive but significant only at the 10% level. In column (2), I drop 13 districts with indications of serious irregularities. As indicated, the magnitude of the relationship rises and the significance increases to the 5% level.

Next, I employ an IV strategy, but the instruments prove much weaker in this case; only distance to Time yields significant results. In columns (3) and (4), I use distance to Time as an IV and include the standard set of controls, plus distance to federal roads. In both cases, the size of the effect increases greatly. However, the relationship is significant only if I drop districts with voting irregularities. To give a sense of the magnitudes, column (2) implies that a one standard deviation increase in Internet growth leads to a 0.5% increase in turnout. Column (4) implies that a one standard deviation increase in Internet corresponds to a 1.5% increase in turnout, or about half the change in turnout between 2004 and 2008.

Specification (5) shows results for identical OLS regression run for the 1995-1999 elections. Magnitudes are similar to OLS estimates (1) and (2), but standard errors are also much greater, leading to lower significance.

2.6.3 Predicted results in absence of Internet

To put the previous results in perspective, I predict the outcome of the 2008 election had there not been any Internet growth in the 2004-2008 period. Table 2.12

[41] See Pepinsky (2007) and Hai (2002) for details.
[42] See data appendix for details of irregularities.

reports the results. Specifications (1) and (2) give the actual result for the 2004 and 2008 elections, respectively. As shown, the opposition captured four additional statehouses in 2008. Specification (3) employs OLS equation (2.3) to predict results assuming zero growth in $IPperVoter$ from 2004 to 2008. The predicted percent of seats captured by the BN increases in all states, and the BN retains one of the four statehouses lost. Column (4) reports the estimated outcome with no Internet growth using the IV specification. In this case, the effect is more pronounced: the BN retains control of three of the four statehouses that switched to the opposition. These results suggest that, without Internet growth between 2004 and 2008, the BN's 2008 election setback would have proven fairly modest, amounting to the loss of only one statehouse.

2.7 Conclusion

This paper contributes to our understanding of the effect of Internet diffusion on democratization. Focusing on the context in which the traditional media is government-controlled, I have argued that the Internet can facilitate evolution toward a two-party system by preventing any single agent from monopolizing information. Malaysia provides a key opportunity to test this idea: ambitious investment in an Internet free of censorship coincided with strict controls on all other forms of media.

This paper's central contribution is to quantify the effects of the Internet on democratic change, in the context of the huge growth in Internet penetration that has accompanied Malaysia's recent electoral upheavals. I present a model, based on Besley and Prat (2006), in which an increase in Internet access undermines an incumbent party's ability to guarantee reelection through media control. In line with the model's main prediction, I find that Internet growth accounts for one-third of the 11% swing against the BN in the 2008 state elections.

To put this number in perspective, I predict the outcome of the 2008 elections had there not been any Internet growth during the 2004-2008 period. IV estimates imply the BN would have retained control of three of the four statehouses that switched to the opposition. Thus the BN's ICT-based development strategy had the unintended consequence of weakening its control.

I go on to test a secondary prediction of the model. I show that Internet growth

can yield increased turnover if Internet access is sufficiently high. Finally, I find evidence that the Internet can help spur higher turnout.

Another important contribution is a novel measure of Internet growth from 2004 to the present. Such a metric is lacking for most countries in the world, including the U.S.A. My measure of Internet connectivity uses IP geo-location data in conjunction with regional Internet registry records. I smooth the IP address point data into a surface using inverse distance weighting interpolation and then normalize by population. Finally I check the accuracy against an independent measure of Internet diffusion from household census data. This measure is central to the paper's results as it allows me to track Internet growth at the state legislature district level. This measure can also extend to research well outside the ambit of this paper. Equivalent IP geo-location data exists for almost every country in the world and is only becoming more accurate as the technology matures.

This paper presents some of the first evidence of the Internet's quantitative effects on political outcomes. However, there is much scope for future work. First, it is important to get a better understanding of the channels of causation. The model suggests that the Internet influenced elections via the media market. In line with the model's predictions, anecdotal evidence suggests a drop in the popularity of BN owned newspapers and even a decrease in bias among some media outlets as a means to reestablish credibility. Finally, it would be fruitful to explore the Internet's consequences in terms of economic development. Malaysia invested heavily in Internet infrastructure to promote an information economy. An important next step would be to gauge whether this investment paid off, as it would imply a relationship between political openness and economic growth.

2.8 Tables

TABLE 2.1
Summary statistics

Variable	Mean	Std. Dev.	N
Dependent variables			
$\Delta BNShare$ 2004-2008	-.1211	.0933	439
$\Delta Turnout$ 2004-2008	.0189	.0360	439
Independent variables			
% Internet 2000	.1657	.1619	439
% Malay 2004	.6339	.2752	439
% Indian 2004	.07676	.0774	439
% Internet 2004	.1677	.1745	439
GDP per capita 2005	16668.72	7141.14	439
Eligible voters 2004	17716.52	7158.29	439
Population density	790.73	1404.24	439
% Urban	.2148	.2390	439
% Rural	.5022	.2501	439
Slope std. dev.	4.030	2.953	439
Road density	.6153	.6272	439
Km to federal road	3.575	4.727	439
Km to major road	1.381	2.134	439
Instrumental variables			
Km to Time	15.349	18.468	439
Km to Fiberail	22.400	28.184	439
Km to Fiberail sq	1294.351	3080.338	439
Km to TM	7.129	7.787	439

Notes. The table reports summary statistics for state legislature districts in peninsular Malaysia, excluding Kuala Lumpur. Variables measured in 2008 unless otherwise stated. See appendix for details on the construction and sources of variables.

TABLE 2.2
Evaluation of Internet penetration measures

	% households with Internet 2004				
	(1)	(2)	(3)	(4)	(5)
IPSumPerVoter 2005	0.360	0.580	0.265	0.621	0.628
IPMaxPerVoter 2005	0.463	0.511	0.373	0.533	0.539
IPAvgPerVoter 2005	0.317	0.459	0.169	0.492	0.515
IPFixPerVoter 2005	-0.055	-0.105	-0.584	-0.135	-0.133
IPSumPerVoter 2004	0.053	0.016	0.085	0.006	0.017
IPMaxPerVoter 2004	0.068	0.033	0.159	0.022	0.034
IPAvgPerVoter 2004	0.058	0.020	0.078	0.010	0.021
IPFixPerVoter 2004	-0.080	-0.128	-0.144	-0.153	-0.152
Kuala Lumpur	Y	N	N	N	N
Sabah	Y	Y	Y	N	N
Peninsular Malaysia	Y	Y	N	Y	Y
N	518	505	60	445	433

Notes. Correlation between percentage households with Internet subscription 2004 and self-constructed Internet penetration measures. Percentage Households with Internet access in 2004 was derived from Household Basic Amenities Survey 2004. See section 2.10 for details on the construction and source of variables.

TABLE 2.3
Relationship between BN share and Internet growth from 2004 to 2008

	Dependent variable is $\Delta BNShare$ 2004-2008			
	(1)	(2)	(3)	(4)
IPperVoter growth	-0.019***	-0.013***	-0.009***	-0.009***
	(0.004)	(0.003)	(0.003)	(0.003)
% Malay 2004		0.164***	0.133***	0.130***
		(0.016)	(0.018)	(0.019)
% Indian 2004		-0.348***	-0.372***	-0.381***
		(0.053)	(0.052)	(0.054)
GDP per capita			-0.036***	-0.034***
			(0.009)	(0.010)
Log eligible voters 2004				0.003
				(0.011)
Population density				-0.002
				(0.003)
N	427	427	427	427
R^2	.441	.695	.71	.711

Notes. The table reports OLS estimates of equation (2.3). All specifications include 11 state trends. IPperVoter growth is natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.4
Relationship between BN share and Internet growth from 1995 to 1999

	Dependent variable is $\Delta BNShare$ 1995-1999			
	(1)	(2)	(3)	(4)
InternetHH 1995-1999	0.021***	-0.015***	-0.018***	-0.020***
	(0.005)	(0.004)	(0.005)	(0.005)
% Malay 1999		-0.299***	-0.294***	-0.285***
		(0.022)	(0.023)	(0.023)
% Indian 1999		-0.212***	-0.206***	-0.198***
		(0.063)	(0.064)	(0.064)
GDP per capita 2005			0.012	0.008
			(0.011)	(0.011)
Log eligible voters 1995				0.012
				(0.017)
Population density 2008				0.001
				(0.001)
N	374	374	374	374
R^2	.269	.576	.577	.579

Notes. The table reports OLS estimates of equation (2.3). All specifications include 11 state trends. 1999 election is from December 1999. Internet growth is the natural log percentage of households with Internet subscriptions in 2000 (Internet access was zero in 1995). See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.5

First stage relationship between distance to backbone and Internet growth

	Growth in IPs per eligible voter 2004-2008: $\triangle IPperVoter$					
	(1)	(2)	(3)	(4)	(5)	(6)
Km to Time*10	-0.115***	-0.092***	-0.094***	-0.082***	-0.082***	-0.082***
	(0.027)	(0.027)	(0.028)	(0.029)	(0.029)	(0.029)
Km to Fiberail*10	-0.194***	-0.177***	-0.177***	-0.158***	-0.156***	-0.157***
	(0.043)	(0.043)	(0.044)	(0.045)	(0.045)	(0.045)
Km to Fiberail*10 SQ	0.021***	0.019***	0.019***	0.018***	0.018***	0.018***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Km to TM*10	-0.015	-0.014	-0.024	0.003	-0.002	-0.002
	(0.059)	(0.058)	(0.058)	(0.064)	(0.064)	(0.064)
% Malay 2004	-0.001	0.295	0.173	0.297	0.294	0.303
	(0.196)	(0.225)	(0.238)	(0.240)	(0.239)	(0.259)
% Indian 2004	0.134	0.405	0.093	0.388	0.457	0.483
	(0.648)	(0.661)	(0.665)	(0.682)	(0.684)	(0.725)
GDP per capita		0.384***	0.465***	0.390***	0.402***	0.404***
		(0.116)	(0.122)	(0.124)	(0.124)	(0.127)
Log eligible voters 2004			0.001	-0.035	-0.035	-0.039
			(0.157)	(0.155)	(0.155)	(0.162)
Population density			-0.077**	-0.148***	-0.133***	-0.133***
			(0.036)	(0.051)	(0.051)	(0.051)
% Urban				0.770**	0.965***	0.962***
				(0.335)	(0.352)	(0.351)
% Rural				-0.033	-0.018	-0.019
				(0.243)	(0.243)	(0.243)
Slope std				-0.001	-0.001	-0.001
				(0.019)	(0.019)	(0.019)
Road density					-0.131	-0.129
					(0.086)	(0.086)
BN share 2004						-0.056
						(0.495)
N	427	427	427	427	427	427
R^2	.307	.325	.333	.349	.351	.351

Notes. The table presents OLS estimates of equation (2.5). It presents first stage results for the relationship between distance to backbone and growth in IP addresses per voter. All specifications include 11 state trends. All specifications include state trends. IPperVoter is natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels. See appendix for details on the construction and source variables.

TABLE 2.6

IV estimates of the relationship between BN share and Internet growth

	Dependent variable is $\Delta BNShare$ 2004-2008					
			IV			OLS
	(1)	(2)	(3)	(4)	(5)	(6)
IPperVoter Growth	-0.036***	-0.028***	-0.029***	-0.037***	-0.036***	-0.009***
	(0.007)	(0.008)	(0.009)	(0.011)	(0.010)	(0.003)
% Malay 2004	0.156***	0.137***	0.133***	0.129***	0.128***	0.120***
	(0.017)	(0.018)	(0.019)	(0.020)	(0.020)	(0.019)
% Indian 2004	-0.331***	-0.350***	-0.362***	-0.374***	-0.371***	-0.396***
	(0.055)	(0.053)	(0.055)	(0.055)	(0.054)	(0.053)
GDP per capita		-0.025**	-0.022**	-0.018*	-0.017	-0.031***
		(0.010)	(0.011)	(0.011)	(0.011)	(0.010)
Log eligible voters 2004			0.009	0.007	0.007	0.003
			(0.012)	(0.012)	(0.012)	(0.011)
Population density			-0.004	-0.006*	-0.005	-0.000
			(0.003)	(0.003)	(0.003)	(0.003)
% Urban				-0.008	0.004	-0.028
				(0.025)	(0.029)	(0.026)
% Rural				-0.009	-0.009	-0.012
				(0.016)	(0.016)	(0.016)
Slope std				-0.003**	-0.003**	-0.003***
				(0.001)	(0.001)	(0.001)
Road density					-0.008	-0.006
					(0.010)	(0.010)
N	427	427	427	427	427	427
R^2	.657	.686	.685	.67	.673	.717
F-Stat	18.1	13.1	12.9	8.5	8.7	
Hansen Test (p-value)	.75	.76	.73	.80	.82	

Notes. Specifications (1) through (5) show results of IV regressions of change in BN vote share 2004-2008 on IPperVoter growth 2004-2008. Instruments are distance to Time, distance to Fiberail, distance to Fiberail squared, and distance to TM. Column (6) reports results from an ordinary least squares regression of BN vote share 2004-2008 on IPperVoter growth 2004-2008. F-stat is the f-statistic of the instruments from the first stage. The *p*-value for the Hansen test is for the Sargan-Hansen test of overidentifying restrictions. The joint null is that the instruments are uncorrelated with the error. All specifications include 11 state trends. IPperVoter growth is the natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.7

Reduced form estimates of distance to backbone on elections

	Dependent variable is $\Delta BNShare$					
	1986-1990		1995-1999			2004-2008
	(1)	(2)	(3)	(4)	(5)	(6)
Km to Time*100	-0.036		0.038*		0.041**	
	(0.026)		(0.023)		(0.016)	
Km to Fiberail*100		-0.055		0.040		0.035
		(0.047)		(0.038)		(0.030)
Km to Fiberail*100 SQ		0.033		-0.034		-0.057**
		(0.035)		(0.031)		(0.023)
Road density	0.257**	0.269**	0.075	0.061	-0.004	-0.003
	(0.120)	(0.120)	(0.104)	(0.105)	(0.010)	(0.010)
Fiberail joint significance		.72		.55		.001
N	325	325	368	368	427	427
R^2	.507	.507	.567	.565	.715	.716

Notes. Reduced form regressions of change in BN share on distance to the backbone are reported. Columns (1) and (2) cover the 1986-1990 elections; columns (3) and (4) cover the 1995-1999 elections; and columns (5) and (6) cover the 2004-2008 elections. Fiberail joint significance presents the p-value of a test of the joint significance of Km to Fiberail and Km to Fiberail squared. All specifications control for ethnicity, GDP per capita, percent of the district that is urban and rural, the log of eligible voters, population density and 11 state trends. GDP per capita is taken from a 2005 estimate in all cases. For all specifications population density, road density, % urban, and % rural are calculated from a 2008 measure. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.8

IV estimates controlling for distance to roads

	Dependent variable is $\Delta BNShare$ 2004-2008								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
IPperVoter growth	-0.036***	-0.033***	-0.032***	-0.035***	-0.033***	-0.032***	-0.040**	-0.035**	-0.035**
	(0.010)	(0.010)	(0.010)	(0.012)	(0.011)	(0.011)	(0.017)	(0.017)	(0.017)
Km to major road*10		0.017*			0.017*			0.017*	
		(0.009)			(0.009)			(0.010)	
Km to federal road*10			0.010*			0.011*			0.010*
			(0.005)			(0.005)			(0.006)
N	427	427	427	427	427	427	427	427	427
R^2	.673	.684	.686	.677	.684	.686	.659	.678	.679
F-stat	8.7	9.2	9	14.5	14.4	14.3	12.5	13	12.8
Instrumental variables	ALL	ALL	ALL	Fiberail	Fiberail	Fiberail	Time	Time	Time

Notes. Specifications (1) through (9) show results of IV regressions of change in BN vote share 2004-2008 on IPperVoter growth 2004-2008. Instruments in (1)-(3) are distance to Time, distance to Fiberail and Fiberail squared, and distance to TM. Instruments in (4)-(6) are distance to Fiberail and distance to Fiberail squared. The instrument in (7)-(9) is distance to Time. F-stat is the f-statistic of the instruments from the first stage. All specifications control for ethnicity, GDP per capita, percent of the district that is urban and rural, the log of eligible voters, population density and 11 state trends. IPperVoter growth is natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.9

Placebo regressions of $\Delta BNShare$ on Internet in a different time period

	(1)	(2)	(3)	(4)	(5)
PANEL A: Dependent variable is $\Delta BNShare$ 1986-1990					
IPperVoter 2004-2008	-0.002	-0.002	-0.002	-0.004	-0.004
	(0.005)	(0.005)	(0.005)	(0.005)	(0.005)
PANEL B: Dependent variable is $\Delta BNShare$ 1995-1999					
IPperVoter 2004-2008	-0.001	-0.005	-0.005	-0.005	-0.005
	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)
Controls					
Ethnicity	N	Y	Y	Y	Y
GDP per capita	N	N	Y	Y	Y
Population	N	N	N	Y	Y
Road	N	N	N	N	Y

Notes. The table reports OLS estimates of equation BN share change on Internet growth in a different period. Panel A reports results of BN share 1986-1990 on IPperVoter Growth from 2004 to 2008. Panel B reports results of BN share 1995-1999 on IPperVoter Growth from 2004 to 2008. IPperVoter 2004-2008 is the natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. All specifications include 11 state trends. Ethnicity controls are % Malay and % Chinese, from each respective period. GDP per capita is taken from 2005 for panels A and B. Population controls for population density, road density, % urban, % rural, and log of eligible voters. Log of eligible voters is for 1986, 1995, and 2004, respectively, while the other controls are from 2008. Road controls for road density and distance to federal roads as of 2008. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels. For expository clarity, coefficients on controls are not reported.

TABLE 2.10
Probit estimates of turnover on Internet

	Probit		IV Probit			
	1999	2008	2008	2008	2008	2008
	(1)	(2)	(3)	(4)	(5)	(6)
IPperVoter 2008		-0.286**	-0.946*	-1.041**	-1.109	-1.011
		(0.140)	(0.557)	(0.513)	(0.702)	(0.617)
InternetHH 1999	0.300					
	(0.222)					
N	255	383	383	383	383	383
Pseudo R^2	.436	.51				
First stage: Dependent variable is $IPperVoter$ 2008						
Km to Time*10			-0.044	-0.040	-0.050*	
			(0.027)	(0.027)	(0.028)	
Km to Fiberail*10			-0.007	-0.001		-0.005
			(0.044)	(0.042)		(0.043)
Km to Fiberail*10 SQ			0.005	0.004		0.005
			(0.004)	(0.004)		(0.004)
Km to TM*10			0.070			
			(0.059)			
F-stat			2.7	3.28	3.47	3.5
N			383	383	383	383

Notes. Probit estimates of turnover on Internet connectivity are reported. Specification (1) regresses turnover from December 1999 on log % households with Internet subscription in 1999, and restricts sample to districts won by the BN in 1995. Specifications (2)-(6) regress turnover 2008 on log IPperVoter 2008, and restrict sample to districts that the BN won in 2004. All specifications control for ethnicity, GDP per capita, percent of the district that is urban and rural, the log of eligible voters, population density, road density, distance to federal roads, and 11 state trends. GDP per capita is taken from a 2005 estimate. For all specifications distance to federal roads, road density, % urban, and % rural are calculated from a 2008 measure. F-stat is the f-statistic of the instruments from the first stage of 2SLS estimate from the equivalent linear probability model. See appendix for details on the construction and sources of variables. Coefficients are reported with standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.11

Relationship between turnout and Internet growth

	$\Delta Turnout$ 2004-2008				$\Delta Turnout$ 1995-1999
	OLS	OLS	IV	IV	OLS
	(1)	(2)	(3)	(4)	(5)
IPperVoter growth 04-08	0.0035*	0.0042**	0.0138	0.0157*	
	(0.0019)	(0.0019)	(0.0087)	(0.0085)	
$InternetHH$ 1995-1999					0.0034
					(0.0026)
Drop irregularities	N	Y	N	Y	N
Time IV	N	N	Y	Y	N
N	427	413	427	413	368
R^2	.649	.641	.604	.586	.301

Notes. Specifications (1) and (2) show results of OLS regressions of change in turnout 2004-2008 on IPperVoter growth 2004-2008. Columns (3) and (4) present results of IV regressions using distance to Time as an instrument. Specification (5) reports results of regressions of change in turnout 1995-1999 on Internet subscription per household growth 1995-1999. Drop irregularities drops districts with irregularities in turnout; see appendix for details. InternetHH 1995-1999 is the natural log percentage of households with Internet subscriptions in 2000 (Internet access was zero in 1995). All specifications control for ethnicity, GDP per capita, percent of the district that is urban and rural, the log of eligible voters, population density, road density and 11 state trends. For specifications (5) GDP per capita is taken from 2005. Road density, % urban, % rural are from 2008. IPperVoter growth is natural log of IP addresses per eligible voter in 2008 divided by IP addresses per voter in 2004. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 2.12

Results of state legislature elections without Internet

State	$BNSeats2004$	$BNSeats2008$	$\widehat{BNSeats2008}_{OLS}$	$\widehat{BNSeats2008}_{IV}$
Johor	.982	.892	.946	.946
Kedah	**.861**	**.388**	**.361**	**.667**
Kelantan	.466	.133	.311	.489
Melaka	.928	.821	.857	.857
Negeri Sembilan	.944	.583	.75	.75
Pahang	.976	.880	.928	.952
Perak	**.881**	**.474**	**.576**	**.644**
Perlis	.933	.933	1	1
Pulau Pinang	**.95**	**.275**	**.35**	**.45**
Selangor	**.964**	**.357**	**.392**	**.554**
Terengganu	.875	.75	.812	.937
N	445	445	445	445

Notes. Table reports fraction of state legislature seats won by the BN alongside estimates in the absence of Internet. Covers all state peninsular seats.$BNSeats2004$ and $BNSeats2008$ are the fraction of state legislature seats won by the BN in 2004 and 2008 respectively. $\widehat{BNSeats2008}_{OLS}$ is the predicted fraction of seats won by the BN in the absence of Internet growth based on OLS equation (2.3). $\widehat{BNSeats2008}_{IV}$ is the predicted fraction of seats won by the BN in the absence of Internet growth based on the IV system of equations (2.4) and (2.5).

2.9 Figures

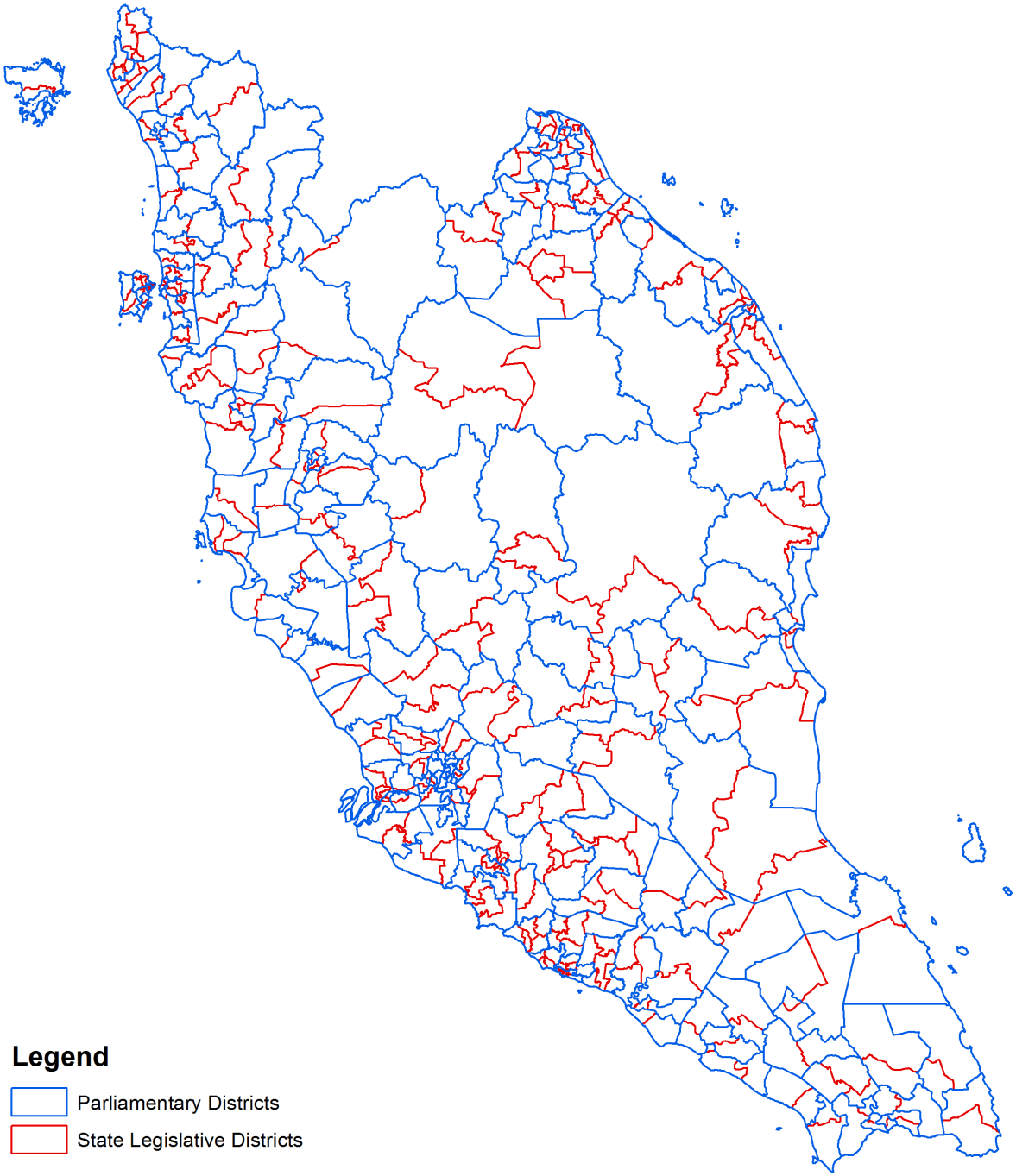

Figure 2.1: Political Boundaries

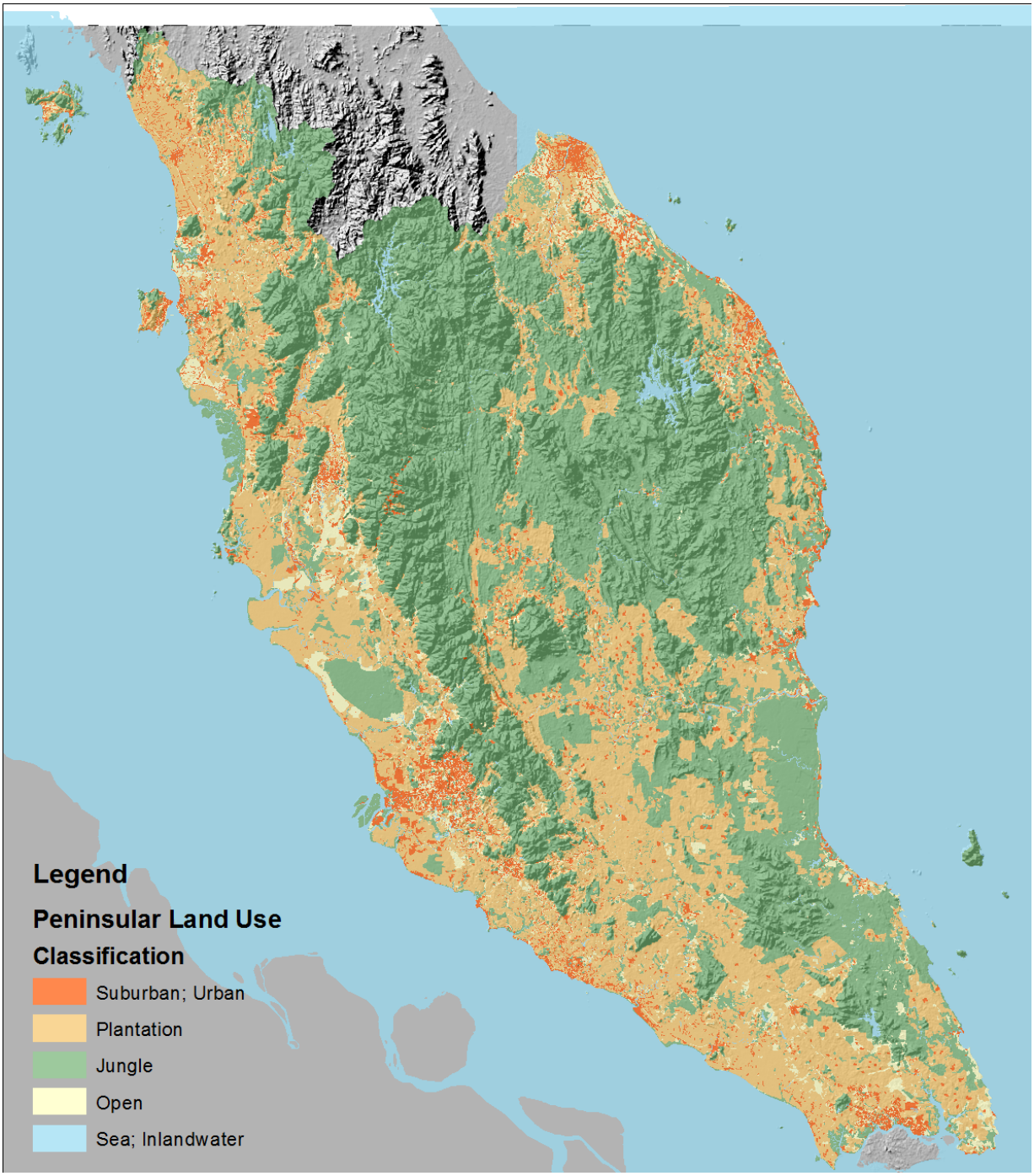

Figure 2.2: Peninsular Malaysia Land Use and Elevation

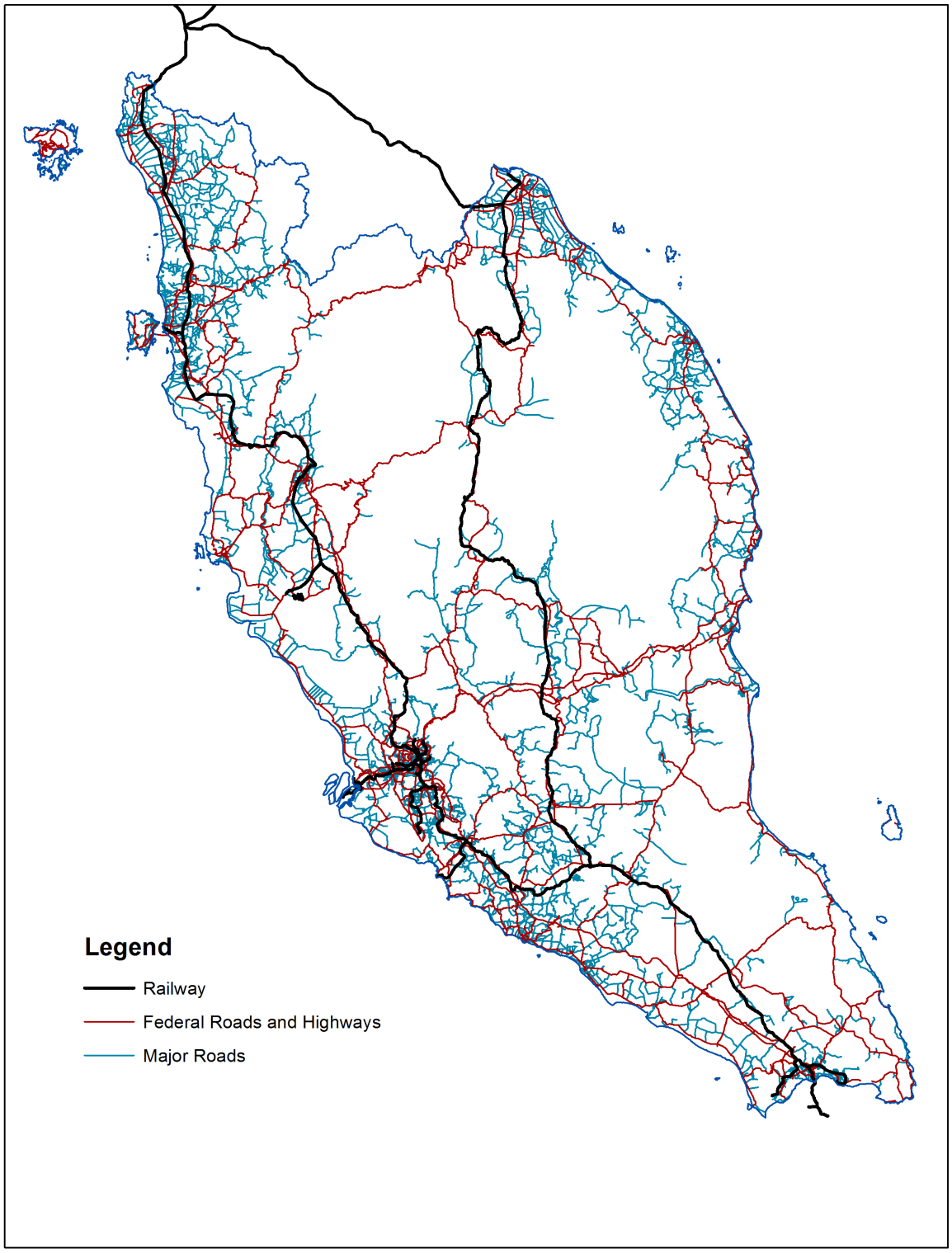

Figure 2.3: Peninsular Malaysia's Road and Railway Network

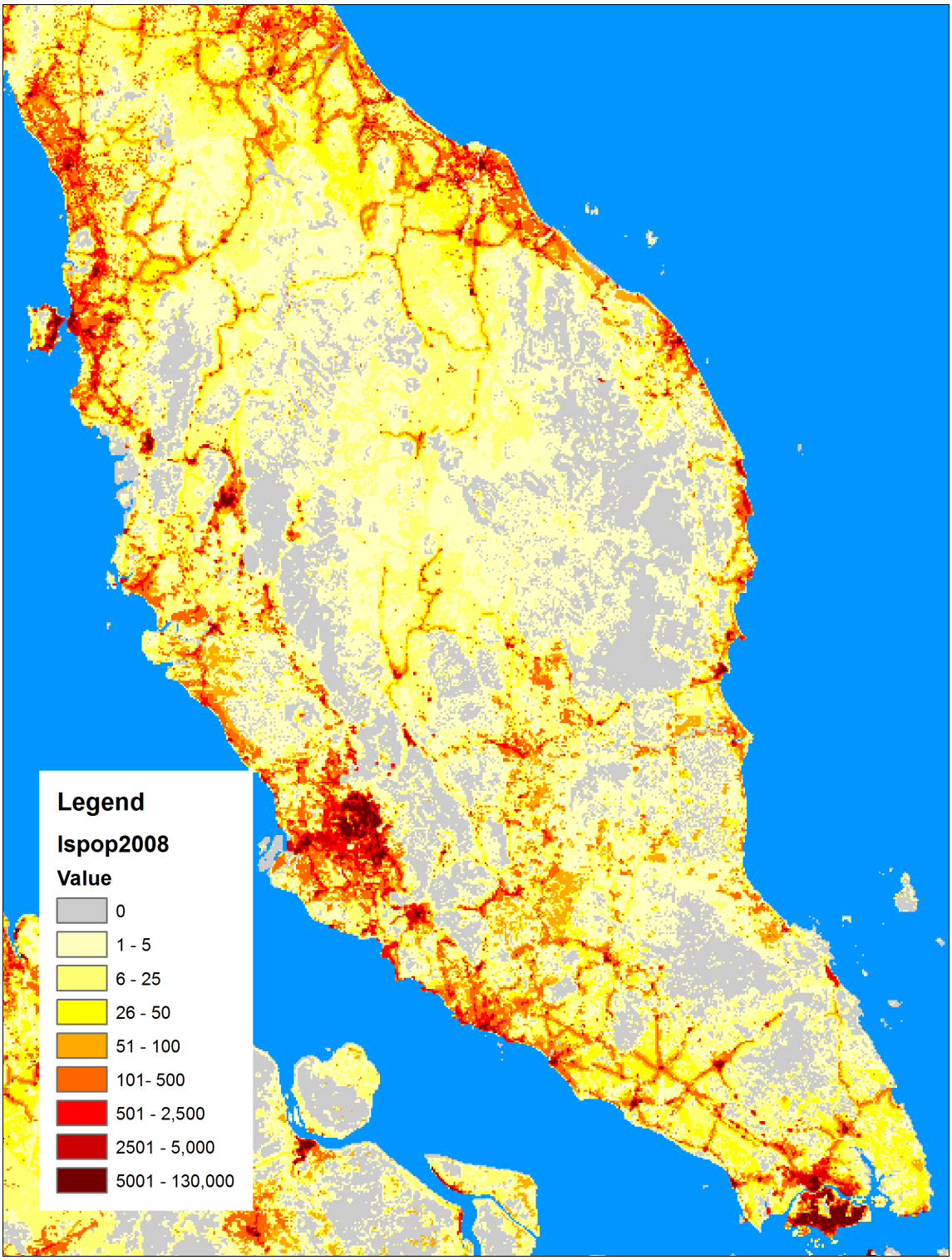

Figure 2.4: Population Distribution

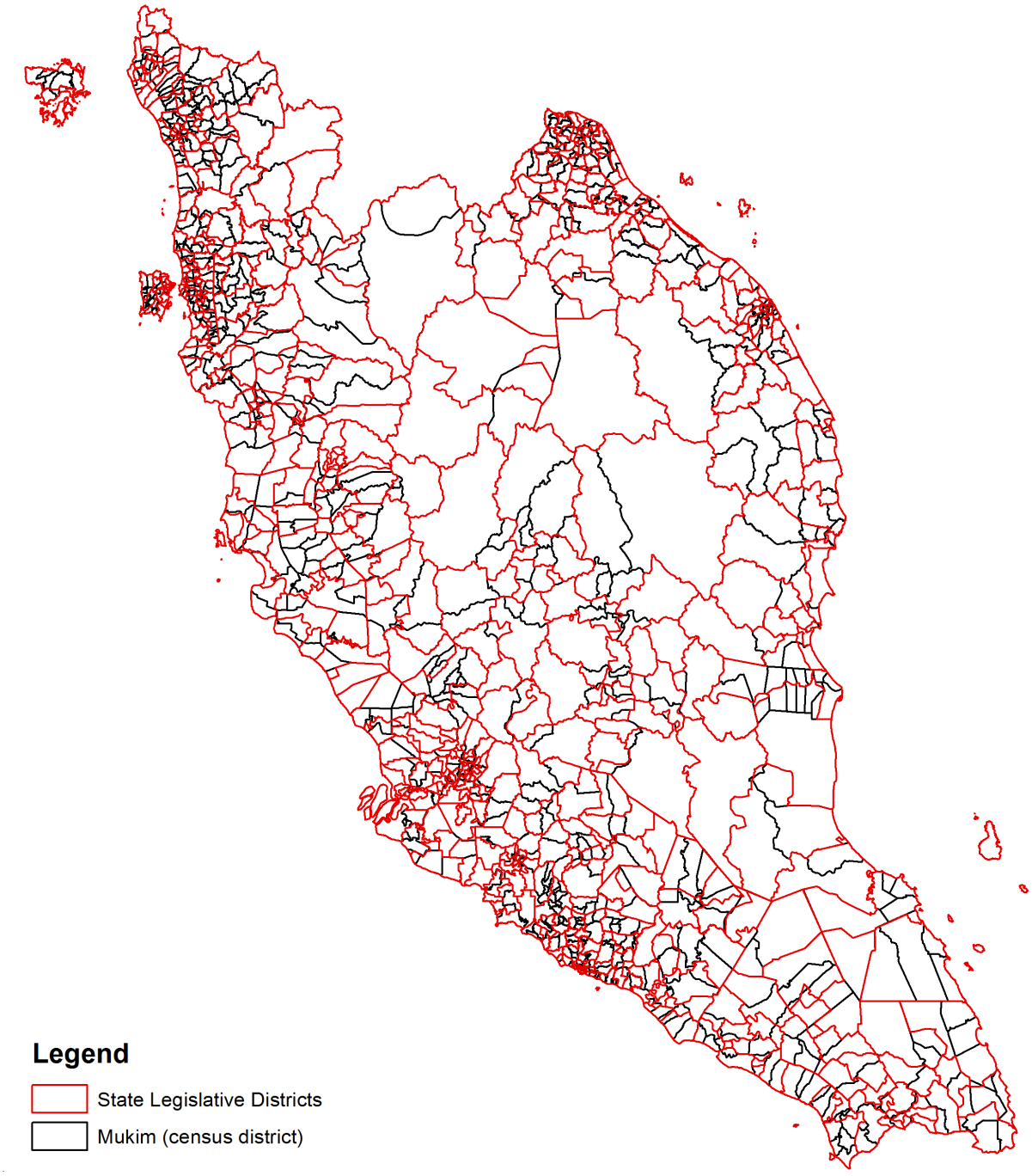

Figure 2.5: Legislative District vs. Census District Boundaries

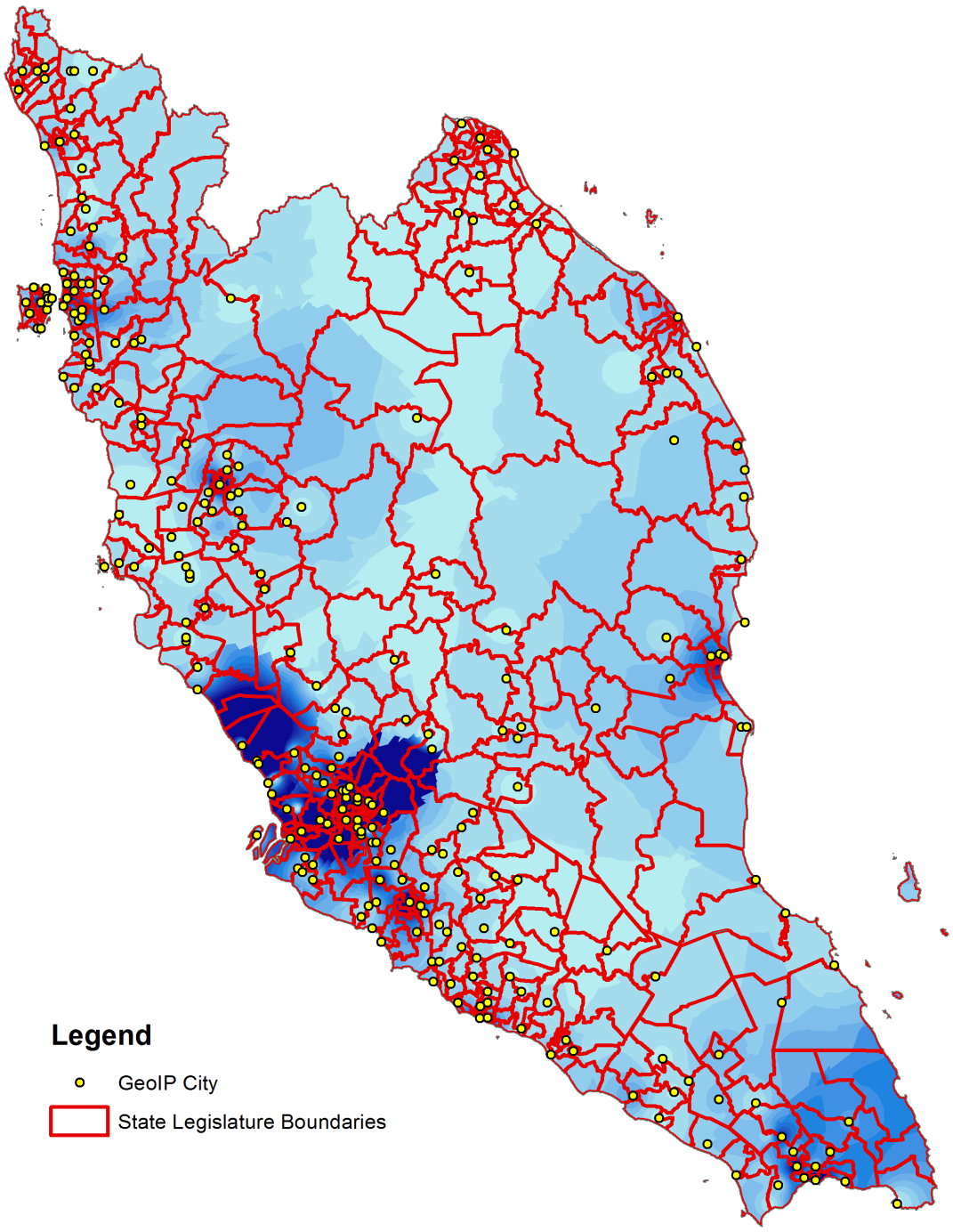

Figure 2.6: Inverse Distance Weighting Interpolation

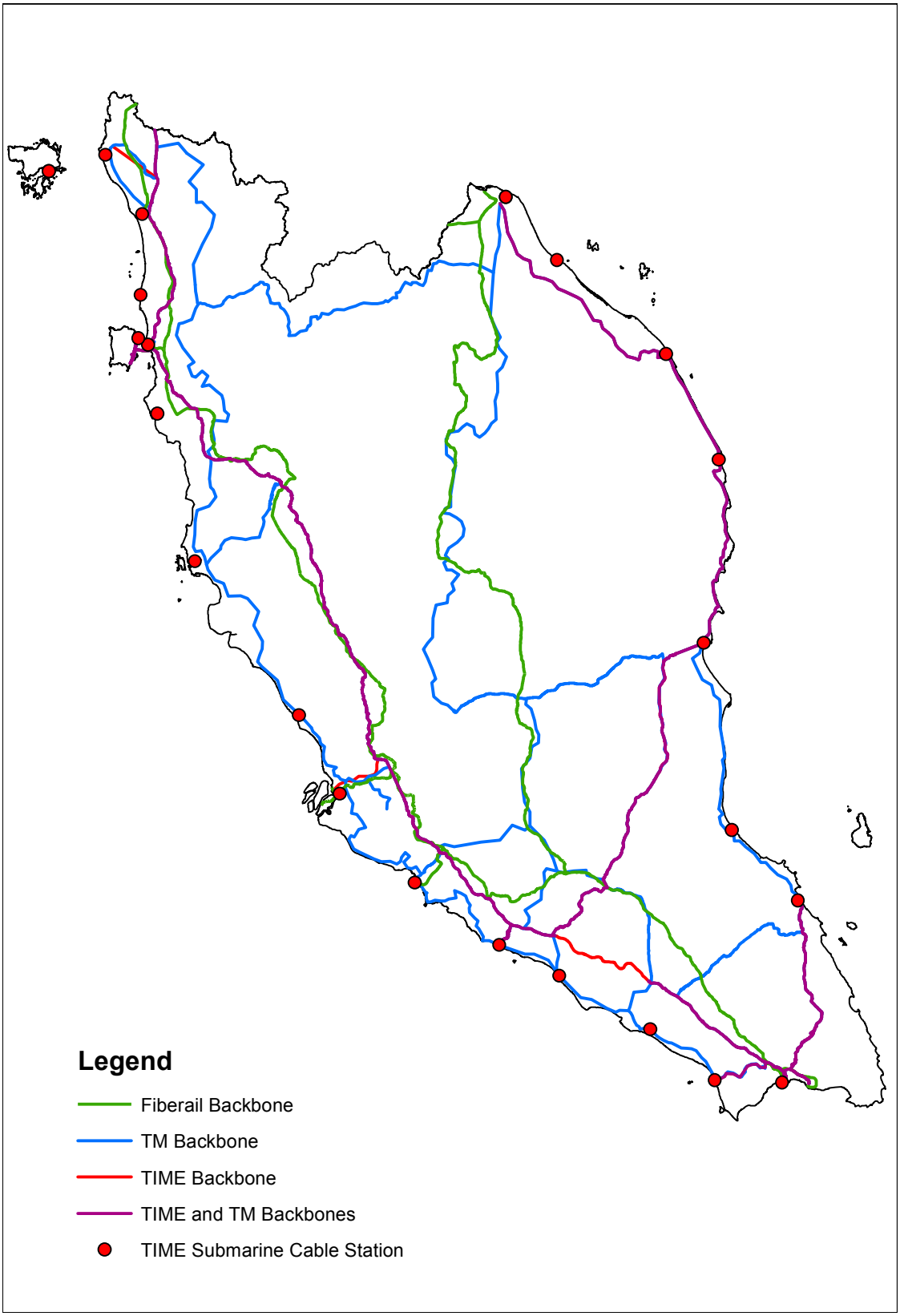

Figure 2.7: Backbone Location

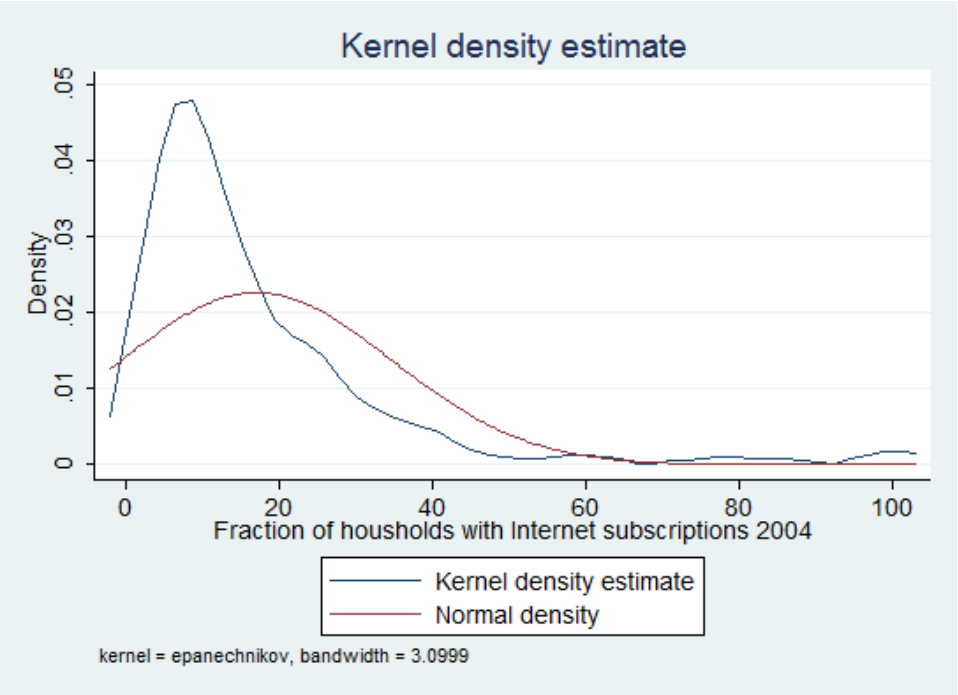

Figure 2.8: Skewness of Households with Internet 2004

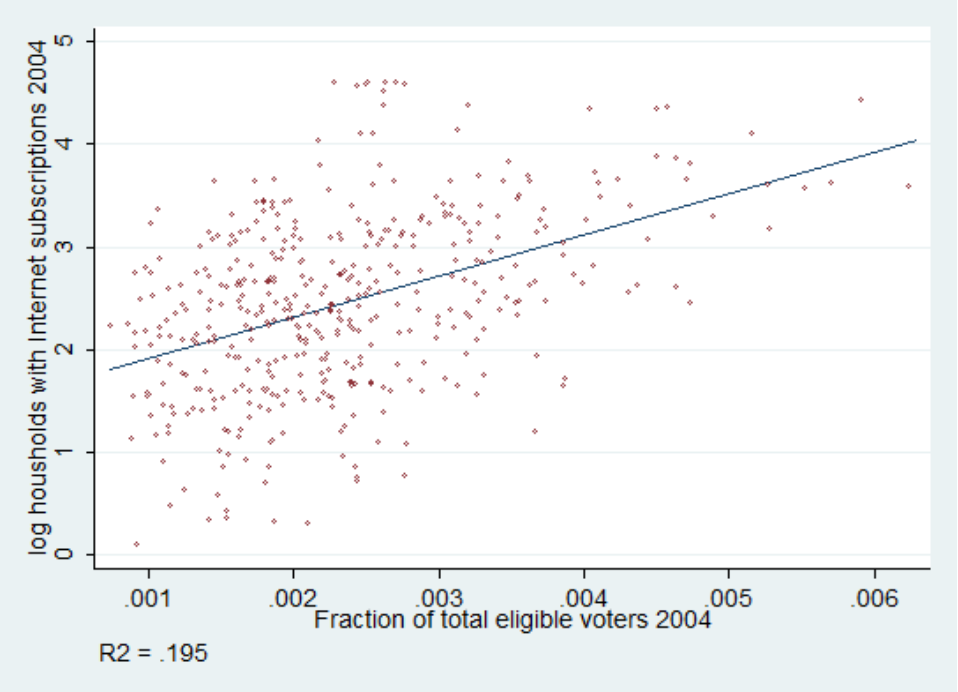

Figure 2.9: Relationship Between Log Households with Internet 2004 and Fraction of Total Eligible Voters

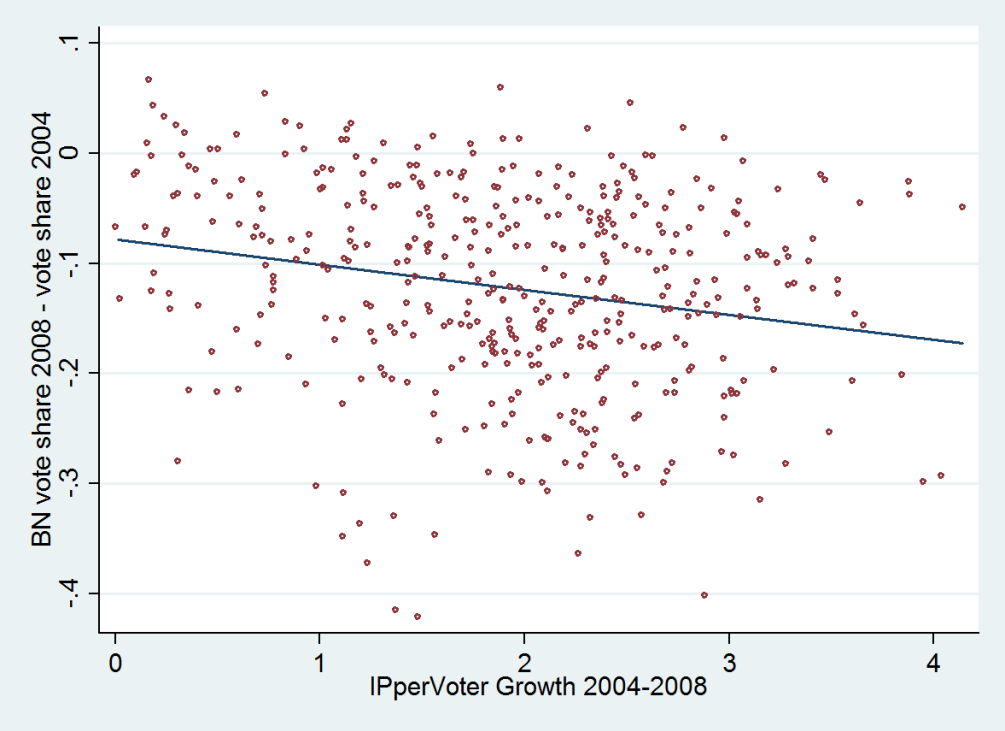

Figure 2.10: Relationship Between Change in BN Vote Share and IP per Voter Growth

Appendix

2.10 Appendix A: Constructing a measure of Internet penetration

Since official statistics do not cover the 2004-2008 period, I construct a novel measure of Internet penetration, $IPperVoter$, at the state legislature district level.

I use the GeoIP City database and APNIC dataset, outlined in section 2.4.3, which together allow me to identify: the initial date of assignment to Malaysia, the ISP managing the IP addresses, and the IP blocks location(s) during the 2004-2008 period. $IPperVoter$ is created by aggregating this data up to the electoral district level and then normalizing by the number of eligible voters.

Challenges In creating this measure I had to address several sources of measurement error:

1. *Change in IP block location over time:* According to Maxmind, in any given month roughly 60% of IP addresses are correctly resolved to a city that lies within 25 square miles of the actual location.[43] To give a sense of what this means, an IP address is not like a static telephone number: although it can remain stable for long periods, it can also change locations without warning. In many cases the IP address is simply reassigned to another computer in the same general area. However, in some cases it may be reassigned to a completely different part of the country. In the context of Malaysia, this would most likely be a problem for smaller cities. Before a city passes a certain level of connectivity, its IP addresses may be routed either to a regional hub or directly to Kuala Lumpur. As a result, this measure will be biased toward larger cities, especially Kuala Lumpur. Indeed, the monthly data is very noisy with smaller cities disappearing and reappearing from month to month.

2. *Change in IP geo-location accuracy over time:* Not only are the locations assigned to an IP address changing over time, but the accuracy of these

[43]This number is periodically updated and can be found at the following address: <http://www.maxmind.com/app/city_accuracy>.

IP/location pairings are increasing as well. The data for 2004 is particularly unreliable.

3. *Geographical Measurement error:* The dataset only provides the coordinates for the city center; it does not specify the city's limits. Thus, an IP address corresponding to a computer in a small town outside a city (and in a different electoral district) may be incorrectly attributed to the city, introducing further bias toward large cities. Furthermore, since only point data is available, the boundaries between cities are unclear. This can most easily be seen in figure 2.6, which presents GeoIP city center data alongside legislative district boundaries. As can be seen, the point locations often appear on the border between two districts, complicating the task of divvying up IP addresses between adjacent districts.

Approach To address these challenges I construct multiple measures of Internet connectivity, each with different strengths and weakness, which I will test against data from the 2004 HBAIS in the next section.

In response to the problem of IP locations changing over time, I test four methods for assigning IP addresses to cities based on data from a year long period.

1. $IPFix$: I limit the sample to IP addresses that never change location over the twelve month period and divide by twelve. This is the most conservative measure, yielding the advantage that the IP addresses are almost certainly assigned to the correct location. The disadvantage is that the majority of the sample is lost with most remaining IPs assigned to Kuala Lumpur.
2. $IPSum$: I sum up the number of IP addresses assigned to each city over a twelve-month period and divide by twelve. This is the second most conservative measure, making equal use of all of the information in the dataset. However, it likely leads to over-counting Kuala Lumpur. For example, there are many cases in which ten or eleven out of twelve observations occur in a single city, with the remainder going to Kuala Lumpur.
3. $IPMax$: I calculate the number of IP addresses assigned to each city for each month. $IPMax$ is the value from the month with the most IP addresses. This measure is meant to account for under-counting of smaller cities. The assumption is that, once an area obtains Internet access, it is unlikely to subsequently have access physically dismantled. If a location does not appear in subsequent

months, this is due to measurement error rather than a subsequent loss of Internet connectivity.

4. *IPAvg:* I, again, calculate the number of IP addresses assigned to each city for each month. I then take the average across months when the city appears in the data. This approach is similar to $IPMax$, but aims to correct for possible over-counting of small cities by $IPMax$.

To address the problem of noisy data for 2004, I create two sets of measures for 2004. The first relies on GeoIP data for 2004. The second uses data from 2005 to infer connectivity at the time of the March 2004 elections. I drop all IP addresses from the 2005 data that were assigned after March 2004 and then calculate the four measures listed above. The assumption is that IP blocks assigned before the March 2004 elections are in roughly the same location in 2005.

To diminish geographical measurement error, I smooth the city point-data into a surface, using inverse distance weighting (IDW) interpolation in ArcGIS. IDW interpolation assigns an IP measure to every point in Malaysia: the value at each interpolated point is a weighted sum of the values in the N known points, where closer points get higher weighting. Figure 2.6 shows an example of IDW interpolation: darker areas have higher numbers of IP addresses.[44] I then calculate the average IP measure for the entire district. Finally, I normalize by the number of eligible voters in the district to generate two sets of measures, one based on 2004 data and the other based on 2005 data: $IPSumPerVoter$, $IPMaxPerVoter$, $IPAvgPerVoter$, and $IPFixPerVoter$.

Choosing the best measure Table 2.2 shows the correlation between the number of Internet subscriptions per household (per the HBAIS 2004) and the IP per voter measures outlined above. Specification (1) includes all state legislative districts for peninsular Malaysia and the Borneo state of Sabah. Kuala Lumpur cannot vote in state legislative elections because it was ejected from the state of Selangor in 1974 and made into a Federal Territory. For the sake of completeness, I include Kuala Lumpur's parliamentary districts. Sarawak, the other state in Borneo, could not be included because it holds its elections for state legislatures on off years.

Turning to the results of table 2.2 we see that the measures based on 2005 vastly

[44]The process can be altered so that values are not calculated for areas over the sea. Since areas over the sea are not included in any calculations, this does not alter the results.

outperform their 2004 counterparts regardless of the specification. This leads me to conclude that whatever may have been lost by relying on 2005 data is made up for by greater accuracy in the dataset all around. Moreover, $IPFixPerVoter$, which counts only IP addresses that haven't changed location, performs very badly and is in fact negatively correlated with Internet subscriptions per household.

In (2) I drop Kuala Lumpur, leading to an immediate increase in correlation for all 2005 measures apart from $IPFixPerVoter$. I interpret this result as arising from the large bias toward Kuala Lumpur mentioned above. Since Kuala Lumpur is measured with such error and does not participate in state legislative elections, I exclude it from my sample.[45] In specification (3), I limit the sample to the 60 state legislative districts in the Borneo state of Sabah. As of 2004, Internet usage in Borneo was sparse relative to the rest of the country. As a result, the GeoIP data contains very few data points for Sabah, which in turn leads to very large bias when performing inverse distance weighting interpolation. Thus, it is unsurprising that the correlations for Sabah alone are low in comparison to (2) (Sabah + Peninsular Malaysia) and (4) (Peninsular Malaysia by itself). Sabah's markedly different ethnic makeup and political structure also make comparison with the mainland problematic. For these two reasons, I exclude Sabah from my sample.

In specification (4), which only counts peninsular Malaysia, $IPSumPerVoter$ 2005 greatly outperforms the other measures. As an additional test, for each measure I generate the corresponding 2008 value and calculate growth from 2004 to 2008. All measures estimate negative growth for a few observations (i.e., the number of IP addresses associated with a legislative district is greater in 2004 than in 2008). In this period, the Internet penetration rate for the country as a whole increased from 40% to 60%, and the number of IP addresses more than doubled.[46] Thus I interpret instances of negative growth as measurement error. Again, $IPSumPerVoter$ 2005 performs the best, with only 12 districts with negative growth out of 445; $IPmaxPerVoter$ has 21, and $IPAvgPerVoter$ 50. For the remainder of this paper, I use the natural log of $IPsumPerVoter$ *2005* for my $IPperVoter$ measure in 2004 and the natural log $IPsumPerVoter$ 2008 for $IPperVoter$ in 2008, and drop the 12 districts in which $IPperVoter$ growth is negative. Unless stated otherwise, results are largely unchanged by including these 12 districts. Of the remaining 433 districts in peninsular Malaysia, I drop 6 regions because of uncontested elections in either

[45]Unless stated otherwise, dropping Kuala Lumpur has no significant effect on my results.
[46]See <data.worldbank.org> for national penetration statistics; APNIC for IP allocation numbers.

2004 or 2008.[47]

Measurement error Specification (5), corresponding to the sample used in the paper, indicates a correlation of 0.63 between $IPperVoter$ and the benchmark. Although this result indicates a strong correlation, it still leaves a large, unexplained difference between the two measures.

There are several explanations for why this difference occurs. First, since HBAIS data only counts households with Internet subscriptions, it most likely underestimates the percentage of households with access to the Internet by omitting people who access the Internet at work or in Internet cafes. $IPperVoter$ should capture IP addresses tied to work and to Internet cafes. However, since $IPperVoter$ contains no metric for intensity of usage, it would not take into account that hundreds of individuals might use the same IP address on any given day.

For reasons stated above, much of this difference likely arises from measurement error. The majority of this error consists of bias toward major cities, first because of the nature of the GeoIP database, which only counts city centers, and second because of the IDW interpolation technique, which treats the centers of major cities as peaks, with connectivity decreasing as we move outwards.

Fortunately, though the measurement error is large, if anything it will lead to an underestimation of my results below. Indeed, in Appendix 2.11, I show formally that the under-sampling of IP addresses farther from major cities yields OLS estimates that are biased toward zero. In section 2.5.4, I use instrumental variables that are uncorrelated with the measurement error to derive consistent estimates.

[47] I've run a separate set of regressions including uncontested seats, counting BN share as 1 if the BN wins unopposed and 0 if the opposition wins unopposed. Results are more significant.

2.11 Appendix B: Derivation of bias from the undersampling of Internet connectivity in areas far from large cities

In this section I draw from Nunn (2008) to show that the undersampling of IP addresses per voter for areas outside the largest cities will result in OLS estimates of the effect of IP per voter growth on incumbent vote share that are biased toward zero.

Denote the true log IP addresses per voter in district i as s^*_{it}, the observed number as s_{it}, the distance to the nearest large city d_i and incumbent vote share by y_{it}. d_i is expressed as deviation from the mean. Assume the true relationship between the change in log of IP addresses and distance from a major city is:

$$\triangle s^*_{it} = -\alpha d_i + \triangle e_{it} \tag{2.11.1}$$

where $\alpha > 0$ is and e_{it} is i.i.d. and drawn from a normal distribution.

Next, turning to the undersampling of regions farther from large cities, assume that the relationship between the observed change in the log of IP addresses, $\triangle s_{it}$, and the distance to the nearest major city is given by:

$$\triangle s_{it} = \triangle s^*_{it} - \gamma d_i + \triangle v_{it} \tag{2.11.2}$$

where $\gamma > 0$ and v_{it} is uncorrelated with e_{it} and d_i.

The true relationship between change in BN share and change in log IP addresses per voter is given by

$$\triangle y_{it} = -\beta \triangle s^*_{it} + \triangle w_{it} \tag{2.11.3}$$

where $\beta > 0$ and w_{it} is uncorrelated with all other variables.

If we perform a simple OLS estimate of the effect of observed IP per voter on BN share, $\triangle y_{it} = b\triangle s_{it} + \triangle u_{it}$, we get:

$$\hat{b} = \frac{\sum_i \triangle s_{it} \triangle y_{it}}{\sum_i (\triangle s_{it})^2} \tag{2.11.4}$$

Substituting (2.11.1) into (2.11.2) and taking the first difference gives:

$$\triangle s_{it} = -(\alpha + \gamma) d_i + \triangle e_{it} + \triangle v_{it} \tag{2.11.5}$$

Substituting (2.11.1) into (2.11.3) gives:

$$\triangle y_{it} = \beta\alpha d_i - \beta\triangle e_{it} + \triangle w_{it} \tag{2.11.6}$$

Finally substituting (2.11.5) and (2.11.6) into (2.11.4) and taking the plim gives :

$$plim\hat{b} = -\beta \frac{\sigma^2_{\triangle s^*} + \alpha\gamma\sigma^2_d}{\sigma^2_{\triangle s^*} + \gamma(2\alpha + \gamma)\sigma^2_d + \sigma^2_{\triangle \nu}} \tag{2.11.7}$$

where $\sigma^2_{\triangle s^*} = \alpha^2\sigma^2_d + \sigma^2_{\triangle e}$

In the case where $\gamma = 0$ we are in the situation of classical measurement error and the result reduces to standard attenuation bias.

If $\gamma > 0$ we are in a situation where the underestimation of IP addresses per voter is increasing in distance from a major city. Since $2\gamma(\alpha + \gamma) > \alpha\gamma$, (2.11.7) will be biased toward zero in this case as well.

2.12 Appendix C: Proofs of propositions

2.12.1 Proof of Proposition 1

As in Besley and Prat (2006), I focus on pure strategy, perfect Bayesian equilibria.

First I start with the voters. Since voting is sincere, each voter observes media reports and updates the posterior probability that the incumbent is good $\hat{\gamma}$. She votes for the incumbent if and only if $\hat{\gamma} > \gamma$. Suppose that viewers observe each of the two signal realizations with positive probability. Then it must be that $\hat{\gamma}(\tilde{s}_i = b) < \gamma < \hat{\gamma}(\tilde{s}_i = \emptyset)$. To see this, suppose there is a pure strategy equilibrium where the incumbent is kicked out if $s = \emptyset$: $\hat{\gamma}(\tilde{s}_i = \emptyset) < \gamma$. If this were the case, then the incumbent would never buy off the media. This in turn would cause voters to update their posterior such that $\hat{\gamma}(\tilde{s}_i = \emptyset) > \gamma$, a contradiction. This means that if $s = \emptyset$ the incumbent party is always reelected and if $s = b$ the incumbent party is reelected if and only if at least half of the voters observe $\tilde{s}_i = \emptyset$.

Now I move to the decision of media firms. Suppose that τ_W is so high that the incumbent will never choose to capture the web based media firm. Next, suppose that the mainstream outlet has been offered t_M to suppress its signal and it knows that the web firm will not suppress its signal. The mainstream firm's payoff is:

$$\pi_M = \begin{cases} a\left(1 - \frac{\phi}{2}\right) & \text{if she rejects} \\ \frac{t_M}{\tau_M} & \text{if she accepts} \end{cases}$$

Thus she accepts if and only if

$$t_M \geq \tau_M a\left(1 - \frac{\phi}{2}\right)$$

Finally, consider the incumbent. The web media is too costly to capture by definition. If $\Phi \geq \frac{1}{2}$ at least half of the voters will receive the bad signal and the incumbent will lose regardless of whether the traditional media is captured. Thus a bad incumbent will not capture either outlet and will be discovered with probability q. If $\Phi < \frac{1}{2}$ a bad incumbent will capture the mainstream firm if the return from

reelection is greater than the cost

$$r \geq \tau_M a \left(1 - \tfrac{\phi}{2}\right)$$

I have thus shown that the incumbent will capture the media under the conditions specified in Proposition I.

2.12.2 Proof of Proposition 2

The effect on turnover and the media is explained in the text. The remaining results follow directly from Proposition 1.

2.12.3 Proof of Proposition 3

The proof is identical to Proposition 1 except for the equilibrium strategies and decisions of the incumbent.

Consider the incumbent. Assume $\Phi \geq \frac{1}{2}$ and either one or both of the following conditions are met, such that $\phi < \frac{1}{2}$ for a majority of districts:

1. Internet penetration is rightly skewed around $\phi = \frac{1}{2}$
2. ϕ and ψ are positively correlated

Then a bad incumbent will capture the mainstream media as the conditions ensure the victory in a majority of seats.

2.13 Appendix D: Data appendix

Data Descriptions and Sources

Variable	**Description**	**Source**
Original Variables		
BNShare	Share of the vote won by a member of the Barisan Nasional.	Election Commission (1986, 1990, 1995, 1999, 2004, 2008)
Turnout	Percentage of eligible voters who voted in a district. Includes spoilt ballots.	Election Commission (1986, 1990, 1995, 1999, 2004, 2008)
Turnover	Dummy indicating whether BN retained seat.	Election Commission (1999, 2008)
Eligible voters	Number of people eligible to vote in a district.	Election Commission (1986, 1990, 1995, 1999, 2004, 2008)
% ethnicity	Percentage of the voters in a district who are of each ethnic group. The omitted category is Chinese/Other. This number is reported as part of the election results in all major dailies.	Malaysiakini (2004), New Strait Times Press (1999)
Generated Variables		
GDP	Measure of average GDP per capita at the mukim (census district) level, generated by the consultancy Booz & Company. Aggregated up to the state legislature district level using ArcGIS and LandScan as outlined in section 2.4.2.	Booz & Company (2005)

InternetHH	Fraction of households with an Internet subscription at the mukim level. Aggregated up to the state legislature district level using ArcGIS and LandScan as outlined in section 2.4.2.	Household Basic Amenities and Income Survey (2004), Population and Housing Census (2000)
% Ethnicity 1991	Percentage of the voters in a district who are of each ethnic group. The omitted category is Chinese/Other. Data is available at the mukim level in 1991 population census. Aggregated up to the state legislature district level using ArcGIS and LandScan as outlined in section 2.4.2.	Population and Housing Census (1991)
Slope std	The standard deviation of the average steepness of land in a district. Calculated from digital elevation satellite imagery using ArcGIS first to calculate the slope at each point and then to derive the average and standard deviation across a district.	Pitney Bowes (2008)
% Urban rural	Percentage of a district that is classified as urban and rural farm using satellite imagery. The omitted category is jungle. Used ArcGIS to calculate percentage at district level.	Pitney Bowes (2008)
Road density	Kilometers of road in a district divided by total area of the district. Calculated using ArcGIS.	Pitney Bowes (2008)
Population density	From Oak Ridge National Laboratory LandScan product, which uses census data in conjunction with satellite information to estimate population at the 1 km resolution. I use ArcGIS to aggregate up to the district level.	LandScan, Oak Ridge National Laboratory (2008)
Km to roads	Distance from the centroid of a district to the closest major road and closest federal road as of 2008. The road data is from Pitney Bowes. ArcGIS was used to calculate the centroid of each district and then derive the distance measure.	Pitney Bowes (2008)

Km to Time	Shortest distance from the centroid of a district to Time dotCom's backbone. Location of Time dotCom's backbone from company records. Distance generated in ArcGIS.	Time dotCom (2004)
Km to Fiberail	Shortest distance from the centroid of a district to Fiberail's backbone, which follows major railroads. Distance generated in ArcGIS.	Pitney Bowes (2008)
Km to TM	Shortest distance from the centroid of a district to Telekom Malaysia's backbone. From 2004 annul report. Distance generated in ArcGIS.	Telekom Malaysia (2004)
IPperVoter	The number of IP addresses per eligible voter. Constructed with ArcGIS from IP geolocation data from MaxMind in conjunction with records from APNIC. See section 2.10 for details.	MaxMind (2004-2008), APNIC (2004-2008)

2.14 Appendix E: Appendix on Election Irregularities

There are several factors that lead to election irregularities in the 2004-2008 period. First, there were some large discrepancies between the number of ballots issued for parliamentary seats and their corresponding state legislature seats. Recall from figure 2.1 that each parliamentary seat is made up of a handful of state legislature seats. Technically the number parliamentary ballots should be the same as the sum of the ballots for all the constituent state legislature seats. However, there were several large discrepancies in this respect, most notably the Kuala Terengganu parliamentary seat where there was a difference of 10,000 between the number of parliamentary ballots issued (around 70,000) and the number of state legislature ballots issued. I drop seats where the discrepancies are suspiciously large: 5 state legislature districts corresponding to the Kuala Terengganu parliamentary seat (Wakaf Mempelam, Bandar, Ladang, and Batu Buruk); and the 4 state legislature districts corresponding to the Setiu parliamentary seat (Batu Rakit, Jabi, Langkap, and Permaisuri).

The 2004 election was marked by unnaturally high turnout rates, greater than 90% in several instances. To deal with this, I drop districts where turnout exceeded 80% in 2004 and turnout differed by more than 10% from its level in the 1999 election. Due to redistricting, boundaries do not perfectly match between 1999 and 2004. In order to generate a 1999 turnout value for a 2004 district I use the population weighted LandScan procedure outlined in section 2.4.2. This rule leads me to drop 6 additional districts: Lunas, Nenggiri, Sungai Udang, Chini, Kuala Nerang, and Sungai Tiang.

3 The Political Impact of the Internet on US Presidential Elections

3.1 Introduction

The Internet is said to have played a key role in the 2008 U.S. presidential campaign: the Obama campaign's online fundraising arm brought in a record $500 billion in small individual donations; and the campaign's heavy use of social media purportedly contributed to the highest rate of youth turnout since voting was extended to 18-year-olds. A common theme has emerged in the U.S. press expressed by the American online magazine, Huffington Post: "were it not for the Internet, Barack Obama would not be president."[1]

We test the extent to which these assertions hold, looking at the effect of Internet diffusion on campaign contributions, turnout and vote share in the U.S. presidential elections. We focus on the 1996-2008 election period, starting from an election in which the Internet had virtually no role, to arrive to the 2008 election in which most commentators saw a powerful and almost decisive role for the Internet and social networks.

We develop a proxy for Internet usage over the 1996-2008 period. To date there is no publicly available data on Internet usage in the U.S. at anything below the state level. We develop a county-level proxy, using data on the number of high speed

[1] See Schiffman (2008)

ISPs that are registered with the FCC in a zip code. We test this proxy against state-level measures and find a high correlation.

We also assemble a novel dataset of political outcomes, combining presidential vote data, with turnout and a full record of campaign contributions going back to 1996.

A study of this sort poses a formidable obstacle: endogeneity of Internet penetration could lead to correlations with the variables of interest, which do not accurately reflect causation. To address the problem of endogeneity, we exploit geographic discontinuities along state borders with different right-of-way laws. We argue that right-of-ways laws provide an exogenous source of variation and determine the cost of building Internet infrastructure.

We compare counties, on either side of state lines, whose unobserved characteristics should be spatially correlated. By taking the difference between regions on either side of the border and instrumenting by right-of-way laws, our methodology controls for unobserved characteristics, providing us with estimates of the causal effect of Internet growth on outcomes of interest.[2]

We find a strong causal effect of Internet access on share of the vote going to the Democratic candidate and on campaign contributions. A one standard deviation increase in Internet access translates into a 1.8 point swing to the Democrats and a $3 increase in donations for every one hundred people. We find some evidence of an effect on turnout—a one standard deviation increase in Internet penetration translates into a .2% increase in turnout—but this is not robust to our IV specification.

This study makes a number of contributions. First, it provides some of the first quantitative evidence of the role of the Internet in U.S. elections. There is very little economic research in this area right now. Gentzkow and Shapiro (2011) uses survey data to show that, opposite to common wisdom, online news consumption is not associated with higher ideological segregation than offline news. From the perspective of political science, Golde and Nie (2010) measures the effect of online news on political participation and polarization. They also find no effect of online news readership on participation or polarization. Andersen, Bentzen, and Dalgaard (2011) looks at the role of the Internet in curbing corruption in the U.S., using lightning strike density as an instrument for Internet diffusion.

From a theoretical perspective, there is very little work looking into the effect of

[2]In our use of spatial differencing to control for unobserved characteristics, our empirical methodology is similar to that of Duranton, Gobillon, and Overman (2011).

online media.Edmond (2011) presents a model of media and regime change that distinguishes between the effect of print and broadcast media and online social media.

Although there are few papers looking into the relationship between Internet-based media and politic outcomes, there is a rich literature on the political economy of traditional media. From an empirical perspective research has been conducted into the effects of newspapers on government responsiveness (Besley and Burgess, 2002); newspapers on reducing corruption (Reinikka and Svensson, 2004); newspapers on federal spending in a district (Stromberg, 2004) ; radio on political violence (Yanagizawa-Drott, 2012); and television on presidential elections (DellaVigna and Kaplan, 2007).

From the theoretical side, Mullainathan and Shleifer (2005) find that increased competition in the media market could lead to increased bias as newspapers slant their news toward their readerships' priors. Alternatively, Besley and Prat (2006), presents a theoretical framework for government capture of the media and shows how increased competition in the media market can yield better information on candidate quality and increased turnover.

This paper relates to a growing literature that employs a methodology from spatial econometrics to achieve identification. Naidu (2012) estimates the effect of 19th century disenfranchisement of African Americans in the U.S. South, comparing adjacent county-pairs on state boundaries. Dube, Lester, and Reich (2010) use the same methodology in an earlier paper to estimate the effects of minimum wages on earnings. This paper employs a somewhat different method for comparing cross border county pairs, supplemented by an instrumental variable approach.

We start in Section 3.2, delving into background on politics, media, and the Internet. We present the anecdotal evidence that suggests a causal role for Internet access in U.S. politics. In Section 3.3, we describe the data sources. Section 3.4 presents basic results from OLS regressions and then moves onto our IV empirical strategy and results. In Section 3.5 we conclude.

3.2 Background

In this section we will provide background on the growth of Internet usage in the United States and then move on to the specific ways in which Internet based tools

were employed by the Obama campaign. We will end with a set of predictions on the effects of the Internet on election outcomes.

3.2.1 Personal Use of Internet

Internet use has grown enormously from 16% in 1996 to 74% in the 2008.[3] This masks an even greater change in the method and sophistication of usage over this period of time. In 1996, AOL was still one of the main methods of accessing the Internet, meaning relatively few people left its portal, effectively a walled garden of curated and licensed content. Moreover, most connections were dial-up and too slow for to display video content. Modern social networks did not yet exist. Bill Clinton did have his own web-page during the 1996 campaign, but it was small, with none of the fundraising and organizational tools that we see in modern campaigns.[4] Unsurprisingly, only a small fraction of the voting population used the Internet for political ends, 4% according to a Pew study.[5] In contrast, in the same study Pew found that as of 2008 a full 44% of the adult population used the Internet as a source of political information, and that it was the primary source of political information for 30% of the population.

Although the stereotype holds that a typical Obama supporter was more wired than a McCain supporter, in practice the opposite is the case: 83% of McCain supporters were Internet consumers as opposed to 76% of Obama supporters.[6] However, much of this is explainable by the gap in average income and education between the two groups. Moreover, whereas the average McCain voter might be more connected, the average Obama voter was much more likely to use the Internet to political ends.

Nowhere is this more evident than in the case of young voters aged 18-29, who, at 72%, were the most politically active online of all age groups.[7] They also swung disproportionately to Obama: 66% of the youth vote went to Obama in 2008 as opposed to 54% to Kerry in 2004.[8] Moreover, turnout among the youth was the highest on record since voting was extended to 18 year-olds in 1972.[9]

[3] See data.worldbank.org
[4] See www.4president.us/websites/1996/clintongore1996website.htm
[5] See Smith and Rainie (2008) p.5.
[6] See Smith and Rainie (2008)p.10.
[7] The number drops to 65% for people aged 30-49, 51% for 50-64, and 22% for 65+. For additional details see Smith and Rainie (2008) p.17.
[8] See Keeter, Horowitz, and Tyson (2008)
[9] See Falcone (2008)

3.2.2 Campaign Use of Internet

New Media The Obama campaign exploited the Internet in a number of ways. First, they made ample use of social media to further their cause both through Facebook and the campaign website my.barackobama.com (MyBO). These tools helped augment traditional campaign tactics: detailed information on supporters helped improve mobilization, especially during caucuses; MyBO supplied tools allowing volunteers to make calls on the campaign's behalf from home; and MyBO and Facebook centered tools also helped volunteers organize their own fundraising events, connecting with friends they hadn't seen for years.[10] The Obama campaign's aggressive action in the social media space played out in exit-survey data. According to Pew, 25% of Obama supporters used social networks for political ends as opposed to 16% of McCain supporters.[11]

The Obama campaign also exploited the potential of online video to get their message to a large audience without having to pay traditional advertising costs for television. For example, a video of Obama's speech on race relations got 6.7 million views by November 2008.[12] Again this translated into a gap among supporters: 21% of Obama supporters shared political videos as opposed to 16% of McCain voters.[13]

All of these trends were especially pronounced among young voters aged 18-29, the largest users of all types of social and online media: where 67% of young voters reported watching online political videos; and 49% used social networks politically, with 40% posting original online content relating to the campaign.[14]

Fundraising The area of Obama's online campaign that has received the most attention is fundraising, where the campaign brought in an estimated 500 million dollars in online donations, eclipsing Howard Dean's previous record of 27 million. In fact the technology used for Obama's online fundraising was developed by veterans of the Dean campaign, a company called Blue State Digital. They built a number of tools to make the effort more "social": people could set their own personal targets; run fundraising campaigns; and watch personal thermometers rise, which gauged

[10]See Talbot (2008)
[11]See Smith and Rainie (2008) p.11.
[12]See Miller (2008)
[13]See Smith and Rainie (2008) p.11.
[14]See Smith and Rainie (2008) p.17.

how well they met their targets.[15] The results in terms of exit polls is dramatic: Pew reports that 15% of Obama voters donated online in contrast to 6% of McCain voters.[16] The results are equally striking in terms of type of donation: of the 6.5 million donations online, 6 million were in increments less than $100 and often from the same person.[17]

3.2.3 Expected Outcomes

For reasons outlined in the above sections, we expect an increase in Internet access to cause:

1. *Increase in donations for Democratic candidate, particularly among small donations less than $500 dollars.* Due to the Obama campaign's use of its online portal to collect a record amount of small-scale donations.
2. *Increased turnout, in particular among youth.* Since the Obama campaign employed extensive tools for interacting with and mobilizing voters, and usage of these tools was highest among people aged 18-29.
3. *Increased support for the Democratic candidate.* Due to the tools used by the Obama campaign both to convince swing voters (e.g. via video) and mobilize voters.

3.3 Data

3.3.1 Census Data

We use county-level census and survey data to generate controls for presidential elections during the 1984-2008 period. Data on ethnicity, age, and sex is available every for every election year. Data on poverty, income, education, and employment is available for 1980, 1988, 1992, 2000, 2004 and 2008 so values for 1984 and 1996 are calculated by taking the average of 1980 and 1988 and 1992 and 2000 respectively.

Alaska, Hawaii, D.C., and outlying U.S. territories are excluded from the sample because their geographic placement makes them outliers and they cannot be used

[15]For more details, see Talbot (2008).
[16]See Smith and Rainie (2008) p.11.
[17]See Vargas (2008).

as part of the identification strategy.[18] There is very little change in counties during this period: the sample size increases from 3076 counties in 1984 to 3077 counties in 2008. Figure 3.1 shows a complete picture of the counties in the mainland U.S. as of 2008.

3.3.2 Political Data

All political variables are likewise available at the county level, covering presidential elections from the 1984-2008 period. Variables on absolute number of votes are derived from FEC data.[19] This allows us to derive a measure of Democratic vote share for each presidential election.

Turnout is also calculated at the county level. It is the ratio of the number of votes cast divided by the estimated voting age population per data from the U.S. Census Bureau.

Donation data is also available for all elections from the FEC. However it does not include donations less than $200 unless these donations are from the same individual and add up to more than $200. As a result, the sample misses a significant amount of smaller donations. However, since the Obama campaign made the greatest inroads among small donors, raising half a billion dollars, if anything this should work against our results. We create measures in per capita terms: total amount donated; total amount donated to Democrats; total amount donated to Democrats less than $500.

3.3.3 Right-of-Way Data

Different levels of broadband penetration across states can be attributed in part to different regulatory regimes concerning Right-of-Ways (ROW) laws. According to Day (2002), the current ROW regime can be traced to the Telecommunications Act of 1996, which allowed municipalities to regulate the public ROW. States passed their own laws concurrent with the Telecommunications Act that either limited or reinforced this municipal right, leading to the significant variation in ROW regulations across states that we see today.

[18] OLS results are robust to the inclusion of these observations.
[19] See David Leip's Election Atlas at uselectionatlas.org.

ROW laws can be broken down into a number of different categories, which influence the cost for ISPs to lay infrastructure:[20]

1. *Jurisdiction:* In some states an ISP has to get permission to build on the public ROW from every single municipality that the project crosses, whereas in other states this is handled by a centralized authority.

2. *Compensation:* Compensation demanded by municipalities in return for granting ROW permission ranges from cost recovery (the cost to the municipality of administering the ROW) to a rental fee (e.g. percentage of gross revenue) to a flat tax.

3. *Timeliness:* Some states have a maximum time for processing permit applications, significantly speeding up the process.

4. *Mediation and Condemnation:* States also vary in how they deal with conflicts between municipalities and ISPs, and private landowners and ISPs. For example, in Vermont, landowner complaints can be heard on a wide range of issues including aesthetics, and decisions are appealable to the Vermont Supreme Court. On the other extreme is Texas, where most factors are not appealable and landowners must pay the ISPs legal expenses if they lose in court.

5. *Remediation and Maintenance:* These laws dictate issues such as in what state of repair ISPs must maintain their facilities. For example, if a sidewalk is torn in order to lay cabling, these laws determine to what extent the sidewalk would need to be restored to its original state and under what time frame.

In 2002 TechNet, an industry lobbying organization that counts almost every major technology company among its members, released a report on the state regulations influencing supply and demand of broadband.[21] As part of this report, the authors compiled an index, ranking the regulatory regime in terms of ROW laws across states.[22] We use this index to capture the regulatory difference in ROW laws across states.

We will return to this issue in Section 3.4.2, where we will show the extent to which ROW laws influence differential levels of broadband diffusion.

[20]See NARUC (2002) for more details.
[21]See Kende and Analysys (2002).
[22]See Kende and Analysys (2002) for details on the construction of this index.

3.3.4 Internet Data

The U.S. lacks comprehensive, publicly available data on broadband usage and availability at anything finer than the state level. ISPs won't release private information on subscribers due to its proprietary nature, and survey data is scant. We use a common alternate measure, the number of broadband providers operating in a zip code, to proxy for broadband usage.

This data was collected by the FCC's Form 477 on a semi-annual basis for 1999-2010 from all high speed providers with more than 250 high speed lines in a state. A provider is counted as high speed if transfer speed is greater than 200 kilobits per second in at least one direction. The data does not differentiate between cable, DSL, satellite, residential, or business providers. A provider is counted if they have at least one subscriber in a zip code.

We make two assumptions that allow us to integrate this data into our analysis. First, the number of providers in zip codes with less than three providers but more than zero is not provided. We take the average and count all these areas as two. Second, there is no information for the 1996 period. Since high speed lines were non-existent at this time outside of a few universities and companies, we assume that high speed connectivity was zero at this time.

There are a number of reasons that this is a good proxy for Internet take-up. In table 3.1 we regress the log of high-speed lines per capita on log providers. As can be seen in specification (1) even without controlling for population or state-year fixed effects, the relationship between log providers and log of high-speed lines is positive and highly significant with an r-squared of .54. The strength of this relationship is unaffected by introducing a control for population in specification (2) and the r-squared increases. In specifications (3) and (4) state-year fixed effects are introduced, barely affecting the coefficient on log providers and yielding markedly higher within r-squared values. Specification (3) can be seen graphically in figure 3.2, which shows a strong positive relationship apart from three outliers in the bottom right corner. In specifications (5) and (6) we drop these outliers, yielding a far more significant relationship and, in the last case, a within r-squared of .96.

One concern is that this relationship is far less significant at the county level. Fortunately Kolko (2010) analyzes the relationship between high-speed Internet adoption and a number of providers at the zip code level, using zip-code level survey data from Forrester. As can be seen in figure 3.3 from Kolko (2010), there is a strong

positive, monotonic relationship between the number of providers in a zip code and its level of high-speed Internet take-up. The only outliers are at the extremes of the distribution for which there are very few observations: zip codes with zero providers and zip codes with greater than 20 providers.

We aggregate the data up to the county level, weighing zip codes by their population.[23] Although the relationship between number of providers and Internet take-up is significant, there are several factors that introduce measurement error. High-speed Internet does not include dial-up connections, which, as we saw in Section 3.2.1, allowed 16% of the population to connect to the Internet in 1996. We don't think this will introduce much error into our results for a number of reasons. First, as mentioned in the same section above, only 4% of the population reported using the Internet for obtaining political information in 1996. Moreover, given the speed limitations inherent in dial-up, many of the modern techniques employed by campaigns such as video streaming are not available. Another concern is that in some cases, this measure may only be proxying for the competitiveness of ISP markets in a county. Although this will certainly introduce error, the strong correlation that we find between the number of ISPs and high-speed Internet adoption reassures us that there will still be substantial variation related to Internet usage.

3.4 Method and Results

Our econometric analysis is based on panel data regressions of the form:

$$y_{ist} = \alpha_i + \beta_t + \mu Internet_{ist} + \xi x_{ist} + \lambda_{st} + \varepsilon_{ist} \tag{3.4.1}$$

where y_{ist} is an outcome variable in county i and state s at time t; $Internet_{ist}$ is the number of providers; x_{ist} are other exogenous variables; α_i is the county fixed effect; β_t is the year fixed effect; and λ_{st} is the year state trend.

The county fixed effect captures county-specific, time-invariant factors. The year fixed effect captures common shocks in particular years. λ_{st} captures trends common across counties within a state.

[23] In almost every case, a zip code fits within a singe county. All results reported below are robust to omitting observations containing zip codes that span multiple counties.

3.4.1 Basic Results

Table 3.2 looks at the relationship between share of the vote for the Democratic presidential candidate and the growth in ISPs from 1996-2008. Column (1) shows a large and significant association between Democrat share and the number of ISP providers, implying that counties with more Internet are also more prone to vote for Democratic candidates. In general Internet is cheaper to install in high density areas, where a large number of households can be connected without having to lay long stretches of cable. Since the Democratic party performs better in urban areas, a primary concern is that we are simply capturing this effect. Column (2) helps allay this concern: we introduce a first set of controls for average age, percent male, and population density. The magnitude of the effect decreases slightly, but it remains highly significant. In column (3) we further control for the ethnic make-up of the county. The coefficient on our Internet measure remains highly significant, although not as strong. In column (4), controls are added for percent of the population below the poverty line and for the log of income. As we can see the results are barely changed.

Column (5) shows estimates with a full set of controls, including controls for education. The strength of the association diminishes, but is still highly significant, showing a strong relationship between Democrat vote share and Internet growth in the face of a wide variety of controls. To give an interpretation of the coefficient, a one standard deviation increase in the number of providers translates into a .3 point swing towards the Democratic candidate.

Finally, there is the possibility of a non-linear relationship between Internet penetration and unobservables. We look into this problem in column (6) by incorporating a lagged dependent variable into the model. We include lagged Democrat election share on the right hand side without fixed effects, exploring the possibility that current election results are influenced by past election results. The coefficient on ISP providers is still positive and significant, but much smaller in magnitude. Per Pischke and Angrist (2010), this provides us with a lower bound on our estimate, providing significant results even in the case where the true model is a lagged dependent model.[24]

[24]Conversely, if the fixed effects model is the true model, the lagged dependent variable specification will underestimate the size of the coefficient of interest. See Pischke and Angrist (2010) (p. 246) for more details.

In Table 3.3 we look at results for the other outcomes of interest with our full set of controls. Column (2) shows a strong association between Internet access and turnout. A one standard deviation increase in internet penetration translates into a .2% increase in turnout.

Columns (3)-(6) move on to per capita donation results. In column (3) we see a strong relationship between total donations per capita to both parties and Internet penetration. To give a sense of magnitudes, a one standard deviation increase in Internet penetration translates into a $3 increase in donations for every hundred people. This relationship remains positive and significant for all outcome variables: donations per capita less than $500 in column (4), total donations per capita to Democrats in column (5), and per capita donations under $500 to Democrats in column (6).

In Table 3.4, we run the same set of regressions with a lagged dependent variable and no fixed effects. Column (2) shows that turnout loses its significance under this specification. Similarly, donations results lose their magnitude and significance in columns (3)-(5). However, column (6) shows that the relationship between Internet access and per capita Democratic donations less than $500 remains significant. This result is in line with the anecdotal evidence that suggests that the greatest effect was on small scale donations to the Democratic candidates.

In sum, we find a strong relationship between Internet access and fraction of the vote for the Democratic candidate, and with donations to the Democratic party under $500 dollars. Strong, but less robust associations have also been found with turnout and donations on other levels.

3.4.2 Endogeneity Concern

While fixed effects control for any time-invariant county characteristics and the state trend accounts for any state level shocks, there is still the possibility of county level shocks. In particular, if the Internet is allocated more heavily to areas that are trending linearly towards the Democrats for unobservable reasons, then the preceding specifications will overestimate the result.[25]

[25]For example, Stephens-Davidowitz (2011) provides evidence that racism played a large role in the outcome of the 2008 election. If racism was on the decline in areas with more Internet penetration, my results would be overestimated.

To address this concern, we use an instrumental variable approach, exploiting exogenous variation across state borders in Internet supply due to different ROW laws as outlined in Section 3.3.3. Limiting our sample to counties on either side of a state border, we use the state level broadband index discussed in Section 3.3.3 as our instrument (W_{ij}). We further limit our sample to border's that have at least five counties on either side, which can be seen in Figure 3.1.[26] Our specification is as follows:

$$y_{ist} = \alpha_{1i} + \beta year_{1t} + \mu Internet_{ist} + \xi x_{ist} + \varepsilon_{ist} \tag{3.4.2}$$

$$Internet_{ist} = \alpha_{2i} + \beta year_{2t} + \sum_{t=1996}^{2008} \theta_{st} BBIndex_{st} * year_{2t} + \xi x_{ist} + \omega_{ist} \tag{3.4.3}$$

The identification assumption is that, conditional on the baseline county characteristics —income, poverty, education, ethnicity, population density, age, gender—the broadband index does not affect change in vote share independently of growth in Internet access. Since the instrument is at the state rather than county level, this approach is equivalent to comparing the average of counties along state borders.

3.4.3 First Stage and Reduced Form

Table 3.5 presents the results of the first stage regressions. The relationship is positive and significant, implying that more favorable ROW laws are associated with higher Internet penetration in a state. In columns (2) to (5), we steadily add controls; the coefficients on our broadband index year interaction terms are for the most part unchanged, except in specification (5) where we control for education as well. In this case we note a small drop in the magnitude of the relationship, but not the significance. The F-Stat is low and unchanged across specifications.

Table 3.6 shows the results of reduced form regressions of Democratic vote share on our measure of ROW laws. In specifications (1) through (5), the coefficients

[26] Our results are robust to a minimum county range of 3-7.

are positive, significant and relatively unaffected by introducing new controls. If anything the relationship becomes more significant as controls are introduced.

3.4.4 IV Results

IV regressions of Democratic vote share on Internet growth are presented in table 3.7. Specification (1) presents basic results without any controls. The relationship is positive and highly significant, implying that increased Internet access is associated with higher Democratic vote share. As in the OLS case, a primary concern with our identification strategy is that we are simply capturing the effects of urbanization. Specification (2) presents a first set of controls for this, in particular population density. As can be seen, the coefficient on the number of providers is largely the same and highly significant. In specification (3) controls for ethnicity are included as well, with almost no change in results. In specification (4), we add controls for income and poverty and the coefficient of interest is unchanged. Finally in specification (5) we round out our full set of controls with variables for education, and there is little effect on the coefficients of interest. Together, the results imply that a one standard deviation increase in Internet access translates into a 2.1 point swing towards the Democratic candidate.

Specification (6) presents OLS results for the limited sample of counties on either side of state lines. The association is still positive and significant, but is smaller in magnitude. We attribute much of this difference to attenuation bias from measurement error. As recounted in Section 3.3.4, the number of ISP providers in a county is a noisy proxy for Internet access.

In table 3.8, we present results of regressions for our other outcomes of interest. The results largely match our earlier OLS results with lag. Specification (2) implies a positive but insignificant relationship between turnout and the number of providers. Columns (3) and (4) show positive but insignificant relationships between Internet penetration and total donations and donations less than $500 respectively. In column (5) we see a large and positive relationship between donations to the Democratic candidate and Internet growth. In this case, a one standard deviation increase in Internet providers leads to a $2 increase in donations to the Democratic Party for every hundred people. Likewise column (6) implies a significant, strong relationship with donations less than $500 to the Democratic Party. Both these results agree

with the anecdotal evidence: that the Democratic Party made better use of online fund-raising, in particular for small donations.

3.4.5 Robustness Checks

Pre-Internet Trend Tests The fundamental worry with our identification strategy is that areas with higher right-of-way index ratings are more likely to swing towards the Democratic candidate for unobservable reasons. We test for this possibility in table 3.9, looking at the reduced-form relationship between change in Democratic candidate vote share and the ROW index for elections in the pre-Internet period. All the specifications include a full set of controls. In specifications (1) and (2), corresponding to the pre-Internet period, we see no evidence of a reduced form relationship between vote share and Internet growth. In columns (3) through (5), corresponding to the 1996-2008 period, we see in each case a strong, positive coefficient on the ROW Index variable. This suggests that the results are not being driven by unobservable characteristics of areas with high ROW index scores in the pre-Internet period.

Placebo Regressions As an additional robustness check, in table 3.10 we test whether areas with higher Internet penetration were more likely to swing towards the Democratic candidate. For the pre-Internet period, we regress change in share of the vote on change in Internet from the 2004-2008 period.[27] Specification (1) looks at the 1988-1992 period and includes a full set of controls. As can be seen, there is no significant relationship between change in vote share and growth in Internet access over the 2004-2008 period. Likewise, column (2) shows that there is no significant relationship for the 1992-1996 period, when Internet usage was very small and had yet to spread into the general population. Interestingly, as can be seen in columns (3) and (4), the coefficient on Internet growth is insignificant for the 1996-2000 and 2000-2004 periods. Specification (5) shows that the coefficient is positive and highly significant only for the 2004-2008 period. This implies that the majority of the effect on vote share is driven by the 2004-2008 period. This result agrees with the anecdotal evidence, which ascribes the majority of the effort by the Democratic candidate to harness the Internet for electoral advantage to the 2004-2008 period.

[27]Results are the same if we use change in Internet for the 1996-2000 or 2000-2004 periods.

3.5 Conclusion

This paper provides some of the first empirical evidence on the effect of the rapid rise in Internet usage in the U.S. on the basic functioning of the political process. We focus on presidential elections for the 1996-2008 period, when Internet usage rose from 16% of the population accessing Internet via dial-up to 74% of the population accessing the Internet primarily from high-speed connections.

This paper's central contribution is to find a strong causal effect of Internet diffusion on presidential vote share and campaign contributions in the U.S. We find that a one standard deviation increase in Internet access translates into a 1.8 point swing to the Democrats and an $3 increase in donations for every one hundred people. We find limited evidence of an effect on turnout, implying that a one standard deviation increase in Internet penetration translates into a .2% increase in turnout. However, this result is not robust to our IV specification.

There is much scope for future research. First, our current identification strategy exploits state level exogenous variation. As a robustness check, we aim to use the identification strategy from Naidu (2012) and Dube, Lester, and Reich (2010), which works at the county level. Second, there is much work to be done in terms of disentangling the effects on sub-groups. For example, anecdotal evidence suggests that voter outreach and donation elicitation was most effective among young voters.

Finally, the effect of Internet diffusion on presidential voting opens many questions into the channels of causality. Did the Internet act primarily as a platform for coordinating the base or did it help convince new voters? Which platform was the most effective and how does the introduction of these platforms interact with traditional media sources? Is there any evidence of the Internet driving increased polarization?

3.6 Tables and Figures

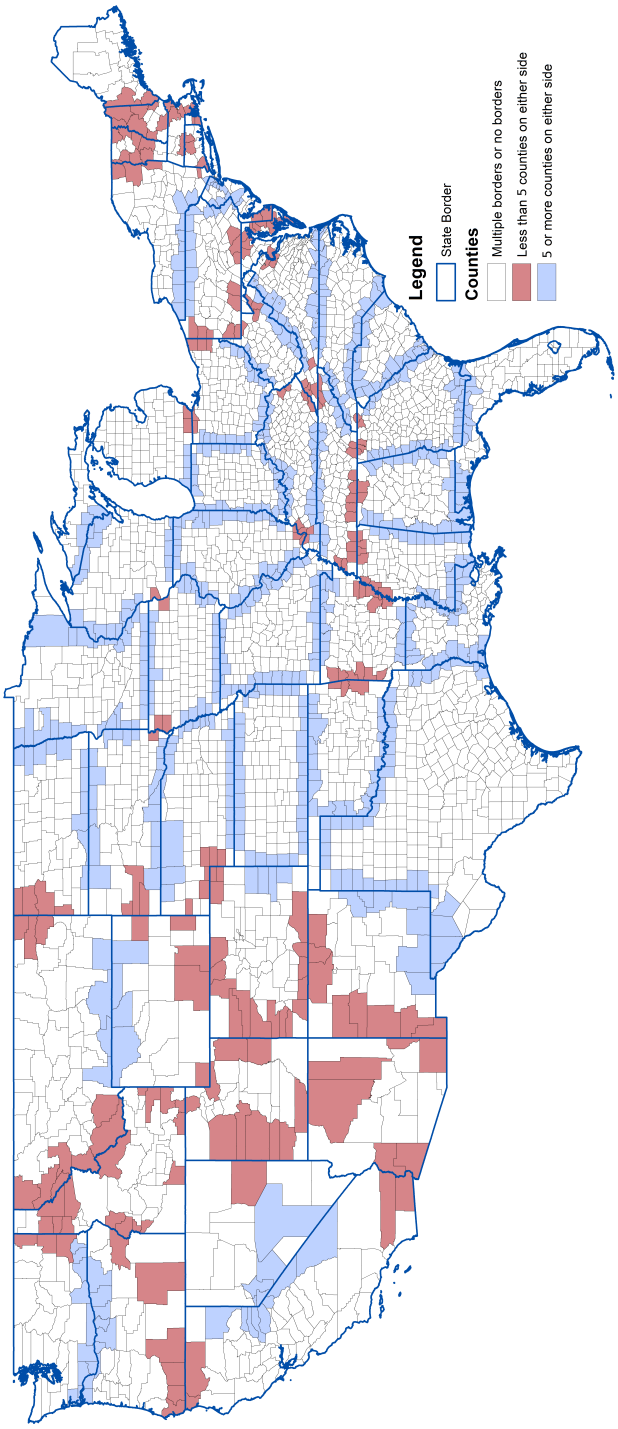

Figure 3.1: Counties in Sample

TABLE 3.1

Correlations between providers and high-speed lines per capita

	Log of high-speed lines per capita					
	(1)	(2)	(3)	(4)	(5)	(6)
Log providers	1.2668***	1.3984***	1.3100***	1.2630***	1.0866***	0.2167***
	(0.109)	(0.116)	(0.163)	(0.166)	(0.045)	(0.067)
Population		-0.0448***		-0.2207		
		(0.007)		(0.144)		
Constant	-1.9937***	-2.1597***	-3.0215***	-1.6887*	-1.3596***	-0.3973***
	(0.384)	(0.387)	(0.353)	(0.939)	(0.153)	(0.148)
N	452	452	452	452	449	449
R^2	.5414	.5754	.7993(W)	.8005(W)	.5613	.9558(W)
State and year FE	N	N	Y	Y	N	Y

Notes. This table presents the regression of the log of the number of ISP providers in a state on the number of high-speed lines per capita. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

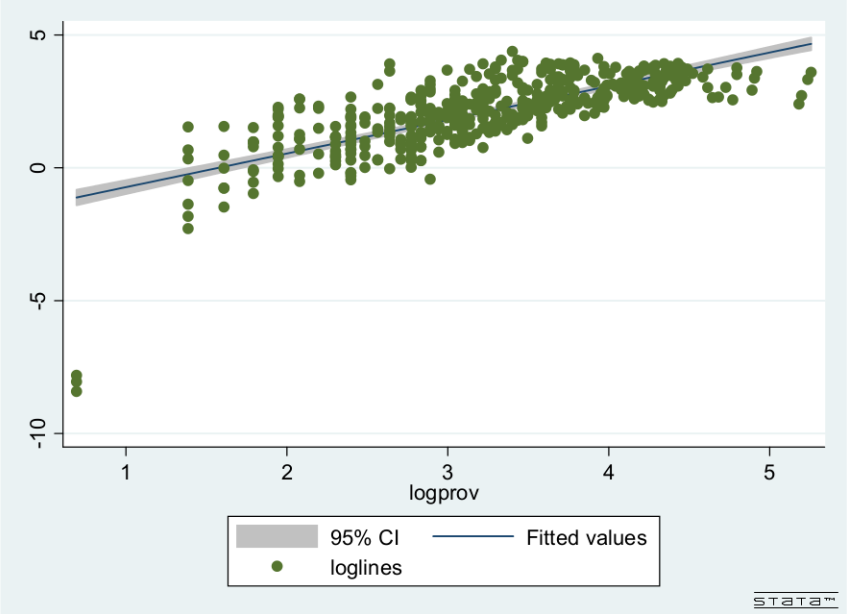

Figure 3.2: Correlation Between Providers and High Speed Lines per Capita

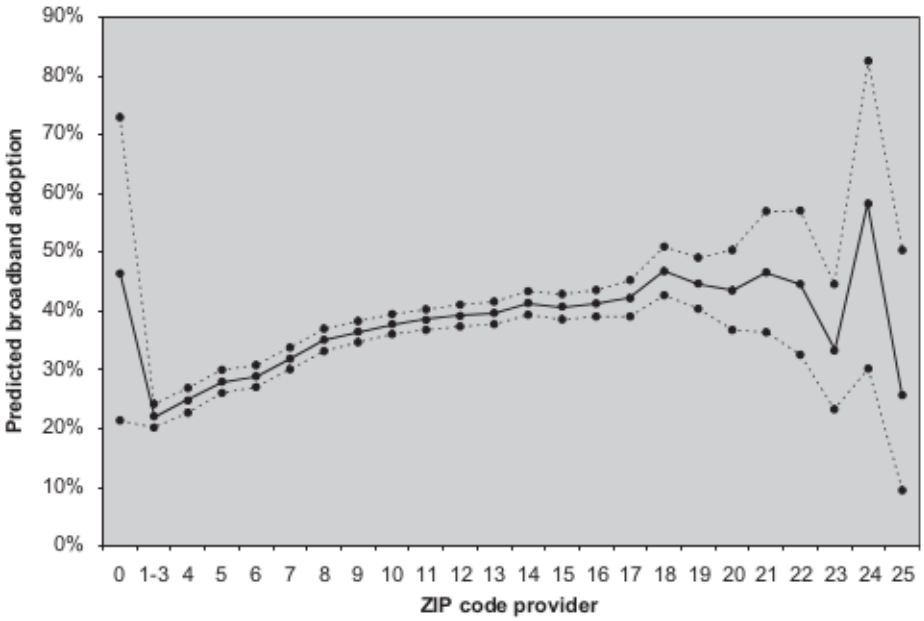

Figure 3.3: Predicted Broadband Adoption by Zip Provider Count. *Note.* Solid line represents predicted values, and dotted lines represent upper and lower bounds of 95% confidence interval. **Source: Kolko (2010).**

TABLE 3.2

OLS estimates of democrat vote share on ISP growth

	Democrat share 1996-2008					
	(1)	(2)	(3)	(4)	(5)	(6)
ISP providers	**0.066*** **	**0.058*** **	**0.045*** **	**0.044*** **	**0.032*** **	**0.006*** **
	(0.003)	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)
Democrat share: t-1						0.949***
						(0.004)
Avg. age		0.003***	0.004***	0.004***	0.003***	-0.001***
		(0.001)	(0.001)	(0.001)	(0.001)	(0.000)
% Male		-0.262***	-0.383***	-0.385***	-0.374***	-0.072***
		(0.069)	(0.073)	(0.073)	(0.072)	(0.015)
Log pop. density		0.066***	0.061***	0.063***	0.046***	0.003***
		(0.010)	(0.009)	(0.010)	(0.009)	(0.000)
% Black			0.404***	0.398***	0.356***	0.086***
			(0.041)	(0.042)	(0.044)	(0.004)
% Hispanic			0.074***	0.074***	0.039**	0.073***
			(0.018)	(0.018)	(0.019)	(0.006)
% Asian			0.688***	0.691***	0.433***	-0.100***
			(0.142)	(0.141)	(0.132)	(0.026)
% Poverty				0.011	-0.042	-0.005
				(0.036)	(0.036)	(0.018)
Log income				-0.014	-0.009	-0.011***
				(0.010)	(0.010)	(0.004)
% Less than HS					0.226***	-0.055***
					(0.039)	(0.011)
% Some college					-0.270***	-0.022**
					(0.043)	(0.011)
% College degree					0.110**	0.090***
					(0.044)	(0.008)
Fixed effects	Y	Y	Y	Y	Y	N
Lag. dep. Var.	N	N	N	N	N	Y
N	12227	12227	12227	12227	12227	12226
R^2	.656	.662	.675	.675	.687	.95

Notes. Specifications (1) through (5) report OLS estimates of equation 3.4.1 with fixed effects and state time trends. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. Column (6) adds a lagged dependent variable and drops fixed effects and time trends. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.3

OLS estimates of vote share, turnout, and donations on ISP growth

	% Democrat	Turnout	Total donats	Donats < $500	Dem donats	Dem donats < $500
	(1)	(2)	(3)	(4)	(5)	(6)
ISP providers	**0.032*** **	**0.024*** **	**1.067*** **	**0.176*** **	**0.065*** **	**0.034*** **
	(0.003)	**(0.003)**	**(0.222)**	**(0.047)**	**(0.023)**	**(0.009)**
Avg. age	0.003***	-0.003***	0.106**	0.044***	0.008**	0.002
	(0.001)	(0.001)	(0.049)	(0.012)	(0.004)	(0.001)
% Male	-0.374***	-0.910***	5.075*	0.516	0.440*	0.128
	(0.072)	(0.136)	(2.916)	(0.654)	(0.242)	(0.102)
Log pop. density	0.046***	-0.020	-2.231***	-0.552***	-0.187***	-0.079***
	(0.009)	(0.013)	(0.649)	(0.133)	(0.064)	(0.026)
% Black	0.356***	-0.308***	-7.370***	-1.618***	-0.863***	-0.320***
	(0.044)	(0.084)	(1.911)	(0.349)	(0.165)	(0.065)
% Hispanic	0.039**	-0.230***	-1.917*	-0.768***	-0.396***	-0.207***
	(0.019)	(0.046)	(1.132)	(0.205)	(0.102)	(0.045)
% Asian	0.433***	0.545***	45.238***	12.510***	7.331***	3.728***
	(0.132)	(0.204)	(13.828)	(2.836)	(1.582)	(0.790)
% Poverty	-0.042	0.054	3.860**	0.705*	0.113	0.066
	(0.036)	(0.034)	(1.530)	(0.364)	(0.163)	(0.080)
Log income	-0.009	0.001	4.583***	1.257***	0.444***	0.202***
	(0.010)	(0.010)	(0.798)	(0.185)	(0.104)	(0.046)
% Less than HS	0.226***	0.135***	14.270***	4.034***	1.410***	0.641***
	(0.039)	(0.039)	(1.958)	(0.495)	(0.226)	(0.097)
% Some college	-0.270***	-0.039	-15.526***	-3.049***	-1.826***	-0.779***
	(0.043)	(0.035)	(3.393)	(0.658)	(0.313)	(0.129)
% College degree	0.110**	0.192***	19.435***	5.438***	1.815***	0.790***
	(0.044)	(0.039)	(3.134)	(0.759)	(0.354)	(0.153)
N	12227	9171	12228	12228	12228	12228
R^2	.687	.763	.257	.441	.343	.4

Notes. The table reports OLS estimates of equation 3.4.1 with fixed effects and state time trends. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. All donation variables are expressed as dollars donated per capita: Total Donats, Donats<500, Dem Donats, Dem Donats<500. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.4

OLS estimates of all outcomes on ISP growth with lagged dependent variable 1996-2008

	% Democrat	Turnout	Total donats	Donats < $500	Dem donats	Dem donats < $500
	(1)	(2)	(3)	(4)	(5)	(6)
ISP providers	**0.006*** **	**-0.000**	**-0.011**	**0.009**	**0.013**	**0.012** **
	(0.002)	(0.002)	(0.198)	(0.031)	(0.013)	(0.006)
y_{t-1}	0.949***	0.812***	0.798***	0.930***	2.195***	2.043***
	(0.004)	(0.015)	(0.212)	(0.037)	(0.252)	(0.432)
Avg. age	-0.001***	0.001***	0.129***	0.029***	0.006***	0.003***
	(0.000)	(0.000)	(0.034)	(0.003)	(0.001)	(0.001)
% Male	-0.072***	-0.253***	-0.046	-0.086	0.102	0.025
	(0.015)	(0.041)	(1.057)	(0.238)	(0.083)	(0.037)
Log pop. density	0.003***	-0.003***	-0.037	-0.026***	-0.001	0.000
	(0.000)	(0.001)	(0.041)	(0.008)	(0.003)	(0.001)
% Black	0.086***	0.077***	0.112	-0.018	-0.022	-0.011
	(0.004)	(0.005)	(0.294)	(0.048)	(0.015)	(0.008)
% Hispanic	0.073***	-0.020***	0.075	0.083	-0.055**	-0.028**
	(0.006)	(0.006)	(0.539)	(0.061)	(0.023)	(0.012)
% Asian	-0.100***	-0.139***	1.921	0.790	0.745**	0.366**
	(0.026)	(0.024)	(3.556)	(0.547)	(0.339)	(0.177)
% Poverty	-0.005	-0.062**	4.967***	0.789**	0.184*	0.093*
	(0.018)	(0.030)	(1.317)	(0.309)	(0.106)	(0.052)
Log income	-0.011***	0.018***	2.079***	0.367***	0.043	0.020
	(0.004)	(0.006)	(0.486)	(0.072)	(0.028)	(0.014)
% Less than HS	-0.055***	-0.041***	4.129***	0.738***	0.187***	0.079**
	(0.011)	(0.016)	(1.259)	(0.208)	(0.063)	(0.036)
% Some college	-0.022**	0.001	-4.224***	-0.649***	-0.460***	-0.261***
	(0.011)	(0.015)	(1.634)	(0.215)	(0.089)	(0.044)
% College degree	0.090***	0.101***	11.905***	2.561***	0.727***	0.381***
	(0.008)	(0.010)	(3.134)	(0.198)	(0.082)	(0.049)
N	12226	6114	9171	9171	9171	9171
R^2	.95	.923	.567	.768	.744	.669

Notes. The table reports OLS estimates of equation 3.4.1 with a lagged dependent variable, and no fixed effects. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. y_{t-1} is the coefficient on the lagged dependent variable. All donation variables are expressed as dollars donated per capita: Total Donats, Donats<500, Dem Donats, Dem Donats<500. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.5

First stage relationship between providers and index of ROW laws 1996-2008

	Number of Internet providers				
	(1)	(2)	(3)	(4)	(5)
ROW index * $year_{2008}$	**0.178*** **	**0.193*** **	**0.188*** **	**0.190*** **	**0.157*** **
	(0.051)	(0.049)	(0.048)	(0.048)	(0.047)
ROW index * $year_{2004}$	**0.041**	**0.051**	**0.043**	**0.042**	**0.023**
	(0.036)	(0.034)	(0.035)	(0.036)	(0.035)
ROW index * $year_{2000}$	**0.015**	**0.017**	**0.008**	**0.014**	**0.008**
	(0.015)	(0.016)	(0.019)	(0.020)	(0.020)
Avg. age		0.003	0.007	0.007	0.001
		(0.007)	(0.007)	(0.007)	(0.007)
% Male		-1.746***	-2.366***	-2.338***	-1.520***
		(0.560)	(0.538)	(0.541)	(0.567)
Log pop. density		0.605***	0.570***	0.619***	0.426***
		(0.093)	(0.088)	(0.089)	(0.081)
% Black			2.569***	2.238***	1.790***
			(0.406)	(0.416)	(0.378)
% Hispanic			0.114	0.039	-0.157
			(0.236)	(0.247)	(0.257)
% Asian			9.070***	8.941***	7.151***
			(2.763)	(2.782)	(2.508)
% Poverty				0.376	0.146
				(0.280)	(0.273)
Log income				-0.392***	-0.389***
				(0.095)	(0.094)
% Less than HS					1.723***
					(0.364)
% Some college					-0.477
					(0.362)
% College degree					1.814***
					(0.436)
F-stat	6.14	6.14	6.14	6.14	6.14
N	2496	2496	2496	2496	2496
R^2	.905	.911	.917	.919	.924

Notes. The table presents OLS estimates of equation 3.4.3. It presents first stage results for the relationship between Internet diffusion and an index of right-of-way laws interacted with year dummies. The measure for Internet diffusion is the number of high speed Internet providers registered in a county. Row Index, is an index gauging the amenability of right-of-way laws in a state to Internet infrastructure investment. The sample has been limited to counties straddling state borders. See figure 3.1 for a map of counties in the sample. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.6

Reduced form relationship between democrat share and index of ROW laws 1996-2008

	Democratic vote share				
	(1)	(2)	(3)	(4)	(5)
ROW index * $year_{2008}$	0.067***	0.066***	0.063***	0.063***	0.058***
	(0.013)	(0.013)	(0.013)	(0.014)	(0.013)
ROW index * $year_{2004}$	0.041***	0.040***	0.036***	0.037***	0.034***
	(0.010)	(0.010)	(0.011)	(0.011)	(0.011)
ROW index * $year_{2000}$	0.016**	0.014**	0.011	0.012	0.011
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Avg. age		0.002	0.003	0.003	0.002
		(0.002)	(0.002)	(0.002)	(0.002)
% Male		-0.209	-0.329**	-0.326**	-0.186
		(0.156)	(0.155)	(0.155)	(0.158)
Log pop. density		0.019	0.016	0.019	-0.013
		(0.022)	(0.021)	(0.021)	(0.021)
% Black			0.541***	0.522***	0.457***
			(0.100)	(0.102)	(0.099)
% Hispanic			0.132*	0.129*	0.104
			(0.069)	(0.069)	(0.068)
% Asian			0.450	0.440	0.065
			(0.319)	(0.316)	(0.306)
% Poverty				-0.002	-0.036
				(0.080)	(0.078)
Log income				-0.026	-0.028
				(0.023)	(0.023)
% Less than HS					0.209**
					(0.088)
% Some college					-0.150
					(0.092)
% College degree					0.361***
					(0.098)
N	2504	2504	2504	2504	2504
R^2	.649	.65	.665	.665	.677

Notes. The table presents reduced form regressions of results for the relationship between Democratic candidate vote share in U.S. presidential elections and index of right-of-way laws interacted with year dummies. Row Index, is an index gauging the amenability of right-of-way laws in a state to Internet infrastructure investment. The sample has been limited to counties straddling state borders. See figure 3.1 for a map of counties in the sample. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.7

IV estimates of democrat share on Internet access 1996-2008

	Democrat vote share					
	(1)	(2)	(3)	(4)	(5)	(6)
ISP providers	**0.323*** **	**0.304*** **	**0.292*** **	**0.292*** **	**0.303*** **	**0.024*** **
	(0.082)	**(0.073)**	**(0.071)**	**(0.072)**	**(0.088)**	**(0.007)**
Avg. age		0.001	0.001	0.001	0.002	-0.000
		(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
% Male		0.327	0.363	0.358	0.283	-0.246*
		(0.202)	(0.228)	(0.227)	(0.209)	(0.128)
Log pop. density		-0.165***	-0.152***	-0.163***	-0.145***	0.019***
		(0.048)	(0.044)	(0.047)	(0.042)	(0.003)
% Black			-0.213	-0.134	-0.091	0.505***
			(0.200)	(0.183)	(0.184)	(0.026)
% Hispanic			0.116*	0.134**	0.171***	0.072
			(0.060)	(0.059)	(0.064)	(0.053)
% Asian			-2.189**	-2.161**	-2.094**	-0.136
			(0.933)	(0.937)	(0.940)	(0.229)
% Poverty				-0.111	-0.079	0.063
				(0.102)	(0.101)	(0.077)
Log income				0.089**	0.091*	-0.074***
				(0.042)	(0.047)	(0.020)
% Less than HS					-0.305	0.016
					(0.189)	(0.071)
% Some college					0.002	-0.241***
					(0.124)	(0.077)
% College degree					-0.173	0.117*
					(0.205)	(0.063)
N	2496	2496	2496	2496	2496	2496
R^2	.145	.248	.298	.306	.28	.687

Notes. Specifications (1) through (5) show results of IV regressions of Democratic candidate vote share in U.S. presidential elections on Internet diffusion from 1996-2008. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. The instrument is Row index, an index gauging the amenability of right-of-way laws in a state to Internet infrastructure investment, interacted with year dummies. The sample has been limited to counties straddling state borders. See figure 3.1 for a map of counties in the sample. Column (6) shows the OLS results of estimating equation 3.4.1 for the limited sample. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.8

IV estimates of other outcomes on Internet access 1996-2008

	% Democrat	Turnout	Total donats	Donats < $500	Dem donats	Dem donats < $500
	(1)	(2)	(3)	(4)	(5)	(6)
ISP providers	**0.303*** **	**0.064**	**5.004**	**1.051**	**0.530****	**0.192****
	(0.088)	(0.040)	(3.759)	(0.655)	(0.243)	(0.092)
Avg. age	0.002	-0.003	0.121	0.041**	-0.005	-0.004*
	(0.002)	(0.002)	(0.108)	(0.017)	(0.007)	(0.002)
% Male	0.283	-0.868***	16.396	1.977	1.224**	0.204
	(0.209)	(0.267)	(10.272)	(1.600)	(0.619)	(0.243)
Log pop. density	-0.145***	-0.083***	-6.097*	-0.874**	-0.466***	-0.105*
	(0.042)	(0.030)	(3.441)	(0.366)	(0.176)	(0.060)
% Black	-0.091	-0.375*	-17.008*	-3.382**	-1.657**	-0.534**
	(0.184)	(0.200)	(9.495)	(1.487)	(0.736)	(0.268)
% Hispanic	0.171***	-0.198*	-0.970	-0.795	-0.467	-0.179
	(0.064)	(0.104)	(4.339)	(0.616)	(0.402)	(0.150)
% Asian	-2.094**	0.273	-13.924	-0.940	-2.342	-0.459
	(0.940)	(0.575)	(41.460)	(6.016)	(3.202)	(1.227)
% Poverty	-0.079	0.220***	-0.641	-0.532	-0.462*	-0.132
	(0.101)	(0.078)	(3.508)	(0.648)	(0.262)	(0.112)
Log income	0.091*	0.042*	6.674**	1.155***	0.558**	0.146
	(0.047)	(0.025)	(3.126)	(0.391)	(0.260)	(0.099)
% Less than HS	-0.305	0.061	9.338	1.505	0.143	-0.101
	(0.189)	(0.104)	(6.023)	(1.420)	(0.441)	(0.178)
% Some college	0.002	-0.040	-1.147	-0.274	-0.668*	-0.321**
	(0.124)	(0.080)	(5.463)	(0.987)	(0.384)	(0.151)
% College degree	-0.173	0.236**	9.747	1.489	0.780	0.136
	(0.205)	(0.097)	(7.412)	(1.713)	(0.586)	(0.242)
N	2496	1872	2496	2496	2496	2496
R^2	.28	.758	.197	.317	.208	.325

Notes. The table presents instrumental variable estimates of equations 3.4.2 and 3.4.3. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. The instrument is Row Index, an index gauging the amenability of right-of-way laws in a state to Internet infrastructure investment, interacted with year dummies. The sample has been limited to counties straddling state borders. See figure 3.1 for a map of counties in the sample. All donation variables are expressed as dollars donated per capita: Total Donats, Donats<500, Dem Donats, Dem Donats<500. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.9

Reduced form estimates of democrat share on ROW index for previous years

	△Dem 88-92	△Dem 92-96	△Dem 96-00	△Dem 00-04	△Dem 04-08
	(1)	(2)	(3)	(4)	(5)
ROW index	-0.008	-0.003	0.018***	0.024***	0.019***
	(0.007)	(0.005)	(0.007)	(0.006)	(0.007)
N	624	624	624	624	624
R^2	.796	.623	.599	.642	.68

Notes. The table presents reduced form regressions of results for the relationship between Democratic candidate vote share in U.S. presidential elections and index of right-of-way laws interacted with year dummies. Specifications (1) and (2) look at the relationship during the pre-Internet election period. Specifications (3) to (5) cover the 1996-2008 period when rapid Internet growth coincided with presidential elections. Row Index, is an index gauging the amenability of right-of-way laws in a state to Internet infrastructure investment. The sample has been limited to counties straddling state borders. See figure 3.1 for a map of counties in the sample. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

TABLE 3.10

Placebo OLS estimates of change in democrat vote share on ISP growth

	△Dem 88-92	△Dem 92-96	△Dem 96-00	△Dem 00-04	△Dem 04-08
	(1)	(2)	(3)	(4)	(5)
ISP providers	0.008	-0.008			0.019***
2004-2008	(0.007)	(0.006)			(0.007)
ISP providers			0.004		
1996-2000			(0.020)		
ISP providers				-0.010	
2000-2004				(0.009)	
N	624	624	624	624	624
R^2	.79	.601	.592	.632	.679

Notes. The table reports OLS estimates of equation 3.4.1 with fixed effects and state time trends. Specification (1) regresses change in Democratic vote share for 1988-1992 on Internet growth for the 2004-2008 period. Similarly, specification (2) regresses change in Democratic vote share for 1992-1996 on Internet growth for the 2004-2008 period. Specifications (3) to (5) regress change in Democratic vote share on change in Internet in the same time period. ISP Providers, our proxy for Internet diffusion, is the number of high speed Internet providers registered in a county. See appendix for details on the construction and sources of variables. Coefficients are reported with robust standard errors in brackets. ***,**, and * indicate significance at the 1%, 5%, and 10% levels.

3.7 Data Appendix

Data Descriptions and Sources

Variable	Description	Source
Original Variables		
% Democrat	Share of the vote won by the presidential candidate from the Democratic Party.	uselectionatlas.org (1988, 1992, 1996, 2000, 2004, 2008)
Avg. age; % Male; Black; Asian; Hispanic	Respectively, average age; percentage male; percentage black; percentage asian; and percentage hispanic. Percentage white is omitted.	U.S. Census Bureau (1988, 1992, 1996, 2000, 2004, 2008)
% Poverty; Log income	Percentage of population below poverty line and log of income. For data not collected on exact year of election applied to closest election year.	U.S. Census Bureau (1989, 1993, 1996, 2000, 2004, 2008)
Log high-speed lines per capita	Log of high-speed lines per capita at the state level.	FCC (1988, 1992, 1996, 2000, 2004, 2008)
Generated Variables		
Log pop density	Log of County population divided by square kilomoters area.	U.S. Census Bureau (1988, 1992, 1996, 2000, 2004, 2008)
% Less than HS; Some college; % College degree	Respectively, Percentage of population with less than High School degree; percentage with some years of College; percentage with College degree. Percentage with High School degree is omitted. Some data not collected on election years. Average between 1980 and 1990 used to calculate 1988 control. Average between 1990 and 2000 used both for 1992 and 1996 control. Average between 2000 and 2006-2010 used for 2004.	U.S. Census Bureau (1980, 1990, 2000, 2006-2010)

Turnout	Number of votes cast divided by estimation of voting age population.	uselectionatlas.org (1988, 1992, 1996, 2000, 2004, 2008); U.S. Census Bureau (1988, 1992, 1996, 2000, 2004, 2008)
Total donats; Donats < $500; Dem donats; Dem donats < $500	Donation data aggregated from zip code to county level and then divided by county population. Respectively per capita total donations to both parties; per capita donations to both parties under $500; per capita donations to Democratic Party; per capita donations under $500 to Democratic Party.	FEC (1988, 1992, 1996, 2000, 2004, 2008)
ISP providers	Number of ISP providers registered to a county. Raw data is at the zip code level. Aggregated to county level via population weighted average. Value is divided by 10.	FCC (1988, 1992, 1996, 2000, 2004, 2008)
ROW index	Index measuring the relative favorability of right-of-way laws to Internet infrastructure investment. Index at state level and is out of one hundred. See Section 3.3.3 for details on laws.	TechNet (2002); Analysis Consulting (2002)

Bibliography

Abbott, J. (2011): "Electoral Authoritarianism and the Print Media in Malaysia: Measuring Political Bias and Analyzing Its Cause," *Asian Affairs: An American Review*, 38(1), 1–38.

Andersen, T., J. Bentzen, and C. Dalgaard (2011): "Does the Internet Reduce Corruption? Evidence from US States and across Countries," *The World Bank Economic Review*, 25(3), 387–417.

Banerjee, A. A., S. Kumar, R. Pande, and F. Su (2011): "Do Informed Voters Make Better Choices? Experimental Evidence from Urban India," .

Barham, T., M. Lipscomb, and A. Mobarak (2011): "Development Effects of Electrification: Evidence from the Geologic Placement of Hydropower Plants in Brazil," *CEPR Discussion Paper No. DP8427*.

Baron, D. P. (2006): "Persistent media bias," *Journal of Public Economics*, 1845(1845), 1–36.

Besley, T., and R. Burgess (2002): "The political economy of government responsiveness: Theory and evidence from India," *Quarterly Journal of Economics*, 117(4), 1415–1451.

Besley, T., and A. Prat (2006): "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, 96(3), 720–736.

Bhuller, M., T. Havnes, E. Leuven, and M. Mogstad (2011): "Broadband Internet: An Information Superhighway to Sex Crime?," *IZA Discussion Paper Series*, (5675).

Burchfield, M., H. G. Overman, D. Puga, and M. a. Turner (2006): "Causes of Sprawl: A Portrait from Space," *The Quarterly Journal of Economics*, 121(2), 587–633.

Case, W. (2001): "Malaysia's Resilient Pseudodemocracy," *Journal of Democracy*, 12(1), 43–57.

Centre for Independent Journalism Malaysia (2008): "Report on the Quantitative Analysis of the Media Monitoring Initiative for the 12th General Elections," pp. 1–41.

Danesh, a., L. Trajkovic, S. Rubin, and M. Smith (1999): "Mapping the Internet," in *Proceedings Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*, vol. 2, pp. 687–692. IEEE.

Day, C. R. (2002): "The Concrete Barrier at the End of the Information Superhighway : Why Lack of Local Rights-of-Way Access Is Killing Competitive Local Exchange Carriers," *Federal Communications Law Journal*, May, 461–492.

DellaVigna, S., and E. Kaplan (2007): "The fox news effect: Media bias and voting," *The Quarterly Journal of Economics*, 122(3), 1187–1234.

Dinkelman, T. (2011): "The Effects of Rural Electrification on Employment: New Evidence from South Africa," *American Economic Review*, 101(7), 3078–3108.

Dube, A., T. W. Lester, and M. Reich (2010): "Minimum wage effects across state borders: estimates using contiguous counties," *The Review of Economics and Statistics*, 92(4), 945–964.

Duflo, E., and R. Pande (2007): "Dams," *Quarterly Journal of Economics*, 122(2), 601–646.

Duranton, G., L. Gobillon, and H. G. Overman (2011): "Assessing the effects of local taxation using microgeographic data," *The Economic Journal*, 121(555), 1017–1046.

Edmond, C. (2011): "Information manipulation, coordination and regime change," *NBER Working Paper 17395*.

Falcone, M. (2008): "Youth Turnout Up by 2 Million From 2004," *The New York Times: The Caucus*, November.

Faloutsos, M., P. Faloutsos, and C. Faloutsos (1999): "On power-law relationships of the Internet topology," in *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication - SIGCOMM '99*, pp. 251–262, New York, New York, USA. ACM Press.

Fomenkov, M., and K. Claffy (2011): "Internet measurement data management challenges," in *Workshop on Research Data Lifecycle Management*, Princeton, NJ.

Gentzkow, M., and J. M. Shapiro (2011): "Ideological Segregation Online and Offline," *The Quarterly Journal of Economics*, 127(4), 1–49.

Golde, S. D., and N. H. Nie (2010): "The Effects of Online News on Political Behavior," *Stanford University*.

Goyal, A. (2010): "Information, Direct Access to Farmers, and Rural Market Performance in Central India," *American Economic Journal: Applied Economics*, 2(3), 22–45.

Hai, L. H. (2002): "Electoral Politics in Malaysia: 'managing' Elections in a Plural Society," in *Electoral Politics in Southeast and East Asia*, ed. by A. Croissant, G. Bruns, and M. John, pp. 101–148. Friedrich Ebert Stiftung, Singapore.

Heidemann, J., Y. Pradkin, R. Govindan, C. Papadopoulos, G. Bartlett, and J. Bannister (2008): "Census and survey of the visible internet," in *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement conference - IMC '08*, pp. 169–182, New York, New York, USA. ACM Press.

Hu, W.-m., and J. Prieger (2008): "Competition in Broadband Provision and the Digital Divide," in *Handbook of Research on Global Diffusion of Broadband Data Transmission, Vol. 1*, ed. by Y. Dwivedi, and E. al., no. October, pp. 241–259. IGI Global, Hershey, PA.

Huffaker, B., M. Fomenkov, and K. C. Claffy (2011): "Geocompare: a comparison of public and commercial geolocation databases," *2011 ISMA Workshop on Active Internet Measurements*.

Jack, W., and T. Suri (2011): "Risk Sharing and Transaction Costs: Evidence from Kenya's Mobile Money Revolution," *MIT*.

Jensen, R. (2007): "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector," *The Quarterly Journal of Economics*, 122(3), 879–924.

Keeter, S., J. Horowitz, and A. Tyson (2008): "Young Voters in the 2008 Election," *The Pew Research Center*, November.

Kende, M., and Analysys (2002): "The State Broadband Index: An Assessment of State Policies Impacting Broadband Deployment and Demand," *TechNet Report*.

Kolko, J. (2010): "A New Measure of US Residential Broadband Availability," *Telecommunications Policy*, 34(3), 132–143.

Kua, K. S. (ed.) (1990): *Mediawatch: the use and abuse of the Malaysian Press*. Resource & Research Centre, Selangor Chinese Assembly Hall, Kuala Lumpur.

Miller, C. (2008): "How Obamas internet campaign changed politics," *The New York Times*, November.

Miner, L. (2012): "The Unintended Consequences of Internet Diffusion Motivation : Evidence from Malaysia," *London School of Economics Working Paper*.

Mullainathan, S., and A. Shleifer (2005): "The Market for News," *American Economic Review*, 95(4), 1031–1053.

Naidu, S. (2012): "Suffrage, Schooling, and Sorting in the Post-Bellum U.S. South," *Columbia University*.

NARUC (2002): "Promoting Broadband Access Through Public Rights-of-Way and Public Lands," *Proceedings of NARUC Summer Meetings in Portland, Oregon*.

Nunn, N. (2008): "The Long-Term Effects of Africa's Slave Trades," *Quarterly Journal of Economics*, 123(1), 139–176.

Nunn, N., and D. Puga (2012): "Ruggedness: The Blessing of Bad Geography in Africa," *Review of Economics and Statistics*, 94(1), 20–36.

Olken, B. A. (2009): "Do Television and Radio Destroy Social Capital? Evidence from Indonesian Villages," *American Economic Journal: Applied Economics*, 1(4), 1–33.

Paluck, E. L. (2009): "Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda," *Journal of personality and social psychology*, 96(3), 574–87.

Pepinsky, T. B. (2007): "Malaysia: Turnover Without Change," *Journal of Democracy*, 18(1), 113–127.

Pischke, S., and J. Angrist (2010): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Reinikka, R., and J. Svensson (2004): "The power of information in public services: Evidence from education in Uganda," *Journal of Public Economics*, 95(July), 1–33.

Salazar, L. C. (2007): *Getting a Dial Tone*. Institute of Southeast Asian Studies, Singapore.

Schiffman, B. (2008): "The Reason for the Obama Victory: It's the Internet, Stupid," *Wired*, November.

Smith, A., and L. Rainie (2008): "The Internet and the 2008 Election," *Pew Research Center*, June.

Snyder, J. M., and D. Strömberg (2010): "Press Coverage and Political Accountability," *Journal of Political Economy*, 118(2), 355–408.

Stephens-Davidowitz, S. (2011): "The Effects of Racial Animus on Voting: Evidence Using Google Search Data," *Harvard University Working Paper*.

Stromberg, D. (2004): "Radio's Impact on Public Spending," *The Quarterly Journal of Economics*, 119(1), 189–221.

Talbot, D. (2008): "How Obama Really Did It," *Technology Review*, September.

Vargas, J. A. (2008): "Obama Raised Half a Billion Online," *The Washington Post*, Novermber.

Yanagizawa-Drott, D. (2012): "Propaganda and Conflict: Theory and Evidence From the Rwandan Genocide," *Harvard Kennedy School*.

Zinnbauer, D. (2003): "Power and Activism in the Context of a Maturing Internet : The Case of Malaysia," Ph.D. thesis, London School of Economics and Political Science.