The Effect of Internet on Political Mobilization*

Preliminary Version – Please Do Not Cite.

Klaus Ackermann† Simon D. Angus‡ Roland Hodler§
Paul A. Raschky¶

June 27, 2018

Abstract

This paper studies empirically the effect of the Internet on protests worldwide. We compile a novel panel dataset that combines geo-referenced data on Internet quality and weekly protests for over 18,907 subnational (ADM2) districts from 236 countries and the years 2006-2012. The Internet penetration data was constructed by combining over a trillion (1.5 X 10^{12}) IP activity (offline/online) observations to a commercially-available, IP-geolocation library. Our identification strategy exploits random weekly variation in global Internet latency to identify the causal effect of the Internet on local protests. According to our estimates, latency-adjusted Internet increases the occurrence of local protests. We show that most of the variation in the effect of the Internet on local protests comes from national differences in political institutions and local differences in Internet penetration.

*Roland Hodler and Paul Raschky gratefully acknowledge financial support from the Australian Research Council (ARC Discovery Grant DP150100061).

†Department of Econometrics and Business Statistics, Monash University; email: klaus.ackermann@monash.edu

‡Department of Economics, Monash University; email: simon.angus@monash.edu

§Department of Economics, University of St.Gallen; CEPR; CESifo; email: roland.hodler@unisg.ch.

¶Department of Economics, Monash University; email: paul.raschky@monash.edu.

1 Introduction

On the night of November 30th 2013, Ukrainian special police units (*Berkut*) brutally attacked and dispersed protesters at a rally at Kiev's Maidan square. Online videos and images of the police violence immediately went viral through Youtube and Facebook and were key in mobilizing people for Ukraine-wide protests and civil unrests that lasted until February 2014.[1] The "Euromaidan" protests ultimately resulted in the resignation of the government and the ousting of President Yanukovich.

This is just one of many anecdotes that highlights the Internet's potential as a "liberation technology."[2] The argument suggest that the Internet can foster political activism, participation and mass mobilization by decreasing transaction costs and increasing information and communication. It enables citizens to report and receive news, expose wrong doings of public officials and share information about economic grievances.[3] It can also help to coordinate and organize grass root movements and eventually help to mobilize supporters during protests and elections.[4]

However, the Internet can also be a bane for political mobilization. Similar to the introduction of TV (e.g., Gentzkow 2006), the Internet might crowd out other, more politically relevant, sources of information and foster citizens' disengagement with the political process. Similarly, the Internet can play a role in decreasing social capital accumulation which is often seen as a pre-requisite for mass mobilization by increasing weak social ties as opposed to strong social ties (e.g. Bond et al. 2012).

Despite the popularity of the liberation technology argument, only a few empirical studies exist that empirically examine the nexus between the Internet and political mobi-

[1] According to a survey (Kurylo and Dumova 2016) around 70 % of the participants in the Maidan protests named the viral videos as a reason for participation.

[2] For example, In 2011, a young women was beaten by Egyptian police during a protest on Cairo's Tahrir square. The video of her ordeal was circulated online and quickly broadcasted by large international news companies. In response, thousands of Egyptians marched into Tahrir Square later in December 2011 including the largest number of women in decades.

[3] One example is the violent death of Sun Zhigang in police custody in Guangzhou, China, in 2003. The news that Mr. Zhigang had been beaten to death in custody quickly spread though chat forums and soon after became a national story, forcing China's central government to launch an investigation that resulted in the conviction of 12 perpetrators (Diamond 2010).

[4] Prominent examples are Barack Obama's 2008 U.S. presidential election campaign, the Five Star Movement ("Movimento 5 Stelle") in Italy, or Emanual Macron's "La République En Marche!" in France.

lization. Enikolopov et al. (2016) find that an increase in the adoption of VK, a Russian social network platform, increased the incidence of local protests in Russia during 2011. Manacorda and Tesei (2016) combine yearly data of mobile phone adaptation rates and protests at a 55km × 55 km grid-cell level for the entirity of Africa. Their results show that higher mobile phone penetration can lead to more protests but only during times of economic downturns. Campante et al. (2017) study the effect of broadband roll-out on voter turnout in Italy. Initially, access to broadband Internet decreased voter turnout, but this effect was subsequently reversed by the emergence of political movements that used the Internet to attract and mobilize supporters. Finally, Falck et al. (2014) find a negative effect of broadband coverage on election turnout in Germany.

We complement the existing literature by analysing the effect of the Internet on local protests at a global scale and by employing Internet data at a high spatial and temporal resolution. In particular, we make use of a unique panel dataset that combines georeferenced data on Internet access and quality and weekly protests for 18,907 subnational ADM2 regions ('districts') from 236 countries and territories over the years 2006–2012. This high spatio-temporal resolution allows us to exploit within year and within subnational district variation and explore the potential heterogeneity of the effect along national and subnational differences.

We focus on latency-adjusted Internet activity on the local level. Latency relates to the travel time that a data package requires to get from its source point to its destination point. In the context of political mobilization, Internet latency is essential because it is strongly associated with the time it takes to upload or download a video or a photo, access particular sites or communicate via voice-over-IP. The basic idea is that Internet access by itself is a necessary but not sufficient condition for political mobilization. What is required is a sufficiently stable and fast local connection ("the entrance to the data highway") as well as a functioning global Internet backbone ("the global data highway itself") along which any data package travels.

Our identification strategy ultimately employs the random weekly variation in the latter component, the global Internet backbone. Our main explanatory variable is an interaction term between local (in the district) Internet activity and global, weekly Internet latency. In combination with the fine spatio-temporal resolution of the protest and Internet data, this enables causal identification. In practice, we specify a regression model that includes district × year fixed effects as well as week-of-the-year fixed effects. Thereby, we compare the effect of within-district-year variation in local Internet activity on within-district-year variation in weekly protests conditional on the level of weekly, global Internet latency. In other words, we examine if a district with high Internet activity is more likely to observe a protest in a week with good global Internet latency compared to a week with large distortions to the global Internet backbone.

This estimation strategy relies on two assumptions: First, large shocks to the global Internet backbone have to have a systematic impact on Internet speed and thereby people's interaction with the Internet. We present evidence that supports the argument that random damages to the regional fibre-optic cable-network had significant, temporary impact on regional Internet speed. We then show that human usage of the Internet decreases during periods of slow global Internet speed.

Second, global Internet latency needs to be orthogonal to local protest activity. As a support for this assumption we provide evidence that shocks to global latency are driven by damages to the physical infrastructure of the Internet (e.g., cuts of submarine cables) and basically unaffected by even the largest global Internet bandwidth intensity events during our sample period. In addition, we provide evidence that Internet latency is unaffected even by the largest live-video stream events and weeks with large media events.

In our empirical analysis, we examine the effect of the Internet on local protests in the very short-run. We find that latency-adjusted Internet activity systematically increases the occurrence of local protests in a given week.[5] The quality of a country's political institutions and local Internet penetration, measured by IPs per capita, seem to explain most of the heterogeneity of the effect.[6]

Our paper makes a number of contributions to the recent literature on the effect of the

[5] We present results based on monthly aggregates in Appendix C.

[6] 'IP' is the common short-hand for 'Internet Protocol address'. Any device with a connection to the internet is assigned a unique IP such that information packages can be sent and received, from or to, the device.

media on political mobilization. Among this literature is a strand that has investigated the effect of more traditional forms of media such as Radio (Stromberg 2004) and TV (e.g., Gentzkow 2006; DellaVigna and Kaplan 2007; Oberholzer-Gee and Waldfogel 2009; Campante et al. 2017) on voter turnout and election outcomes.

While more recent contributions focus explicitly on the role of the Internet on election outcomes (e.g., Campante et al. 2013; Falck et al. 2014) and political protests (e.g., Enikolopov et al. 2016; Manacorda and Tesei 2016), they typically rely on data from individual countries. A notable exception is the recent study by Manacorda and Tesei 2016 who use data from all African countries.

Our first contribution is that we take the analysis to a global level using data from 236 countries and territories with large variation in economic development and political institutions. Exploiting in particular the heterogeneity in the level of democracy, we show that the effect of the Internet on the incidence of local protests is more pronounced in less democratic countries. As such, this study provides the first globally generalizable, empirical support for the claim that the Internet is a liberation technology, helping to mobilize citizens in countries with otherwise limited options for political participation.

Second, we augment the existing literature by focusing on the mobilization effect of the Internet in the very short run. A few exceptions aside (e.g., Enikolopov 2016), most studies rely on cross-sectional or yearly variation in both the outcome variable (i.e. voter participation or protest incidence) and the Internet or ICT variables. To identify the effect, the majority of these studies have to rely on instrumental variables (IVs). In contrast, the fine temporal granularity of our Internet data allows us to estimate effects directly, and with a less stringent set of identifying assumptions.

Third, our data allows us to distinguish between the role of Internet activity *per se* and latency-adjusted Internet activity. Our findings imply that it is access to the Internet *in combination* with good quality/latency of the network that matters for political mobilization. As such, our results support the approach that is applied by many papers that use access to high-speed Internet for their identification strategy (e.g., Falck et al. 2014). Further, our findings also have important implications for the ongoing policy debate about the importance of net neutrality.[7] If Internet speed matters for political mobilization, any systematic discrimination in the handling of different Internet packages, could potentially have consequences for political participation and political outcomes.

The remainder of the paper is organized as follows: Section 2 presents the data on local protests, Internet activity and latency. Section 4 discusses the main empirical strategy and present the results using weekly data, while section 4 presents the results using monthly aggregates. Section 5 concludes.

2 Data

We construct an unbalanced panel dataset at the district-week level for the period 2006–2012. Our panel unit i are 18,907 subnational districts at the second federal level (ADM2) from 236 countries and territories c. ADM2 districts correspond to U.S. counties, Australian LGAs or districts in India etc. Our time unit t are weeks. We have compiled data for (almost) every week t in month m and year y starting from week 9 in 2006 until week 52 in 2012.

2.1 Protests

Our data on protests comes from the open-source GDELT 1.0 database. GDELT collects daily news event information from "the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages" (Leetaru and Schrodt 2013). News reports from other languages are translated into English through a collaboration with Google Ideas. Each news report is fed into a parsing algorithm that automatically extracts information about the time and location of the event as well as to classify it into categories and define the actors involved.

GDELT uses 20 main event classifications based on the CAMEO (Conflict and Mediation Event Observations) coding scheme. We use all events from the category "140:

[7] 'Net neutrality' refers to the existing principle, adopted by all carriers in all jurisdictions globally, that every packet routed on the Internet is given equal standing with respect to routing priority, regardless of origin, destination, or contents.

Protests", which defines events as follows: "All civilian demonstrations and other collective actions carried out as protests against the target actor not otherwise specified." Category 140 is further subdivded into sub-categories which allows us to distinguish between demonstrations, strikes and boycotts, and violent protests. Figure 1 plots the weekly number of protest events recorded in the GDELT database in our sample period.

Figure 1 about here

A main advantage of GDELT data is the fine temporal scale (daily) of the event data which allows us to construct monthly aggregates of protest events and exploit within-year and country variation of those events. Another advantage is that the GDELT data has been georeferenced using an ADM2 boundary shapefile. Other daily event datasets (e.g., ACLED or UDCP) are often limited in geographic coverage or do not contain information about protests.

One caveat of the GDELT data is that it is sourced from online news reports and relies on automatic coding. This means that the data coverage and data quality can vary over time. To account for global, seasonal, shocks in data coverage and quality we include week-of-the-year dummies. To control for country as well as district and year specific shocks in coverage and quality we also include a large vector of district-year dummies.

Our main outcome variable, $protest_{it}$, is a dummy variable that switches to one if a protest event was recorded in GDELT in week t and region i. We only consider an event a local protest if the precision code in the GDELT data suggests georeference precision at the subnational (rather than country) level and the country code of the event matches the country code of the ADM2 region.

2.2 Internet Data

To associate historical Internet activity and latency with a given ADM2 region and week, two datasets are required. First, an Internet *activity* database provides observations on the online/offline status of individual IPs which have been scanned (or 'probed') frequently over time. In this study, Internet activity data were provided by the University of Southern

California (USC) PREDICT internet-security database who conducted two types of IP scans: 'census' scans, where the universe of 2^{32} unique IP addresses were scanned over a period of time; and 'survey' scans, where 1% samples of IP addresses were scanned at high frequency (11 min) (Heidemann et al., 2008). In each case, the most basic package (a 'ping') is sent over the Internet to the target IP, effectively querying the target's online/offline status, with the query yielding a success indicator, and a return time for the query to complete.[8] In the case that an IP address is not online, or unreachable due to firewalls or other prohibitions, the nearest router or host will respond to indicate the result. Due to the vast scale of the Internet addressing space, and the frequency of scans, the Internet activity database used in this study comprised over 1.5 trillion observations.

On its own, such an Internet activity database provides highly granular observations on the activity and latency of individual IPs within the abstracted routing network (the Internet) only. To associate these observations with human localities a second dataset is required, namely, a historical geo-IP catalogue which provides a precise geo-location (lat/lon) for each IP, or IP-range. Since the geo-location of IPs (or IP-ranges) is not static over time,[9] the best geo-IP catalogues are updated bi-weekly to ensure maximum locational accuracy. For this purpose we used a commercial product, the *NetAcuity* database from Digital Envoy,[10]. This database has consistently been recognised as the most accurate geo-IP catalogue of its type (Gharaibeh et al., 2017), and provides a geo-location for the universe of IPs, updated on average every two weeks.

The mechanics of joining the activity and geo-location datasets are involved and are discussed in the Appendix. In summary, we associate an historically accurate geo-location from the geo-IP database to each online/offline IP observation in the IP activity database. From this intermediate we produce two datasets. First, for each month, the number of unique IPs that were found to be online at least once during the month were counted in each of the 40,000 ADM2 (subnational districts) boundaries. This formed the basis of our measure of Internet activity, IP_{im}, being the (log of) the number of unique IPs per 1,000

[8] Return times are typically in the order of 50 to 750 milliseconds.

[9] Internet service providers (ISPs) routinely purchase or release tranches of IPs as required for their customer needs, associating the IP to physical users by updating routing tables.

[10] See https://www.digitalenvoy.com/.

inhabitants for each district i and month m.

Figure 2 about here

Second, we utilise the ping return time (in milliseconds, ms) which is recorded with every online IP observation in the Internet activity database. The higher the ping return time the worse is the carriage speed of packets being sent and received by the given IP (i.e. the 'slower' the Internet for that IP). Figure 3 plots the weekly measurements of average global ping speed in our sample period from 2006 to 2012. We use the inverse of the average global ping return time to calculate our measure for global Internet latency in week t, $ping_t$. For an alternative specification we also calculate the inverse of the average weekly ping return time for each country c and week t, $ping_{ct}$.

Figure 3 about here

Two key factors are required for our identification strategy: first, that large shocks to the Internet backbone have a systematic impact on local Internet speed, and so, individual interaction with the Internet; and second, that variation in global latency is orthogonal to protest activity, or in other words, Internet activity associated with protests have no impact on the global backbone. We shall take each factor in turn.

Table 2 presents the summary statistics of our key variables.

Table 2 about here

2.3 Global Latency is Supply Driven

The most common source of major shocks to global latency are random breakages and technical faults to the Internet's physical infrastructure (overland and submarine fibre-optic cables.). For example cuts to submarine fibre-optic cables caused by ship anchors and trawling activity are well known in the telecommunications industry.[11] Other sources

[11] One industry source has estimated 'about 100' submarine cable cuts occur each year, see http://www.wired.co.uk/article/vulnerable-undersea-cables.

of shocks are due to natural disasters or accidental cuts.[12] Figure 6 shows the effect of three major cable cuts on regional ping speed, including: (a) the Taiwanese earthquake of 26 Dec 2006; (b) the series of submarine cable cuts off Egypt of 30 Jan – 4 Feb 2008; and (c) the Christmas Day submarine cable cut of 2011 off the Arab peninsula. In each case, the internet backbone disturbance is indicated by a vertical grey bar, with coloured lines given the average change in ping speed, by country, relative to the average over the series.

Figure 6 about here

Second, we show that, despite common perceptions, even spikes in simultaneous mass user demand for content over the internet has no impact on global, or regional, latency. To begin, it should be noted that the majority of Internet bandwidth consumed during our sample period is used for peer-to-peer (P2P) file sharing.[13] In 2007, over 70% of the global Internet traffic originated from P2P file sharing exchanges (Ipoque 2007).[14] Hence, large, concentrated events, where a vast number of people use bandwidth heavy Internet services, such as video streaming, are of relatively trivial scale when compared to this P2P background demand.

As an illustration, Figure 4 plots the weekly regional latency around two of the biggest live video stream events in our sample period: The 2011 Royal Wedding in Great Britain (panel a) and the finals of the 2012 League of Legends (an online computer game) Championships.[15] The periods when the event took place are highlighted by the gray vertical bars. In both events, we are unable to detect a significant decrease in the Internet latency of the countries expected to be most affected by the event. To place, for instance, the 2011 Royal Wedding in context, it is estimated that 1.3 Tbps of streamed content was

[12] One example is the infamous Hayastan Shakarian cut (Bachmann et al. 2016). Hayastan Shakarian, an elderly women from Georgia cut through an underground fibre-optic cable while digging to scavenge copper. This triggered an Internet blackout for more than three-and-a-half million people in Georgia as well as neighboring Armenia and Azerbaijan.

[13] For example BitTorrent.

[14] Among the files shared the largest fraction is made up by current movies, pornography, and music.

[15] Over its course, the live streams of the finals attracted an average of around 1.1 million concurrent viewers accumulating a total of 24,230,688 hours online streaming. https://www.polygon.com/2012/10/23/3542424/league-of-legends-final-attracts-8-2-million-viewers-season-2 .

demanded at the peak of the event which compares with around 1.1 Tbps during the inauguration of President Obama. Together, these two events are rated as the highest streaming pressure events of this period, but neither had any noticable impact on global, or regional, Internet latency.[16]

Figure 4 about here

As a further illustration, one might consider that large news events suddenly increases the Internet activity globally because people search simultaneously for news related information online, causing congestion. To analyse if there is a systematic impact between large news events and global Internet latency, we compared data on weekly news pressure (Stromberg 2004) with weekly fluctuations in global latency.[17] The scatter plot in Figure 5 shows that there is no systematic relationship and the simple correlation between those two measures is also very low (Pearson's rho: 0.072)

Figure 5 about here

Taken together, we have presented evidence that exogenous shocks to the Internet backbone have dramatic and measurable impacts on latency and that there is little evidence of any kind of congestion-effect even by the most bandwidth intensive Internet events that occured during our sample period. As such, we think it is reasonable to assume that shocks to global Internet speed are supply side driven and unaffected, even by large protest considered in this study.

2.4 Global Latency affects Human Internet Activity

In this section we provide some evidence that global latency affects human Internet activity. We accessed publicly available data by O'Neill et al. (2016) who compiled a unique dataset on gaming behaviour of over 108.7 million users of the Steam community over the

[16] See https://techcrunch.com/2017/01/23/trumps-inauguration-broke-live-video-streaming-records/ . As pointed out in the linked article, one should be careful comparing peak stream levels over distant years, given the huge expansion in bandwidth capacity and streaming quality (intensity of information per second) over time.

[17] The idea is that major news events in general increase news pressure. For more information on how this variable is constructed see Stromberg (2004).

years 2013 and 2014. Steam is one of the world's largest gaming communities and this dataset allows us to track a particular form of human Internet activity, online gaming, over time.

From this dataset, we use information about the total time played on online games (in minutes) over the past two weeks globally, and regress it on the average global ping speed over the past two weeks.

The results in Table 1 reveal that higher ping in the past two weeks systematically reduces the average time spend playing online games. We take this result as support for our claim that the global ping speed measured in our dataset systematically impacts human Internet activity.

Table 1 about here

This analysis supports our claim that shocks to the global ping speed, which is used in this study, have a systematic impact on people's use of the Internet. During our sample period of 2006–2012, we find that in weeks with relatively low Internet quality, individuals spend less time online.

3 Empirical Analysis

3.1 Baseline Results

We commence our analysis by presenting correlations between protest incidence in week t and district i and Internet activity in the corresponding month m and district i. In a first step we estimate the following baseline model:

$$protest_{it} = \alpha_i + \beta_1 IP_{i,m-1} + \mathbf{T}_t + \epsilon_{it} \tag{1}$$

where $protest_{it}$ is a dummy that switches to one if there is a protest in district i in month m and week t. IP is natural log of the number of active IPs per 1,000 inhabitants in district i and the previous month $m-1$. We use the first lag of Internet activity to avoid issues related to reversed causality.[18] α_i are district fixed effects which capture any time-invariant differences between districts that determine variation in Internet activity and protest likelihood. Finally, $\mathbf{T}_t$ is a vector of week-of-the-year fixed effects that accounts in a very flexible way for any seasonal variation in Internet activity and protests. After estimating the baseline specifications, including only week-of-the-year fixed effects as well as region and week-of-the-year fixed effects, we are going to estimate specifications with a more conservative set of fixed effects. First, we will augment equation (1) by a vector of country-year fixed effects to account for any shocks that are common to all districts in a given country and year. Second, we will replace district fixed effects by a vector of district-year fixed effects. This vector of fixed effects allows us to control for any unobservable shocks that are common to the weekly observations in each district i and year y.

We estimate equation (1) as a linear probability model (LPM) using OLS. The results are presented in columns 1–4 of Table 3. The estimates in columns (1) and (2) show that, controlling for week and region fixed effects, Internet activity in district i and month $m-1$ are positively and significantly correlated with protest incidence.

In column (3), we further include a vector of country-year fixed effects. The coefficient is still statistically significant but the effect is basically zero. Once we replace the district and country year fixed effects from column (3) by district-year fixed effects in column (4), the coefficient also loses statistical significance.

Table 3 about here

These baseline results reveal that, once we control for country-year specific as well as district-year specific shocks, monthly variation in Internet activity does not seem to be systematically correlated with protest activity in the very short run.

This could indicate that increased availability and usage of Internet by itself might not be sufficient condition to increase coordination and political mobilization, at least, in the very short run.

Therefore, in the next step, we expand our baseline specification and include the Internet latency as additional regressor:

[18] Using the contemporary values of Internet acitvity in month m yields very similar results.

$$protest_{it} = \alpha_i + \beta_1 IP_{i,m-1} + \beta_2 ping_{ct} + \mathbf{T}_t + \epsilon_{it} \tag{2}$$

In equation (2), $ping_{ct}$ denotes the inverse of the average ping speed in country c and week t. However, any correlation between the incidence of protest and weekly variation in a country's ping speed could be spuriously driven by other country wide events in a given week, such as a general government crack-down on opposition and wide-spread Internet censorship. In order to avoid this problem, we eventually replace average country ping speed $ping_{ct}$ by our preferred Internet latency variable, average global ping speed $ping_t$.

The results in columns (5) and (6) in Table 3 show that weekly improvement in country-wide and global Internet latency have are systematically positively correlated with protest occurrence.

We now combine the insights from the baseline analysis so far to specify a model that allows for a causal interpretation of the effect of the Internet on the incidence of protest in the very short-run. In particular, we exploit the exogenous variation in global Internet latency and construct a measure for local, latency-adjusted Internet activity. The basic idea is that, better Internet allows for the online dissemination of more information with better quality in a more timely manner. This improves short-run mobilization of protesters and, consequently, the incidence of protests.

To implement this, we specify our main model as follows:

$$protest_{it} = \lambda_{iy} + \beta_1 IP_{i,m-1} + \beta_2 ping_{ct} + \delta(IP_{i,m-1} \times ping_t) + \mathbf{T}_t + \epsilon_{it} \tag{3}$$

where λ_{iy} is a vector of district-year fixed effects and $ping_t$ is the inverse of the global average ping speed in week t. We will also be presenting results using the average country ping speed in week t, $ping_{ct}$, instead of $ping_t$. δ is the parameter of key interest, reflecting the effect of latency-adjusted Internet activity on the incidence of protests in district i and week t. Again, we estimate equation (3) with OLS and cluster the standard errors at the year-week level.[19]

Table 4 reports our main results. Column (1) presents the specifications using average

[19] The results are robust to alternative levels of clustering. See robustness check in Table 7

country ping speed. We find that latency adjusted Internet activity, $IP_{i,m-1} \times ping_{ct}$, has a positive and statistically significant effect on the incidence of protests in the same week.[20]

It is possible that the δ using $ping_{ct}$ is subject to issues of reversed causality and omitted variable bias. As such, in column (2), we replace $ping_{ct}$ by $ping_t$, the average global Internet latency in week t. The coefficient of latency adjusted Internet activity is again positive and statistically significant at the 1% level.

How large is the effect of latency adjusted Internet on protest incidence? As reported in Table 2 the probability of a protest in a given week in our sample is 0.09. A one standard-deviation increase in latency-adjusted IP activity, which compares to an increase in latency-adjusted IP activity in Port Philip (VIC) between late 2008 and early 2012, translates into an increase in the probability of protest to 0.14.

Column (3) uses Internet activity in the current month, IP_{im}, as opposed to the first lag, and the results stay qualitatively and quantitatively the same. Columns (4) to (7) use several other dependent variables. In column (4), we replace the protest incidence by the (log of) the number of protests and we a similar pattern to the main results in column (2). In columns (5)-(7), we split the protest variable into it's three sub-categories, demonstrations, strikes & boycotts, and violent protests. We find that our results are mainly driven by events labeled as demonstrations. The estimated coefficient for latency adjusted Internet activity in columns (6) and (7) are positive but not statistically significant and very small in magnitude.

These results are intutive given that strikes are the result of some longer planning and do not necessarily rely on short-run mobilization via the Internet. The eruption of violence during a protest are mainly the result of ad-hoc dynamics during the protest rather than short-term online mobilization.

Table 4 about here

[20] Note that the number of observations in column (1) is smaller than in the other columns. The reason for this difference is that for some country-week observations, there was not sufficient IP scan data available to calculate average country ping speed.

3.2 Robustness Tests

We now present a number of exercises to show the robustness of our main specification in column (2) of Table 4. In Table 5 column (1) we include the lagged dependent variable as an additional control variable. In column (2) we exclude all observations that had only 1 recording of a protest event in the GDELT database. The idea is to check how sensitive the results are to the exclusion of marginal protest events, that only received one news report in total. The coefficient slightly drops in magnitude and is still highly statistically significant. On the one hand this shows that the short-run effect of the Internet on protest incidence is stronger for the marginal protest event.[21] On the other hand, this robustness check provides further support for our argument that our analysis estimates a mobilization rather than a news reporting effect. Recall, that our dependent variable is constructed from the GDELT archive that collects news reports about protest events. If there is indeed a reporting effect, this one should mainly occur for small, marginal protests (with only one entry in the GDELT database).

In column (3) we exclude all district and week observations with more than 10 protest event recordings in the GDELT database to check if our results are robust to the exclusion of very large events and potential outliers. The estimated coefficient remains similar to the coefficient estimate in our main specification, again supporting the idea that Internet increases the short-run mobilization for relatively smaller protest events.

In column (4) we exclude all observations with zero IP activity in a given district and week. The results suggest that our main findings are not driven by districts and weeks that switch from zero to positive IP activity (e.g., districts that have received access to the Internet recently). Finally, we exclude all observations with more than 100 unique IPs per 1,000 inhabitants. Again, our results are robust.

Table 5 about here

In the next step, we investigate the dynamics around global changes in ping speed. We do that by estimating specifications including lagged values *ping*, in particular the

[21] Our analysis basically estimates the effect on the extensive margin (occurrence of protests) rather than the intensive margin (size of the protests)

global ping speed in the previous week, $ping_{t-1}$ as well as leads, which is the ping speed in the following week, $ping_{t+1}$.

A specification using deeper lags would enable us to learn more about the temporal extent of the mobilization effect. In contrast, global ping speed in the following week[22] should have no systematic effect on the incidence of protest. A statistically significant and positive coefficient would indicate that we capture a reporting rather than a mobilization effect.

Table 6 about here

In Table 7 we present specifications with alternative levels of clustering. In particular, we cluster at the country-year (Column 1), country (2), district-year (3), district level (4). Changing the cluster units does not effect our results.

Table 7 about here

3.3 Heterogeneity

We now want to uncover some of the heterogeneity of the effect. In a first step, we exploit the global coverage of our dataset and investigate whether there are systematic differences of the effect of Internet on short-run mobilization.

In Table 8 we add various interaction terms between our main explanatory variable and dummy variables that assign countries to different country groups depending on their level of development, degree of Internet censorship, and geographic location.[23]

Table 8 about here

In Table 9 we look a the heterogeneity of the Internet effect along a number of national indicators. In particular, we build interaction terms between our main variable of interest, $IP_{i,m-1} \times ping_{ct}$, and the quality of political institutions *polity*, dummy variables for

[22] Controlling for global ping speed in the current week to account for potential autocorrelation in global ping speed)

[23] Censor countries include: China, Iran, India, Iraq, South Korea, North Korea, Russia, Pakistan, Vietnam, UAR, Tajikistan, Uzbekistan, Turkey, Egypt, Syria.

democracies (*Democ*), anocracies (*Anoc*), and dictatorships (*Dict*), and indicator if the country has above median GDP per capita (*HighGDP*), and if the country is currently in a recession (*Recession*), respectively.

The results in Table 9 reveal the following pattern: Overall, the effect of latency adjusted internet activity on short-run mobilization is stronger with weaker political institutions (columns 1, 5, and 6), but this effect seems to be mainly driven by dictatorial regimes and less so by anocracies (column 2). The effect is more pronounced in higher income countries (columns 3 and 4) and weaker if the country currently experiences a recession (columns 4 and 6).

Table 9 about here

Finally, in Table 10 we examine the heterogeneity of the Internet effect along a number of subnational indicators. We build interaction terms between our main variable of interest, $IP_{i,m-1} \times ping_{ct}$, two sets of measures which vary at the district and year level. The first one is the nighttime light intensity in district i and year y, $HiLight_{iy}$ (if light is above median light) and $Light_{iy} > 58$ (if light is above 58). Nighttime light intensity is used as a proxy for district level economic development given the absence of any other official measure of district level economic development with a global coverage. The second set of measures are dummy variables to indicate if the Internet penetration in the district is above a certain threshold.

Table 10 about here

4 Conclusion

We study the effect of the Internet on local protests using data at the district-week level. We find that latency-adjusted Internet activity increases the occurrence of local protests. The effect is stronger in more democratic countries and ADM2 regions with higher local Internet penetration. The information effect is strongest in the 3 months prior to the protest. We also find some preliminary evidence of a crowding-out effect in the information channel.

References

Ackermann, Klaus, Simon D. Angus and Paul A. Raschky (2017), "The Internet as Quantitative Social Science Platform: Insights from a Trillion Observations," arXiv:1701.05632.

Bachmann, Ivana, Patricio Reyes, Alonso Silva, and Javier Bustos-Jiménez (2016), "The Hayastan Shakarian Cut: Measuring the Impact of Network Disconnections," arXiv:1601.02465.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D.I. Kramer, Cameron Marlow, Jaime E. Settle and James H. Fowler (2012), "A 61-Million-Person Experiment in Social Influence and Political Mobilization," *Nature* 489(7415), 295–298.

Campante, Filipe, Ruben Durante and Francesco Sobbrio (2017), "Politics 2.0: The Multifaceted Effect of Broadband Internet on Political Participation," *Journal of the European Economic Association*, forthcoming.

DellaVigna, Stefano, and Ethan Kaplan (2007), "The Fox News Effect: Media Bias and Voting," *Quarterly Journal of Economics* 122(3), 1187–1234

Diamond, Larry (2010), "Liberation Technology," *Journal of Democracy*, 21(3), 69–83.

Eisensee, Thomas, and David Strömberg (2007), "News Droughts, News Floods, and U.S. Disaster Relief," *Quarterly Journal of Economics* 122(2), 693–728.

Enikolopov, Ruben, Aleksey Makarin and Maria Petrova (2016), "Social Media and Protest Participation: Evidence from Russia," CEPR Discussion Paper 11254.

Falck, Oliver, Robert Gold, and Stephan Heblich (2014), "E-lections: Voting Behavior and the Internet," *American Economic Review* 104(7), 2238–2265.

Gentzkow, Matthew (2006), "Television and Voter Turnout," *Quarterly Journal of Economics* 121(3), 931–972.

Gharaibeh, Manaf, Anant Shah, Bradley Huffaker, Han Zhang, Roya Ensafi and Christos Papadopoulos (2017), "A Look at Router Geolocation in Public and Commercial Databases", *Proceedings of the 2017 Internet Measurement Conference*, 463–469.

Heidemann, John, Yuri Pradkin, Ramesh Govindan, Christos Papadopoulos, Genevieve Bartlett and Joseph Bannister (2008), "Census and Survey of the Visible Internet," *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, 169–182.

ipoque (2007), *Internet Study 2007*, Leipzig, Germany.

Kurylo, Anastacia and Tatyana Dumova (2016), *Social Networking: Redefining Communication in the Digital Age*, Fairleigh Dickinson University Press.

Leetaru, Kalev, and Philip A. Schrodt (2013), *GDELT: Global data on events, location, and tone, 1979–2012*, ISA Annual Convention 2(4).

Manacorda, Marco, and Andrea Tesei (2016), "Liberation Technology: Mobile Phones and Political Mobilization in Africa," CEPR Discussion Papers 1419.

Oberholzer-Gee, Felix, and Joel Waldfogel (2009), "Media Markets and Localism: Does Local News en Español Boost Hispanic Voter Turnout?" *American Economic Review*, 99(5), 2120–2128.

O'Neill, Mark, Elham Vaziripour, Justin Wu and Daniel Zappala (2016), "Condensing Steam: Distilling the Diversity of Gamer Behavior," *Proceedings of the 2016 Internet Measurement Conference*, 81–95.

Strömberg, David (2007), "Mass Media Competition, Political Competition, and Public Policy," *Review of Economic Studies*, 71(1), 265–284.

Figures and Tables

Figure 1: Weekly number of protests events recorded in GDELT, 2006–2012

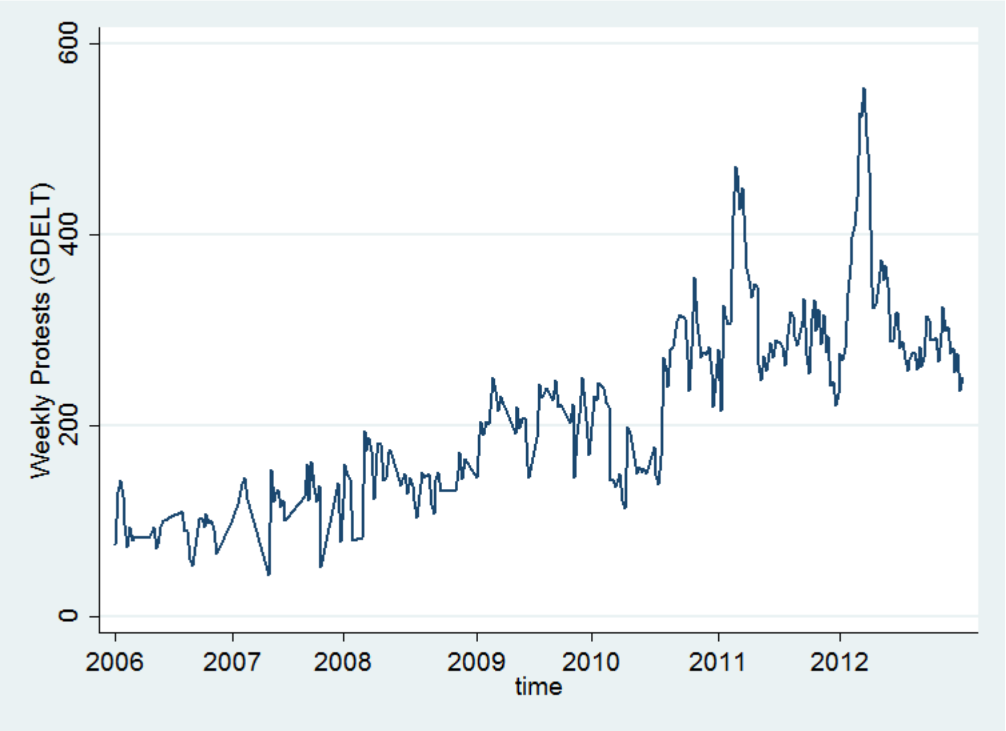


Figure 2: Global unique IPs, 2012

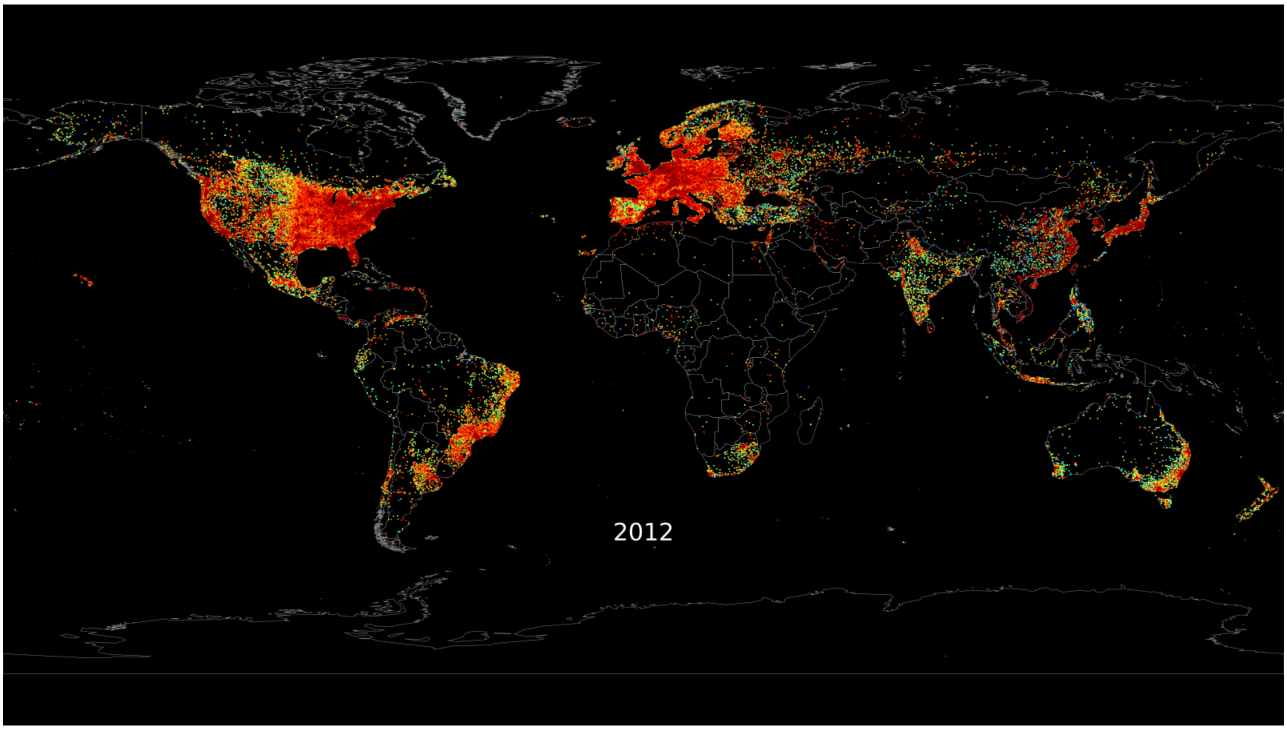


Figure 3: Weekly Variation in Global Ping, 2006–2012

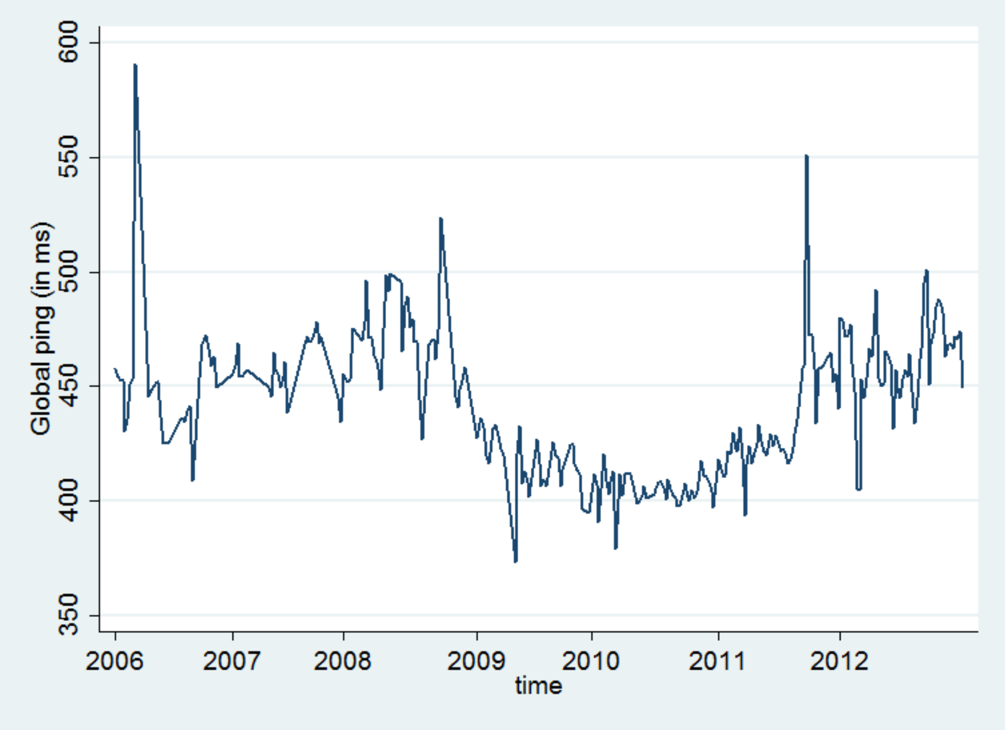
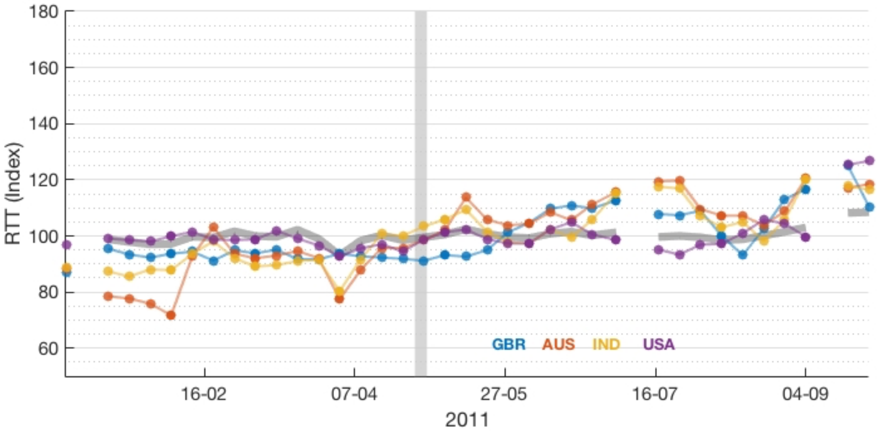


Figure 4: The Effect of the Large Stream Events on Global Ping

a) Royal Wedding 2011



b) League of Legends Championships 2012

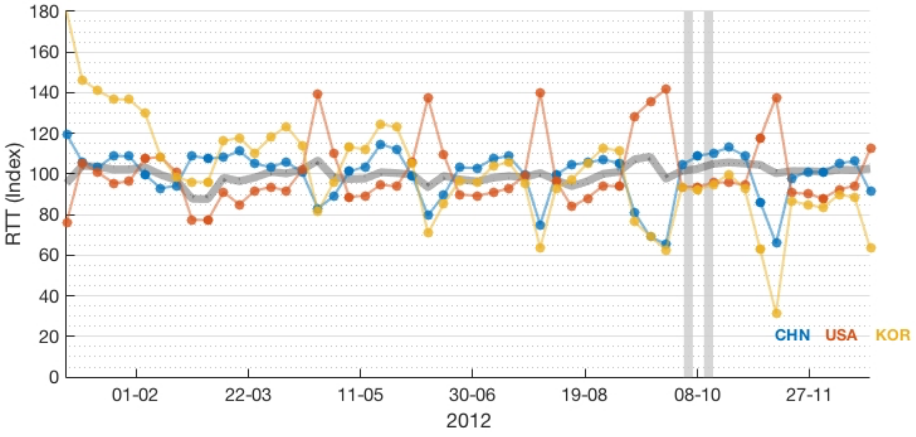


Figure 5: News Pressure and Global Ping

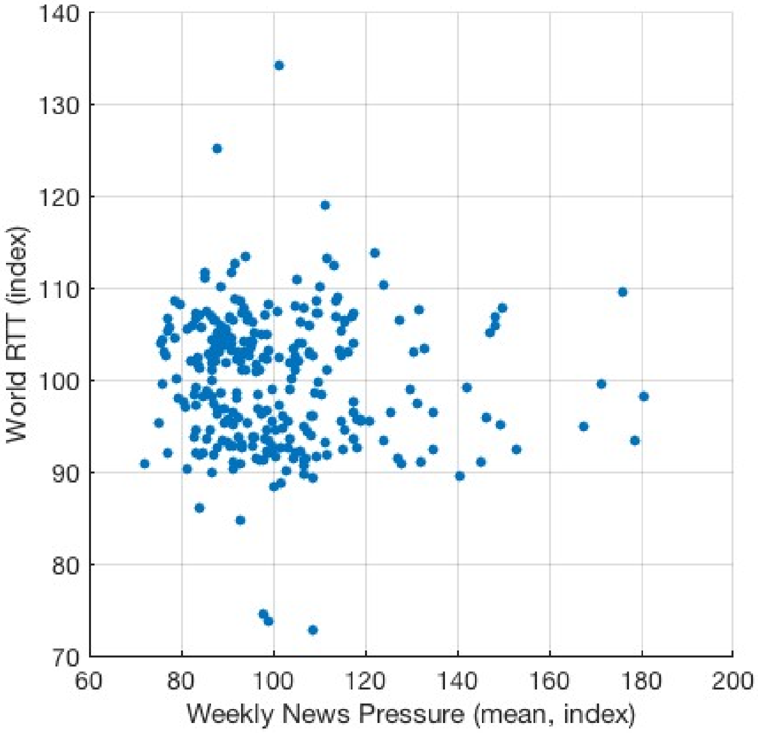
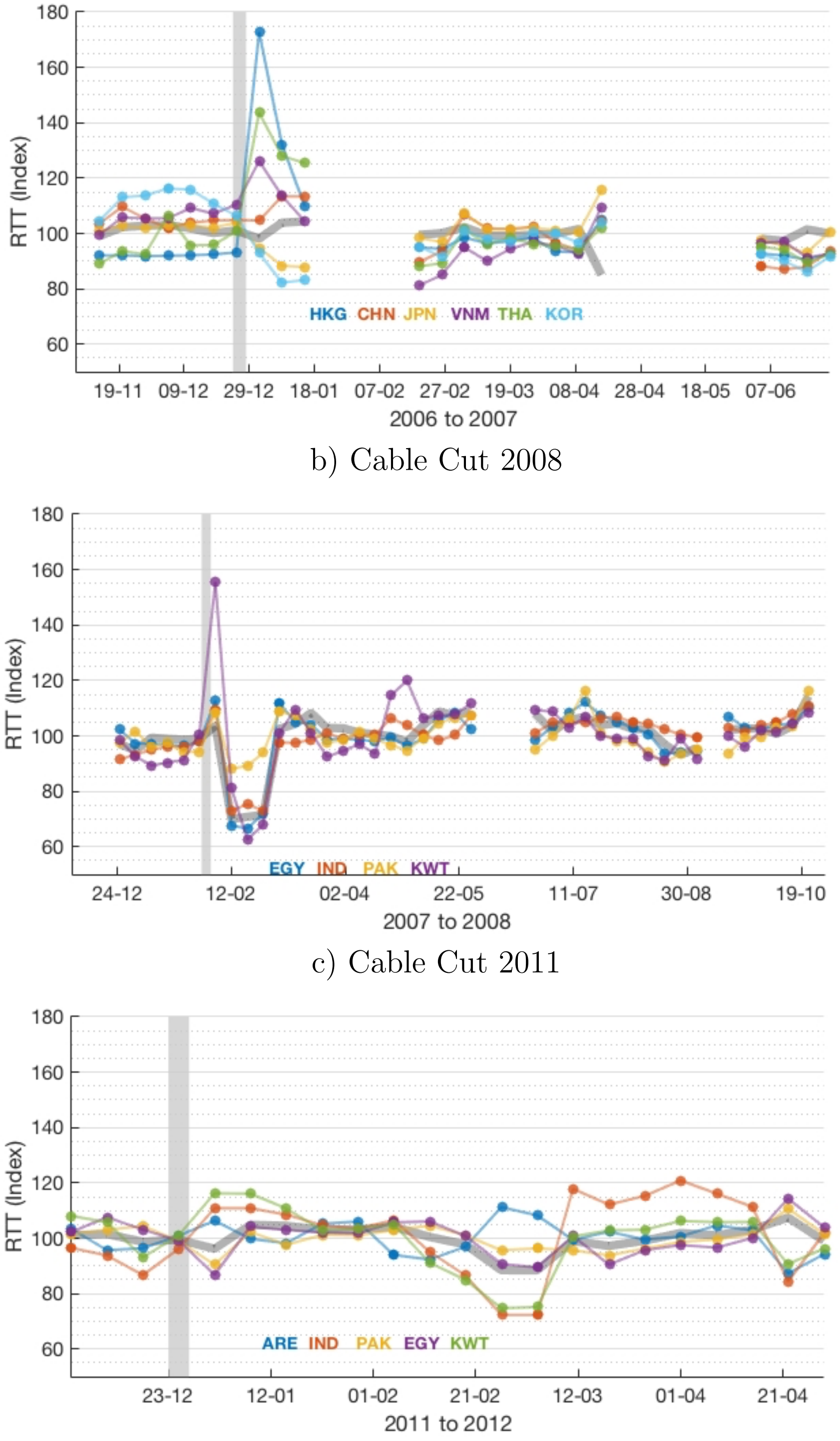


Figure 6: The Effect of Cable Cuts on Regional Ping



a) Taiwanese Earthquake 2006

b) Cable Cut 2008

c) Cable Cut 2011

Table 1: Internet Latency and Global Online Gaming Behaviour - May-Nov 2013

	(1)	(2)	(3)	(4)
	Playtime past 2 weeks		ln(Playtime past 2 wks)	
Average ping past 2 weeks	-9.3784**	-8.7967**		
	(3.6452)	(3.6008)		
ln(average ping past 2 weeks)			-0.5140**	-0.5066**
			(0.2083)	(0.2009)
Day of the Week FE	No	Yes	No	Yes
N	171	171	171	171

Notes: All specifications include month fixed effects and quadratic time trends. *** (**, *): significant at the one (five, ten) percent level. Robust Standard errors in parenthesis.

Table 2: Descriptive Statistics

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
$protest_{iymt}$	5,464,123	0.023	0.151	0.000	1.000
$\#protest_{iymt}$	5,464,123	0.131	3.037	0.000	2421.000
$Demo$	5,464,123	0.108	2.490	0.000	1849.000
$Strike$	5,464,123	0.007	0.196	0.000	54.000
$Violent$	5,464,123	0.011	0.529	0.000	511.000
$Events\ abroad$	5,464,123	0.001	0.113	0.000	174.000
$IP_{i,m-1}$	5,464,123	-1.135	3.832	-4.605	11.936
$ping_{cymt}$	5,450,164	0.040	0.023	0.000	0.296
$ping_{ymt}$	5,464,123	0.228	0.017	0.169	0.312

Table 3: Correlations

Protest	(1)	(2)	(3)	(4)	(5)	(6)
$IP_{i,m-1}$	0.0041***	0.0019***	-0.0001***	0.0000	0.0000	0.0001*
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$ping_{ct}$					0.0735***	
					(0.0055)	
$ping_t$						0.0665***
						(0.0042)
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	No	Yes	Yes	No	No	No
Country-Year FE	No	No	Yes	No	No	No
Region-Year FE	No	No	No	Yes	Yes	Yes
N	5464123	5464123	5464123	5464123	5450164	5464123

Notes: Robust standard errors in parenthesis. *** (**, *): significant at the one (five, ten) percent level.

Table 4: Main Results

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
		Protest		# Protests	Demo	Strike	Violent
$IP_{i,m-1}$	-0.0007***	-0.0062***		-0.0339***	-0.0058***	-0.0003	-0.0005
	(0.0001)	(0.0017)		(0.0097)	(0.0016)	(0.0004)	(0.0003)
$IP_{i,m-1} \times ping_{ct}$	0.0152***						
	(0.0031)						
$ping_{ct}$	0.0623***						
	(0.0220)						
$IP_{i,m-1} \times ping_{t}$		0.0279***		0.1526***	0.0264***	0.0011	0.0023
		(0.0077)		(0.0436)	(0.0073)	(0.0017)	(0.0015)
$ping_{t}$		0.1436***	0.1372***	0.7937***	0.1305***	0.0107	0.0183***
		(0.0353)	(0.0358)	(0.1994)	(0.0331)	(0.0076)	(0.0063)
IP_{im}			-0.0056***				
			(0.0018)				
$IP_{im} \times ping_{t}$			0.0257***				
			(0.0079)				
N	5450164	5464123	5464123	5464123	5464123	5464123	5464123

Notes: All specifications include district-year and week-of-the-year fixed effects except for column (1) which only includes district-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering.

Table 5: Robustness Tests 1

Protest	(1)	(2) Protests>1	(3) Protests<10	(4) $IP_{i,m-1}$ >0	(5) $IP_{i,m-1}$ <100
$Protest_{i,t-1}$	0.0566***				
	(0.0020)				
$IP_{i,m-1}$	-0.0066***	-0.0037***	-0.0061***	-0.0062***	-0.0042***
	(0.0018)	(0.0014)	(0.0017)	(0.0023)	(0.0011)
$IP_{i,m-1} \times ping_{ymt}$	0.0300***	0.0167***	0.0277***	0.0285***	0.0188***
	(0.0079)	(0.0064)	(0.0076)	(0.0101)	(0.0051)
$ping_t$	0.1485***	0.0900***	0.1413***	0.1472***	0.1044***
	(0.0359)	(0.0282)	(0.0346)	(0.0408)	(0.0245)
N	4934727	5414712	5449365	3071177	4997918

Notes: All specifications include district-year and week-of-the-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering.

Table 6: Robustness Tests 2 - Lags and Leads

Protest	(1)	(2)	(3)	(4)	(5)
$IP_{i,m-1}$	-0.0065***	-0.0063***	-0.0086***	-0.0088***	-0.0104***
	(0.0017)	(0.0019)	(0.0019)	(0.0019)	(0.0021)
$IP_{i,m-1} \times ping_{t-1}$	0.0300***		0.0203**		0.0125
	(0.0076)		(0.0085)		(0.0094)
$ping_{t-1}$	0.1669***		0.1217***		0.0938**
	(0.0344)		(0.0435)		(0.0459)
$IP_{i,m-1} \times ping_{t+1}$		0.0285***		0.0160	0.0133
		(0.0085)		(0.0123)	(0.0133)
$ping_{t+1}$		0.1492***		0.0799*	0.0741
		(0.0333)		(0.0476)	(0.0512)
$IP_{i,m-1} \times ping_t$			0.0190**	0.0235*	0.0216
			(0.0089)	(0.0133)	(0.0136)
$ping_t$			0.0828*	0.1227**	0.0802
			(0.0440)	(0.0558)	(0.0602)
N	4934727	4934727	4934727	4934727	4405331

Notes: All specifications include district-year and week-of-the-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering.

Table 7: Robustness Tests 3 - Clustering

Protest	(1)	(2)	(3)	(4)
$IP_{i,m-1}$	-0.0062***	-0.0062***	-0.0062***	-0.0062***
	(0.0024)	(0.0021)	(0.0004)	(0.0004)
$IP_{i,m-1} \times ping_{ymt}$	0.0279**	0.0279***	0.0279***	0.0279***
	(0.0111)	(0.0101)	(0.0017)	(0.0017)
$ping_t$	0.1436***	0.1436***	0.1436***	0.1436***
	(0.0464)	(0.0480)	(0.0073)	(0.0075)
Cluster Unit	cy	c	iy	i
N	5464123	5464123	5464123	5464123

Notes: All specifications include ADM2-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering.

Table 8: Heterogeneity by Country Groups

Protest	(1)	(2)	(3)	(4)
$IP_{i,m-1}$	-0.0235***	-0.0344***	-0.0110	-0.0409***
	(0.0073)	(0.0100)	(0.0258)	(0.0119)
$IP_{i,m-1} \times ping_t$	0.0995***	0.1550***	0.0731	0.1859***
	(0.0330)	(0.0449)	(0.0233)	(0.0536)
$IP_{i,m-1} \times ping_t$	0.0729*			
$\times OECD_c$	(0.0425)			
$IP_{i,m-1} \times ping_t$		0.0263		
$\times Censor_c$		(0.0521)		
$IP_{i,m-1} \times ping_t$			0.0408	
$\times Africa_c$			(0.1983)	
$IP_{i,m-1} \times ping_t$			0.0788	
$\times Asia_c$			(0.1202)	
$IP_{i,m-1} \times ping_t$			-0.0432	
$\times Europe_c$			(0.1143)	
$IP_{i,m-1} \times ping_t$			0.1819	
$\times Americas_c$			(0.1431)	
$IP_{i,m-1} \times ping_t$			0.0741	
$\times Oceania_c$			(0.1242)	
$IP_{i,m-1} \times ping_t$				-0.1643***
$\times ECA_c$				(0.0476)
$IP_{i,m-1} \times ping_t$				-0.0396
$\times EAP_c$				(0.0450)
$IP_{i,m-1} \times ping_t$				0.2022
$\times MENA_c$				(0.2442)
$IP_{i,m-1} \times ping_t$				0.7062*
$\times SA_c$				(0.3910)
$IP_{i,m-1} \times ping_t$				-0.1364
$\times SSA_c$				(0.1754)
$ping_t$	0.6300***	0.7399***	0.4377	0.8480***
	(0.1683)	(0.2014)	(0.5262)	(0.2330)
N	5464123	5464123	5419039	5419039

Notes: All specifications include ADM2-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering. Censor countries include: China, Iran, India, Iraq, South Korea, North Korea, Russia, Pakistan, Vietnam, UAR, Tajikistan, Uzbekistan, Turkey, Egypt, Syria. Other interaction terms between country group variables and $ping_{ymt}$ as well as country group variables and IP_{im-1} included in specifications but not reported.

Table 9: Heterogeneity by National Indicators

Protest	(1)	(2)	(3)	(4)	(5)	(6)
$IP_{i,m-1}$	-0.0122***	-0.0006**	-0.0045***	-0.0084***	-0.0128***	-0.0115***
	(0.0039)	(0.0002)	(0.0016)	(0.0023)	(0.0039)	(0.0039)
$IP_{i,m-1} \times ping_t$	0.0499***		0.0197***	0.0379***	0.0527***	0.0465***
	(0.0170)		(0.0070)	(0.0104)	(0.0173)	(0.0172)
$IP_{i,m-1} \times ping_t$	-0.0231				-0.0437**	-0.0089
$\times Polity_{cy}$	(0.0178)				(0.0192)	(0.0202)
$IP_{i,m-1} \times ping_t$		0.0274***				
$\times Democ_{cy}$		(0.0079)				
$IP_{i,m-1} \times ping_t$		0.0417				
$\times Anoc_{cy}$		(0.0288)				
$IP_{i,m-1} \times ping_t$		0.0777***				
$\times Dict_{cy}$		(0.0203)				
$IP_{i,m-1} \times ping_t$			0.0109		0.0204**	
$\times HighGDP_c$			(0.0090)		(0.0097)	
$IP_{im-1} \times ping_t$				-0.0373***		-0.0363***
$\times Recession_{cy}$				(0.0112)		(0.0118)
$ping_{ymt}$	0.1979**	0.0460**	0.1326***	0.1858***	0.2106***	0.1895**
	(0.0771)	(0.0218)	(0.0361)	(0.0468)	(0.0772)	(0.0775)
N	5352921	5464123	5321502	5464123	5244981	5352921

Notes: All specifications include ADM2-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering. Other interaction terms between national indicator variables and $ping_{ymt}$ as well as national indicator variables and IP_{im-1} included in specifications but not reported.

Table 10: Heterogeneity by Subnational and National Indicators

Protest	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$IP_{i,m-1}$	-0.0044***	-0.0062***	-0.0029***	-0.0038***	-0.0021**	-0.0030***	-0.0123***	-0.0135***
	(0.0013)	(0.0017)	(0.0010)	(0.0011)	(0.0008)	(0.0010)	(0.0039)	(0.0039)
$IP_{i,m-1} \times ping_t$	0.0203***	0.0279***	0.0136***	0.0174***	0.0102***	0.0139***	0.0505***	0.0555***
	(0.0056)	(0.0077)	(0.0043)	(0.0049)	(0.0036)	(0.0043)	(0.0170)	(0.0172)
$IP_{i,m-1} \times ping_t$	0.0104**				0.0047		0.0046	
$\times HiLight_{iy}$	(0.0051)				(0.0039)		(0.0039)	
$IP_{i,m-1} \times ping_t$		-0.0052				-0.0141*		-0.0156*
$\times Light_{iy} > 58$		(0.0075)				(0.0080)		(0.0082)
$IP_{i,m-1} \times ping_t$			0.0499***		0.0477***	0.0511***	0.0508***	0.0542***
$\times HiIP1_{iy}$			(0.0167)		(0.0163)	(0.0171)	(0.0158)	(0.0166)
$IP_{i,m-1} \times ping_t$				0.0517***				
$\times HiIP2_{iy}$				(0.0169)				
$IP_{i,m-1} \times ping_t$							-0.0534***	-0.0545***
$\times Polity_{cy}$							(0.0193)	(0.0193)
$IP_{i,m-1} \times ping_t$							0.0121	0.0117
$\times HighGDP_c$							(0.0081)	(0.0082)
$ping_t$	0.1039***	0.1443***	0.0863***	0.1018***	0.0618***	0.0878***	0.1899**	0.2140***
	(0.0263)	(0.0349)	(0.0221)	(0.0248)	(0.0183)	(0.0217)	(0.0751)	(0.0763)
N	5380862	5380862	5464123	5464123	5380862	5380862	5161879	5161879

Notes: All specifications include ADM2-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Week clustering. Other interaction terms between subnational and national indicator variables and $ping_t$ as well as subnational and national indicator variables and $IP_{i,m-1}$ included in specifications but not reported.

Appendix

A Joining the IP Activity and geo-IP datasets: details

With a smaller problem the computational challenge of joining the IP Census and geo-location data sets can be efficiently carried out with standard SQL processing. However with the present dataset, comprising trillions of observations, a different approach is necessary. The problem is threefold. First, depending on the IP Census time-stamp, the appropriate location database has to be selected. Second, an individual IP in the Census has to be lined up with an IP range entry in the location database from which the geographical location is then obtained. Finally, an individual geo-located IP needs to be associated with an internet service provider range.

In a map-reduce framework, the join problem has to be specified differently. In the map phase for each element, the join key gets derived, in the reduce phase an assigned reducer collects all items with the same join key and performs the merge. One possible solution to the time matching is to create a virtual key as there are 516 location database revisions from 2005 to 2012. A look up of all time stamps can now be used to create a common key for the IP Census and the location database. However, for the range join required in this study, where $7.0 \cdot 10^7$ ISP ranges have to be merged with $5.2 \cdot 10^9$ location records and then have to be joined with $1.5 \cdot 10^{12}$ observations a different strategy is necessary. The standard approach would involve creating a cross product between the data-sets pre-selected by the virtual key of the matching time followed by elimination of each record which does not correspond to the range. This would work for a smaller data-set but is not feasible for the given cross-product's $5.5 \cdot 10^{28}$ possible combinations.

To proceed, an alternative approach was developed which, to the best of our knowledge, is novel to the field. The condition that needs to be satisfied for our approach is that the ranges per location database and ISP time revision are not overlapping. If that condition holds our approach can be used for any kind of dataset, where an association by range is needed.

The main idea of our two developed algorithms is to segment and sort the data set intelligently, which allows the sub-problem to be solved in small junks on a single reducer task. For the merging of the ranges between location and ISP, the difficulty is that the ISP ranges are blocks of IP addresses for a whole country, regardless of the location within. The ISP range needs to be joined onto the location ranges, creating possible gaps. As the ISP database only contains 70 million rows, at first the ISP database is replicated to pre-associated the closest matching location database out of all 516.

The individual IP and location join operates differently in the first step. The IP ranges have to be partitioned by revision using a modified quintile algorithm, which returns the next real value rather than a value between 2 observations. Finally, for both algorithms, the individual partitions can be merged from both data-sets as a sorted list.

Essentially, two sorted lists are merged sequentially within a partition by iterating both lists one element at a time to find a match, or alternatively, skip an input. Depending on the number of partitions generated the join can be parallelized across various nodes of the distributed clusters as the conditions and sorting guarantees that only possible matching observations are contained in the same partition. It is this scaling effect of the modified join algorithm which delivers the substantial computational advantage required over the otherwise computationally expensive cross-product.

B Internet Latency and Internet Use - Evidence from Online Gaming Behaviour

Valve Corporation, the company that owns and operates Steam, provides a REST API, called the Steam Web API, for gathering information about users' profiles, friendships, game ownerships and playtimes, group memberships, and more. O'Neill et al. (2016) used this API to crawl 108.7 million Steam accounts, to gather information about games played and play times for each account. This dataset comprises all Steam accounts available at the time of collection.

The dataset has been made publicly available as an sql database and can be accessed from https://steam.internet.byu.edu/. From that database we use table "Games 1" which provides results from an API request conducted between May and Novmember 2013. The table contains information about the games that each steam user owns. Each row in the table presents a unique steam user and game combination (i.e. A user that owns 4 games has 4 rows in the table) and the total time (in minutes) the user has run each game in the two-week period leading up to the date of the API request.

From an initial number of around 100 million user and game combinations, we only keep entries where the game was one of the five[24] main online games played during this period (which require the Internet to be played).

We then collapsed the date at a daily level and build the sum of minutes dedicated to playing any of those five online games over the past weeks.

[24] Those games (Steam ID) are: Counter Strike: Source (240) , Team Fortress (440), DOTA (570), and Counter Strike: GO (730)

C Empirical Results based on Monthly Aggregates

To analyze the information channel, we aggregate the weekly variables up to the monthly level (averages) and re-write equation (2) as follows:

$$protest_{im} = \lambda_{iy} + \gamma_1 IP_{i,m-1} + \gamma_2 ping_{y,m-1} + \mu(IP_{i,m-1} \times ping_{y,m-1}) + v_{im} \qquad (4)$$

Our preferred specifications include lags 1-3 and lags 1-6, respectively. Those specifications are then used to calculate the cumulative marginal effects of Internet quality over the past 3 and 6 months, respectively.

The main results are presented in Table 11. Columns (1) and (2) present specifications that include IP activity and latency adjusted IP activity in the previous month, $m-1$. The estimated coefficient of the interaction term is positive, statistically significant at the 1%-level and of similar magnitude as in the weekly specification. In columns (3) and (4) we present a model that adds the second and third, and all lags up to the sixth, respectively, to the specification presented in column (1). Again the coefficient is positive and statistically significant at the 1 and 5%-level.

This coefficients present the cumulative effect of better latency-adjusted Internet over the past 3 and 6 months, respectively. The cumulative effect is likely to reflect the impact of the Internet on protests through the information channel. Interestingly, this effect is larger than the mobilization effect, but the cumulative effects decreases with the inclusion of longer lags. In other words, the effect of better Internet on protests is actually negative in some months. We interpret this as an indication that the crowding out effect of the Internet (for example through its large content of non-political (i.e. entertainment) information) might outweigh the information and mobilization effect.

Table 11: Monthly Aggregates

$protest_{iym}$	(1)	(2)	(3)	(4)
			Marg. cum. effect	
			3 months	6 months
$protest_{iy,m-1}$		-0.0527***		
		(0.0019)		
$IP_{i,m-1}$	-0.0004***	-0.0004***		
	(0.0002)	(0.0002)		
$(IP_{i,m-1} \times ping_{m-1})$	0.0029***	0.0030***	0.0289***	0.0131**
	(0.0007)	(0.0007)	(0.0035)	(0.00537)
$ping_{m-1}$	0.1962***	0.2042***		
	(0.0117)	(0.0117)		
N	1,417,755	1,417,755	1,417,755	1,417,755

Notes: All specifications include ADM2-year fixed effects. *** (**, *): significant at the one (five, ten) percent level. Standard errors (in parenthesis) are adjusted for within Year-Month clustering.