The Impact of News Aggregators on Internet News Consumption: The Case of Localization*

Preliminary and incomplete. Please check with authors for most recent version before citing.

Susan Athey
Harvard University and Microsoft Research

Markus Mobius
Iowa State University and Microsoft Research

February 24, 2012

Abstract

This paper analyzes the impact of news aggregators on the quantity and composition of internet news consumption. In principle, news aggregators can be a substitute or a complement to the news outlets who invest in the creation of news stories. A policy debate centers around the decrease in the incentives for news creation that results if readers choose to consume their news through aggregators without clicking through to the news websites or generating any revenue for the outlets. This paper provides a case analysis of an example where Google News added local content to their news home page for users who chose to enter their location. Using a dataset of user browsing behavior, we compare users who adopt the localization feature to a sample of control users who are similar to the treatment users in terms of recent internet news consumption. We find that users who adopt the localization feature subsequently increase their usage of Google News, which in turn leads to additional consumption of local news. Users also navigate directly to the new sites they have discovered, further increasing their local news consumption. The increase in local news consumption diminishes over time, however, and in the longer run most of the additional local news consumption derives from increased Google News usage. Patterns of news consumption change: users read a wider variety of outlets, more outlets that are new to them, and a larger fraction of their news "home page" views come from Google News rather than the home page of other news outlets. Thus, the inclusion of local content by Google News had mixed effects on local outlets: it increased their traffic, especially in the short run, but it also increased the reliance of users on Google News for their choices of news, and increased the dispersion of user attention across outlets.

*The authors acknowledge support from Microsoft Research. We would like to thank seminar participants at Microsoft Research and the Toulouse Network for Information Technology for helpful comments, as well as Josh Feng for exceptional research assistance.

1 Introduction

A recent policy debate concerns the impact of the internet on the news media. Many authors have noted a series of stylized facts about the industry that suggest the impact of the internet has been quite negative: for example, the Newspaper Association of America reports that from 2000 to 2009, newspaper advertising revenue declined by 57% in real terms, advertising other than classifieds fell by 40%, and circulation fell by 18%. One particularly contentious point in this debate is the role of news aggregators, which we will define here to include sites that do not produce much original content, but rather curate content created by others using a combination of human editorial judgement and computer algorithms. The results are presented with a few sentences and perhaps photos from the original article; to read the full article, users can click through and go to the web site of the original content creator.

The top 6 internet news websites and their monthly unique users are listed below (Schneider, 2012):

Outlet	Unique Monthly Visitors
Yahoo! News	110 million
CNN	74 million
MSNBC	73 million
Google News	65 million
New York Times	59.5 million
Huffington Post	54 million

"Pure" aggregators, such as Google News, generally do not make any payments or have any formal relationship with the original authors of the news content; rather, they create their page by "crawling" the web and then using statistical algorithms together with editorial judgements to organize and rank the content. Only in a few cases does Google News have a direct commerical relationship with the outlets (e.g. Google News had a relationship with the Associated Press, as analyzed by Chiou and Tucker (2011)). In contrast, sites like Yahoo! News and MSN primarily show content from contractual partners. Sites like the Huffington Post may use a hybrid strategy of curating blogs and aggregating news from other sources.

Why are aggregators so controversial? Less than half of page views on the Google News home page result in visits to any online newspapers; thus, users may read their news from Google News without ever generating any page views or revenues for any of the content creators. Clearly, this undermines the incentive of newspapers to invest in journalism.

In addition, news aggregators can substitute for the home page of an online news outlet like the New York Times. The aggregator can index not just the content of the New York Times but all other news outlets, giving it an advantage in coverage. It may then take over the "curation" function that gives the New York Times its differentiation from other news outlets (and this curation may as a result change from being primarily driven by human judgements to being primarily determined by computer algorithms). In addition, since advertising revenues per page are typically much higher than average on the home page of the newspaper, the online newspapers can lose especially valuable page views.

These concerns have been articulated by industry participants. The points have been made perhaps most colorfully by Rupert Murdoch. In a speech at the FTC in 2009, Ariana Huffington quoted Murdoch's various speeches referring to aggregators as "parasites, content kleptomaniacs, vampires, tech tapeworms in the intestines of the Internet, and, of course, theives who steal all our copyrights." At the same conference, Murdoch called the internet a "disruptive technology" and stated:[1]

> When this work is misappropriated without regard to the investment made, it destroys the economics of producing high quality content. The truth is that the 'aggregators' need news organizations. Without content to transmit, all our flat-screen TVs, computers, cell phones, i-Phones and blackberries, would be blank slates. ...To paraphrase a famous economist – there's no such thing as a free news story.

On the other hand, aggregators may complement newspapers. Consumers incur costs (time and effort) in seraching for news that may interest them. They will compare the expected benefit from visiting a news site to the expected search cost, where that cost includes becoming aware of the existence of the site and finding how to navigate it. A typical user may forget about smaller niche sites, such as local sites, or may decide that the benefits of visiting do not outweigh the cost. Rather, a user may focus on "big-name" sites such as CNN and the New York Times, together with a few personal favorite sites. The impact of aggregators in this case would be to decrease the concentration of user news browsing, increasing the share of consumption on smaller outlets and decreasing the share on larger outlets. Aggregators might also allow consumers to become informed about the quality of a wider range of outlets, leading them to find outlets that are a better match for their interests. This may, in turn, increase a user's overall consumption of news as well as their consumption of news aggregators.

Even if some outlets benefit from an increase in traffic, outlets may also see a change in the composition of the traffic. Outlets may see more "casual" users and fewer "loyal" users, which has implications for the ability of the outlets to target ads (since they have more information about the preferences of loyal users) as well as competition in advertising markets. Indeed, Athey, Calvano, and Gans (2011) show that when consumers switch more often among outlets, advertising is less efficient (more users see duplicate ads) and that this effect is exacerbated by competition among outlets, leading to lower advertising prices.

[1]For coverage of the conference, see Albanesius (2009). For Murdoch's speech see FTC (2009b) and for Huffington's speech see FTC (2009a).

The goal of this paper is to provide empirical evidence about the impact of news aggregators on news consumption. We focus on the effect of aggregators on local news outlets. This example is perhaps an especially favorable one from the perspective of the aggregators, since aggregators may play a more important role in the discovery of local news sites, since they are generally small. We study the introduction of the "local news" feature in France in late 2009. Google enabled a feature where if users enter their zip code, on all subsequent visits where the user is recognized (their cookie is recognized by Google, which will generally be the case if the user has the same machine), the user sees news from local outlets prominently featured on the page.

Our dataset is a sample of all browsing events for 9.3 million computer users in France who use a Microsoft toolbar and have opted in to allowing their data to be used for research purposes.[2] Of those users, 18% meet a criteria we call "consistent" users, and 2% of those consistent users use Google News home page at least twice per month. We compare the news consumption of users who enable this feature ("treatment users") to the contemporaneous usage of a set of "control" users who are similar in terms of recent internet usage frequency and intensity, consumption of local news, use of Google news, and geographical location. For local news consumption, we match based on behavior within the past two weeks as well as the last two days prior to the opt-in decision of that treatment user, in order to control for the fact that a treatment user's interest in news may be trending upward just prior to the opt-in decision and the opt-in decision is thus correlated with an unobserved increase in interest in news. We also incorporate additional control variables for robustness, such as the internet searches for local, national, and other news topics in the 24 hour period just prior to the Google local signup decision. In our main results, we focus on users with consistent toolbar usage and an established prior history of using Google news; this yields a sample of 1,800 "treatment" users. Our main analysis focuses on the two weeks subsequent to the local news signup event, but we also present results for 8 subsequent weeks.[3]

Our empirical analysis proceeds as follows. We consider the effect of the local news signup on a number of outcome variables. In each case, we examine two types of specifications, one including a control for the total Google news utilization after signup, and one that does not. The idea is that a key channel for changes in overall news consumption is increased utilization of Google news. If the local news feature makes Google news more attractive, and thus news consumption more attractive relative to other activities, then we should see an increase in visits to both Google news and all news outlets. Controlling for Google news utilization allows us to isolate additional increases in news consumption beyond consumption that comes directly from Google news.

A preliminary result is that users who adopt the local news feature on Google news subsequently increase their utilization of the Google news home page substantially – by

[2]The data is subject to stringent privacy restrictions and at all times resides only on secure servers, and only aggregate statistics and the output of statistical models can be reported. However, we are able to construct the variables for analysis using the fully disaggregated data.

[3]Longer term analysis is difficult to do because users disappear from our sample.

more than 50%. This could be due to the user getting more utility from a page with more personalized and relevant content.

Our main result is that adoption of Google news leads to greater consumption of local news, both unconditionally (by more than 26%) and conditional on Google news page views, though Google news explains a substantial portion of the increased traffic to local sites. We see a 5% increase in direct navigation to local outlets (bypassing Google news altogether, presumably because the user has learned that they like the outlet and actively chooses it in the future), and a 13% increase in clicks on local outlets from the Google news home page. We show that when we consider the longer-term effects of Google local adoption (over an 8 week period), the local news treatment effect is attentuated somewhat, but remains highly statistically significant and still substantial in magnitude (over 14%). However, the effect becomes smaller and is no longer statistically signficant once we control for Google News home page usage, which suggests that over time the incremental local news consumption derives primarily from increased utilization of Google News.

We also look at the composition of news browsing. We see more than 12% increase in the number of local outlets used. In addition, treatment users visit 10% more new local outlets than control users, while there is no significant difference between the number of old outlets visited between treatment and control users. Thus, the local feature does seem to introduce users to new local news sources which they then continue to visit.

Finally, we examine the extent to which the Google local news feature cuts into the "curation" role of newspapers. We see that the ratio of news outlet home page views to Google news home page views falls by more than 16%. Many of the incremental page views on outlets originate on Google and users are sent to the article directly, bypassing the profitable home page of the news outlet. They may subsequently read other articles in the outlet through following links they see on the same page as the original article, and thus their browsing may never take them to the outlet's home page.

Even though our results broadly support the hypothesis that news aggregators are complements for local news outlets, it is important to emphasize that the impact on local news outlets is mixed overall. Some outlets gain more than others, users spread their consumption over a larger number of outlets, and the curation role of news outlets is diminished. In addition, the long run substitution patterns cannot be fully determined using the empirical strategy in this paper. Finally, industry participants have noted that there are potentially problematic interactions between search engines and their own news aggregators, since users who use Google to find news are directed primarily to the Google news page rather than to news outlets directly. In France, Google's market share in internet search is above 90%.[4]

[4] For example, the Media Institute submitted a white paper to the FTC entitled "Google And The Media: How Google Is Leveraging Its Position In Search To Dominate The Media Economy," where it argued that Google search is a dominant search engine, and that Google search sends traffic to Google news rather than directly to news outlets. In turn, they argue that Google news appropriates others' content, and not all users click through. Finally, they note that news rankings are a combination of

Despite the intrinsic policy importance of the news industry and the close attention this issue has received from regulators, there is very little existing empirical evidence on the impact of aggregators. The paper closest to this one is Chiou and Tucker (2011). They study a " natural experiment" where Google News had a dispute with the Associated Press, and as a result, did not show Associated Press content for about seven weeks. The paper has aggregate data about page views to Google News as well as the sites visited immediately after Google news. They use views to Yahoo! News as a control. The paper finds that Google News is a complement to news outlets: taking the Associated Press content away from Google News lead to fewer visits to news outlets (where Associated Press articles are featured). Our paper is complementary to theirs: our main result is consistent with theirs, even though the nature of the change experienced by users is different (addition of local content versus removal of Associated Press content). The nature of our data enables more nuanced analysis: we observe individual level data, and can thus answer questions beyond the aggregate effect on referrals from Google News to outlets. We are also able to measure the impact of Google News on direct navigation to outlets at the individual level, which was not possible with their aggregate data.

2 The Google News Local Feature

As described in the introduction, we analyze the addition of the local news option on the French Google News website. Starting on November 2^{nd}, 2009, users in France could add a local news section to their Google News Home Page (GNHP).[5] Users are presented with a small "personalization" box on the screen, where they can choose to enter their zip code on the GNHP in order to access the section. Once a user signs up, the local news section appears as one of the regular sections right below the box on subsequent visits.

The following shows the localization box and the change that it induces for users who sign up.

human editorial and algorithms and they argue that Google can and does retaliate against outlets that opt-out of Google news by delisting them from search.

[5] See http://googlenewsblog.blogspot.com/2009/11/local-news-now-available-in-france.html

Figure 1: Before November 2: The blue box on the top right of the partial screenshot gives the user the option to customize their own news section

Figure 2: After November 2: The blue box is now titled "Actualites locales."

We identify users through the presence of the "zx" parameter in the redirect URL that users go to after entering their zip codes. The format of the news homepage remained the same through late 2011.

3 The Data

3.1 Data Sources

Our main source of data comes from user-level online browsing logs covering the period from October 2009 through October 2010. The dataset includes users who used a browser with a certain toolbar installed, unless the users opted to avoid sharing their data, and it includes all browsing events using the browser with the toolbar. Of course, individuals use multiple computers and so our sample is not necessarily representative of all internet usage (e.g. smartphone browsing or browsing using corporate computers may be underrepresented in our sample). Similar limitations apply to most data sources about internet usage, since few corporations enroll in consumer panels like ComScore, and other large firms like Experian Hitwise (the source for Chiou and Tucker (2011)) rely on data from internet service providers, which also underrepresents corporate use and smartphones.

These browsing logs contain the user ID, URL, time, dwell time, and referrer URL for each browsing event. We process these logs to create a user summary dataset and a news event dataset. The user summary dataset contains user location and general usage statistics, such as overall daily page views and time spent online. The news event dataset contains browsing events related to news, including Google News related events and any browsing on news outlet websites.

All of our summary statistics and analysis are constructed on a subsample of users that we call "consistent toolbar users." A user can be a consistent toolbar user in one month but not another, and they meet the definition in a given month if they have a recorded toolbar event in at least 25% of the days. The idea is that we want to exclude users who are only using the toolbar rarely, since their browsing may not be representative of their overall internet usage.

To construct a list of French news websites, we take advantage of French Wikipedia. We wrote a program to extract official website URL's from all Wikipedia concept pages related to the general news category. Wikipedia also contains a local news category, which we use to mark news outlets as local. More details about how we identify and categorize news sites can be found in Appendix B.1.

We have a few more decisions to make in interpreting the data. "Dwell time" measures the amount of time a user spends on a particular page. However, long dwell times are prone to measurement error, since users may switch tasks, leave their computer, etc. To alleviate this concern we cap dwell times at 30 seconds for individual page views.

We also need to implement a definition for what we call a "session." For example, a user can click on a news article from the GNHP, and after having read the article, and proceed to click links to other articles published by the news outlet. To define news sessions, we implement the following algorithm:

1. Group the browsing data by user and news outlet, and order by timestamp within each group
2. Create a new session if there is no referrer or the referrer is from a different domain or if there is a 30 minute gap between events in the user-news outlet data.
3. Otherwise, assign the same session ID to browsing events that are connected.

We next define the concept of "direct navigation." The concept we attempt to capture is that the consumer directly navigates to a site rather than receiving a referral from a news aggregator. We define an event as direct navigation if we don't see a referrer at the beginning of a session. It is important to emphasize that we systematically miss certain types of referrals: referrals arising from secure sessions (https), those implemented with javascript, etc. We define an *immediate* referral from Google News as a referral where the user clicked on the Google News page and then arrived at an outlet.

Although our data are quite rich, there is one significant data issue that requires discussion. The internet browsing data is only retained for a limited time, and due to the timing of the start of our research project, we were not able to capture all of the relevant data before it was deleted. We are missing data from the following dates: November 9th - November 19th, 2009; December 24th - December 28th, 2009; and January 1st - January 22nd, 2010. Since we gather data before and after the signup date for each user, this affects different treatment users (and their associated controls) in different ways. We deal with the issue by simply counting time in our data as if those days did not exist, so that all variables are constructed using the same number of calendar days, but those days may not be contiguous in calendar time due to the gaps. As a robustness check, in Section 6.2 we present results using only treatment users who signed up after March, 2010 so that there are no gaps. The results are similar.

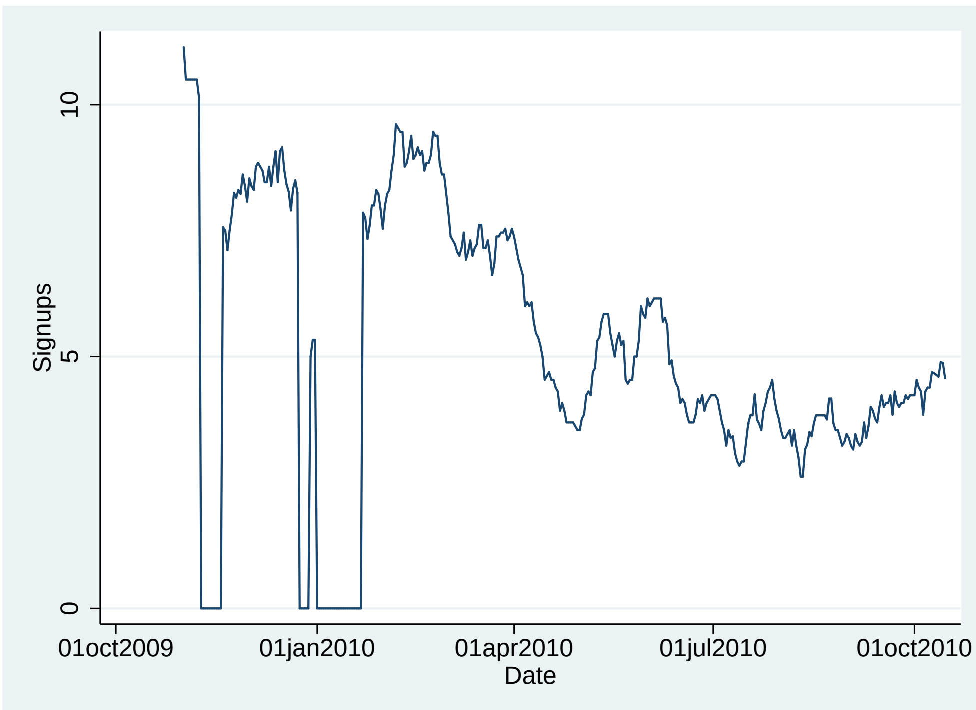

Figure 3: The number of sign-ups over time. Periods with zero signups are due to the missing data.

3.2 Variable Description

In the user-outlet-day dataset, we keep aggregates of the following variables (we cap all variables at the 95th percentile for our empirical analysis):

- Page Views: The number of browsing events on the news outlet's website (for the user on the given date)

- Dwell Time: The total time spent on the news outlet's website, capped at 30 seconds
- Distinct Sessions: The number of sessions (determined by the aforementioned News Session Algorithm) on the news outlet's website
- GNHP Clicks: The number of browsing events on the outlet's website that are the result of a click on the GNHP
- Direct navigation to the news website: The number of browsing events on the outlet's website that have no referrer (either the user typed in the URL or used a bookmark)
- GNHP Clicks: The number of browsing events within sessions where the first event in the session was a result of a click from the GNHP.
- Direct Navigation: The number of browsing events within sessions where the first event in the session had no referrer
- Index Page views: The number of browsing events on index pages (see Index Page Identification algorithm in the Appendix)
- Index Page Dwell Time: The total time spent on index pages
- Categorization of the outlet (local, non-local)

In the user-day dataset, we keep aggregates of the following variables:

- User Zip code: Five-digit French zip code for the user
- Total Browsing page views: The total number of browsing events (for the user on the given date)
- News Page Views: The number of browsing events on news publisher websites
- News Dwell Time: The total amount of time spent on news publisher websites
- GNHP Clicks: The number of browsing events on the outlet's website that are the result of a click on the GNHP
- Direct navigation: The number of browsing events on the outlet's website that have no referrer (either the user typed in the URL or used a bookmark)
- Overall GNHP: The number of browsing events within sessions where the first event in the session was a result of a click from the GNHP.
- Overall Direct: The number of browsing events within sessions where the first event in the session had no referrer
- Outlets Used: The number of distinct news outlets visited

- News Searches: The number of internet searches conducted where the user clicked on a link to a news outlet
- Versions of these news-related variables just for local news outlets ("Local" prefix)

Tables 1 and 2 show summary statistics for the top French news outlets and top local news outlets in our data.

4 Empirical Approach

Our goal is to identify the causal effect of being exposed to the local news content on the Google News Home page, on a variety of outcomes. To identify this effect, we clearly cannot simply track the treatment group over time. A variety of confounding factors could bias the results, including secular trends in news consumption, local elections, seasonal trends in local news consumption, and individual trends. Of particular concern is the fact that adopting Google local is a signal of interest in news at a point in time: the user needed to visit Google News home page on that date, and their decision to enter their zip code is a particular signal of engagement in news.

Although it will be impossible to provide an unassailable empirical strategy, the richness of our data allows us to find "control" users who are similar to treatment users along the key dimensions that could confound our empirical strategy. We focus on consumption of news and local news in the time just prior to signup.

We then compare the outcomes for treatment users to those of control users in the time after signup.

The next subsection provides details of our procedure for selecting control users.

4.1 Selecting Control Users

Our procedure for selecting control users is specific to each treatment user, since each treatment user has a different signup date, and the metrics we use for matching are constructed relative to a particular date.

We begin by eliminating from both the treatment group and the potential control group all users who do not have at least 2 days with internet activity in each of the four weeks surrounding the signup (two before and two after). This ensures that we analyze consistent internet users.

We then match controls to treatment users based on similarity in total local news page views, GNHP visits, and internet usage in the two weeks leading up to the treatment user's sign-up date, as well as short-term local news usage (in the 2 days before the signup) and geographical distance between the pair of users. After the matching is complete, we counterfactually assign the treatment user's sign up date to her corresponding control users.

Finally, we aggregate the usage variables in the periods before and after the sign-up date for each user to create pre and post variables for our regression analysis. To focus in on the longer term effect on the treatment, we do not include data that is within three days of the sign-up date in the aggregation.

More formally, we take as our initial set of treatment users all users who sign up for Google local between November 2nd, 2009 and October 17th, 2010. We then use the following algorithm to select the candidate control users for a given treatment user:

1. We eliminate the treatment user if they do not use the internet on at least 3 days each of the four weeks surrounding the sign up (two before and two after).
2. We find all potential control users who have at least 3 days in each of the four weeks (two before and two after) surrounding the treatment user's sign up date.
3. We filter this list by keeping all potential control users who are within 50% of the treatment user in terms of internet usage days, total browsing page views, local news page views, and GNHP views in the two weeks before the sign up date. As a means to control for time trends in local news usage, we also match on local news page views in the two days before the signup plus the local page views on the day of the signup before the addition of the local section (also at a 50% tolerance). Finally, we also eliminate users that are more than 400 kilometers away from the treatment user.
4. For each pair of possible matches, we generate a proximity score:

   $$\delta_{days}^2 + \delta_{bpv}^2 + \delta_{lpv}^2 + \delta_{gpv}^2 + \delta_{slpv}^2 + (\frac{35}{400}D)^2$$

   where the δ represent the percentage difference between the control and the treatment users' usage in the matching variable and D represents the distance in kilometers between the users.

At the end of the process, we have a set of candidate control users for each treatment user, but some control users are eligible for multiple treatment users. As a final step, we randomly assign up to 20 control users to each treatment user in the following manner. First, we randomly order possible treatment users. In each round of selection, we traverse the treatment users in order, and pick for each treatment user the available control user with the lowest proximity score. After a control user is selected, they lose their eligibility to be selected as a control for another treatment user. We run 20 rounds of selection to obtain our control group.

For some treatment users, fewer than 20 potential control users exist. We eliminate treatment users with less than 4 control users (this is less than 5% of observations), and for the remainder, we construct weights so that the total weight for all control group users for a given treatment user is 20.

As a final step, we counterfactually assign each control user the sign-up date of her corresponding treatment user.

Table 4 shows summary statistics for our matching algorithm. We see that on several dimensions, Google News home page views, local news page views, and short-term local page views, the median control observation is an exact match for the treatment observation. Overall internet usage (browsing days, browsing page views) are within 10% on average, with medians of 3.4% and 5.6%. The median control user is within 13.4 km of its treatment user, with an average distance of 81.1km. Thus, even before controlling for these variables in our regression analysis below, we have constructed a dataset where control users are extremely similar to treatment users along our matching dimensions. However, some discrepancies remain on other variables; in future versions of the paper we plan to modify the matching strategy to achieve similarity across a greater number of dimensions.

4.2 Summary Statistics

The basic summary statistics of our data are shown in Table 6 (separately for control and treatment users). As expected, control and treatment users look very similar along the matched pre-treatment dimensions. However, they also look similar along post-treatment dimensions where we would not expect a significant treatment effect. Both groups exhibit very similar toolbar usage after the treatment. The mean toolbar days and toolbar page view differ by only 3.5% and 1.7%, although there is a slight trend down in the control group and a slight trend up in the treatment group from pre to post period.

As far as local and non-local news consumption is concerned, we see a 23 percent increase in the mean page views and mean dwell time for local news (1.966 to 2.411 page views and 156.9 to 192.9 seconds of dwell time). Non-local news consumption and dwell time remains essentially unchanged (37.5 versus 38.2 mean page views).

For the control group, we see a slight decrease in all forms of news consumption: mean local news page views go down by 12% (137.0 to 122.7) and dwell time by 11% (137 to 122.7 seconds). Non-local use consumption also goes down slightly by about 9% (34.80 to 31.84). This decline is in line with overall decline in toolbar usage of about 8% between the pre and post-signup period for control users.

These patterns suggest that treatment users' overall browsing behavior does not change significantly after signup. However, their local news consumption increases significantly compared to control users and their non-local news consumption escapes the decline experienced by control users.

We control for the matching variables in our empirical work, so differences that remain between treatment and control groups can be accounted for. However, in future versions of the paper we plan to further explore the sources of the downward time trend in news consumption for the control group and refine our matching criteria if appropriate.

4.3 Empirical Model

Our primary empirical model can be written as follows

$$y_{i,post} = \beta_0 + \beta_1 y_{i,pre} + \beta_2 x_{i,pre} + \tau D_i + \epsilon_{i,post}$$

where y represents relevant usage variables such as local news consumption, GNHP consumption, and overall news consumption, i is the individual, and D_i is a dummy variable indicating whether the user is a treatment user. We assume that the $\epsilon_{i,t}$ are independent across users, except that we allow for correlation between the ϵ 's for the treatment user and their associated control users (we cluster the observations when calculating standard errors in our regressions). The interpretation of this assumption is that a user's consumption follows an autoregressive process, where today's consumption is correlated with yesterday's consumption but the user's consumption experiences a new shock in each period.

We estimate the above equation using linear regression. The key assumption required for the estimate of τ to be an unbiased estimator for the effect of the treatment is that the shock to the post-period outcome is uncorrelated with the treatment dummy. The assumption might be violated if the user's consumption is trending up in unobserved ways, and this is correlated with the Google local news signup decision. In our empirical implementation we include additional "pre-period" outcome variables as controls, which are designed to provide measures of the user's interest in news in the period just prior to signup (these are the same measures we used for matching, as well as news-related (local and non-local) and other internet searches in the 24 hours prior to signup).

Note that our functional form is different from a standard difference-in-difference estimator, where we would take as the left-hand side of the regression the difference $y_{i,post} - y_{i,pre}$, or else include an observation for each time period and include fixed effects for the users, a dummy variable for the "post" period as well as a dummy variable for the interaction of "post" and the treatment. The latter specification would require an assumption that a user's deviation in outcomes from the user's mean outcome levels and aggregate time trend are independent over time. We chose the above specification because when viewed over longer time periods, we see serial correlation in deviation from mean outcomes. That is, individual browsing behavior tends to follow an autoregressive process, rather than a process where browsing behavior has serially independent shocks around a mean. In Section 6.5 we show the results from the more standard difference in difference specification.

5 Results

For each outcome of interest, we consider four specifications. Continuous variables are specified in logarithms where we add 1 to the base variables to account for observations of 0. The first specification does not include any of the matching variables as controls (no x's). The second specification includes all of the variables we use for matching as controls, as well as the number of local, non-local, and other internet searches that were conducted on Google in the 24 hours prior to signup. The third specification includes the *post-period* page views on Google News Home Page as an explanatory variable, while the fourth adds an interaction of the treatment with the post-period page views on Google News Home Page (to account for the fact that propensity to click will depend on the treatment, since the treatment changes the types of news available).

Tables 7 and 8 summarize our estimates of the treatment effect τ from equation 4.3 for local and non-local news outcome variables. Each row specifies the outcome of interest, while each column corresponds to one of the four specifications just described. The full regression results are available in Appendix A.

An important note is required about the interpretations of the magnitudes of the treatment effects. Since our specification takes the logarithm of the outcome of interest *plus one* as the dependent variable, the coefficients on the treatment dummy can be interpreted only in terms of percentage changes in that transformed variable. However, the coefficients still provide a lower bound for the percentage change in the dependent variable[6] In subsequent versions of the paper, we will also calculate predicted percentage changes to the underlying variables.

We also emphasize that many of the outcome variables are zero for more than half the users (as indicated by the median being 0). For example, median GNHP use as well as pre and post local news use is zero for both treatment and control users. This raises some functional form issues. In this version of the paper, we also estimate linear-probability variants of our regressions with a dummy for local and non-local news usage as the outcome variable, and future versions will report additional functional form. We will see below that the linear probability model gives larger treatment effects than the model based on logarithms, consistent with the fact that a lot of the variation in the dependent variables takes the form of changes from 0 to 1.

We also observe that all treatment effects are fairly precisely estimated: the standard errors in the page view regressions are about .025, implying that effects above about 5% will be distinguishable from zero at the 95% confidence level. Regressions with time

[6] If $1+x$ increases by a fraction α then the relative change in x equals $\alpha\frac{1+x}{x} > \alpha$ since x is positive.

change as the outcome have standard errors that are about .07, but the estimated effects are also larger.

5.1 Aggregator Usage

We begin by examining the impact of the signup on visits to the GNHP. We see that the treatment effect is .504. This substantial increase in GNHP views is crucial to understanding the rest of the results in this paper, since users click through to news outlets after visiting GNHP about half the time. As we will see below, the treatment also impacts the distribution of clicks, since the treatment exposes local news. We will see that the treated users are relatively more likely to click on local news, relatively less likely to click on non-local news, and more likely to click overall than the non-treated users after signup.

5.2 News Usage

With these basic facts in mind, we now consider the impact of the treatment on our outcomes of interest. We focus on specifications (2) and (4), which include all of our matching variables as well as controls for internet searches just prior to signup, and where specification (4) includes controls for GNHP page views in the "post" period as well as those page views interacted with the treatment. Thus, specification (4) decomposes the treatment effect into the part that derives from direct navigation to outlets, and the part deriving from increased usage of the GNHP together with changed clicking patterns on the GNHP.

Local News

We begin by comparing local news usage of treatment and control users. We use two measures of local news usage – local news page views and a local news indicator which is 1 if the user visits *any* local news site after the treatment. The treatment effect in specification (2) is 0.212 for local page views and 0.815 for the news indicator. Once we control for GNHP usage in specification (4) the treatment still has a strong positive effect of 0.359 through the news indicator. For the page view measure the direct treatment effect becomes insignificant but there is a strong positive interaction effect between GNHP usage and the treatment. We interpret the result as showing that when users have the local news feature enabled, they are more likely to click through to local news sites when they go to GNHP. Thus, most of the increase in local news derives from the increased GNHP usage together with the changed behavior conditional on viewing GNHP. We see that there is some difference in the treatment effect from specification (4) between the indicator measure and the local news page views (in logarithm) measure. In a future version of the paper we hope to provide additional evidence on the sources of the discrepancy as well as additional robustness checks.

When we study dwell time, the results qualitatively mirror the result for local page views: we also see large significant effects in specification (2), at .676. The effect decreases to almost zero (.014) once we control for GNHP usage and its interaction with the treatment in specification (4).

Non-local News

Non-local use also increases in the treatment group: the effect is 0.247 in specification (2). However, the treatment effect disappears once we control for GNHP usage in specifications (3). When we add interact GNHP usage with the treatment dummy in specification (4) the effect becomes negative: this indicates that treatment users substitute away from non-local news as their use of GNHP increases. We see a similar substitution pattern with the other non-local measures, leading to treatment effect estimates that are smaller in specification (3) (without the treatment x GNHP interaction) than in specification (4).

We find analogous results for other measures of non-local news consumption. When we use an indicator for non-local news outlets as the dependent variable we see again a large treatment effect in specification (2) which becomes smaller and insignificant once we control for GNHP usage. Dwell time on non-local news also increases significantly for treatment users by 0.556 but is no longer significant once we control for GNHP use in specification (3), though it is larger (.23) and significant in specification (4).

5.3 Outlet Usage Concentration

We then turn to look at how the treatment affects the composition of local news consumption. The coefficient on the number of different local outlets visited in specification (2) is .128. The increase is again driven by increased GNHP usage of treatment users as specification (4) shows.

We next distinguish between "new" and "old" outlets. Old outlets are those that the user had already visited in the pre-treatment period. New outlets are visited for the first time after signup. We are interested whether the overall increase in local page views comes at the expense of older outlets. Specification (2) shows that the treatment users discover significantly more new local news than the control users. However, they also read more local news in old outlets. The effect for new outlets is larger than for old outlets (0.093 versus 0.506) but both are strongly significant. When we control for GNHP usage we see that both effects are mediated by increased aggregator usage.

When we turn to non-local news outlets we see an increase 0.214 for treatment users in number of non-local outlets visited. However, once we control for GNHP usage it becomes insignificant. An interesting pattern emerges when we distinguish between new and old non-local outlets: users appear to discover new non-local news outlets through increased usage of GNHP. This effect is 0.302 in specification (2) and remains positive at 0.126 and strongly significant even after controlling for GNHP usage in specification (4).

In contrast, treatment users visit old non-local outlets less by -0.175 after controlling for GNHP.

This suggests that the increased aggregator use triggered by localization has two effects on non-local news consumption: (1) it mechanically scales visits to all non-local sites up; and (2) users discover new non-local outlets and they substitute away from old non-local outlets.

5.4 Role of News Outlet as Curator

An important role of newspapers is to act as a curator of news. In offline newspapers, editors select which news makes it to the front page and how prominently each story is displayed. In online newspapers, the index page takes the role of the front page. Aggregators like Google News bypass the publisher's index page – they essentially replace the publisher's front page with their own index page.

We identify index pages among all local and non-local page views using a procedure outlined in Appendix B.2. We then use the share of index pages among local and non-local news as a left-hand side variable. The index page ratio falls by 16% for local news users in specification (2) which is explained by the increase in GNHP usage when we control for Google News usage in specification (4).

5.5 Long-Term Effects

So far, we have fixed the post-signup observation period to 2 weeks. In this section we check whether the effects of localization on news consumption persist over a longer period. We construct a sub-sample of "long-term" treatment users whom we observe for 2 periods before signup as well as 8 weeks after signup. Just as before, we focus on consistent users who use toolbar for at least 3 days a week in each of the 10 weeks. We select an analogous set of potential control users and use the same matching algorithm as before to construct a set of control users for every treatment user. This provides with 685 long-term treatment users compared to 1,904 treatment users in the full data set.

For both local and non-local news outcomes we re-estimate all 4 specifications for three of the main outcomes: page views, news indicator and number of news outlets. In each case, we estimate separate regressions for weeks 1-2, weeks 3-4, weeks 5-6 and weeks 7-8. The results are shown in table 9 for local news and table 10 for non-local news.

For weeks 1-2 the estimates from the smaller long-term sample are almost identical to the estimates from the full regression (for both the local and non-local outcomes). The same holds for the pattern of significant coefficients. For example, for local news and specification (2) the treatment effects are 0.220 versus 0.212 for page views, 0.750 versus 0.815 for news indicator and 0.140 versus 0.128 for number of outlets.

The overall treatment effects captured through specification (2) diminish over time: 0.220 to 0.123 for local page views, 0.807 to 0.414 for local news indicator and 0.140 to 0.0744 for number of local outlets. However, the treatment effects remain strongly significant even in weeks 7-8. When we control for GNHP usage, the direct treatment effect disappears over time and the additional local news consumption derives primarily from increased Google News usage.

6 Robustness Checks

In this section we perform a number of robustness checks.

6.1 Excluding Paris

One particular concern about measuring local news usage with French data is the fact that about 18 percent of the French population lives in the metropolitan area of Paris. While our Wikipedia-based approach does distinguish between local newspapers in Paris and national newspapers, our classification is likely less clear-cut than for the rest of France. We therefore estimate our main regressions by excluding the Paris department (74) as well as the three departments that border it (92, 93, 94).[7]

Table 11 shows the estimates for local left-hand side variables and table 12 shows the results for non-local variables.

Our local news estimates are very close to the estimates with the full sample. For specification (2), we estimate 0.243 versus 0.212 for local page views, 0.842 versus 0.815 for local news indicator and 0.148 versus 0.128 for the number of local outlets (all treatment effects are also strongly significant as in the full regression). Specification (4) actually becomes stronger because now both the pure treatment effect *and* treatment interacted with GNHP usage are strongly significant (in the full regression this is only the case for the local news indicator).

For non-local news, the estimates are also very similar for specification (2). Controlling for GNHP usage makes the treatment effect insignificant in specifications (3) and (4) except for the non-local news indicator.

6.2 Gaps in the Data

In our full data set we lack observations for a number of weeks in November and December 2009 and January 2010. We deal with this problem by cutting these gaps out of the calendar. To check the validity of this procedure, we estimate our main regressions

[7] English Wikipedia has a map of French departments: http://en.wikipedia.org/wiki/Departments_of_France

by only using data collected after January 22, 2010 (from that date on the data has no more gaps).

Table 11 shows again the estimates for local left-hand side variables and Table 12 shows the results for non-local variables.

For local news outcomes, the estimates are extremely close to the estimates with the full data for all specifications. For example, for specification (2), we estimate 0.216 versus 0.212 for local page views, 0.842 versus 0.815 for local news indicator and 0.133 versus 0.128 for local news outlets.

For non-local news, we also get similar estimates and the same patterns of significance for non-local page views non-local outlets. For non-local news indicator the treatment effect survives after we control for GNHP usage while it becomes insignificant with the full data.

6.3 Early Adopters

Google News enabled localization in November 2009 while our data coverage stops in October 2010. We therefore include treatment users who enabled localization in November 2009 as well as those who enabled it in September of 2010. We check whether there is difference between "early adopters" and "late adopters" by estimating our main regressions for early adopters only (those who sign up before January 1, 2010).[8]

The local news results are again very similar to our estimates from the full data set across all specifications. For non-local news, the pattern of significance is the same as in the full regression. Note that the early adopters also have data confounded by gaps, as discussed in the prior section.

6.4 Functional Form

To be done.

6.5 Standard Differences in Differences Specification

To be done.

[8] This robustness check is complementary to the exercise in the previous section where we looked mostly at late adopters.

7 Conclusions

In this paper, we have presented evidence about the short and medium term effects of the introduction of the Google Local News feature in France in late 2009. We have found that the introduction lead to a significant increase in local news consumption, in part due to additional usage of Google News, but also due to an increase in users directly navigating to local sites that they initially discovered through Google news. The gains diminish over time, but remain positive. On the other hand, non-local outlets that a user previously visited see a lower number of page views conditional on the number of visits a user makes to the GNHP.

We also demonstrated that as a result of increased utilization of Google news, users have more dispersed browsing patterns, and that users rely more on Google news to curate their news relative to the home pages of news outlets.

Although these results suggest that some of the negative short-term effects that local news outlets might have feared are not present, we caution that our results should not be interpreted as providing evidence that news outlets are better off because Google news exists, either in the short run or the long run. As discussed in the introduction, there are a number of longer-term threats to news outlets created by news aggregators, including loss of the curation role which affects the brand perception of the news outlet as well as its ability to promote news that is for any reason not selected by Google news. The increased switching and tendency to bypass the home page of the news outlets induced by aggregators also creates challenges for generating advertising revenue.

In future work, we intend to explore the supply side responses of news outlets to the demand changes induced by aggregators. An understanding of how supply and demand forces interact to determine the news that is created and consumed is crucial to addressing the pressing policy issues faced by the news industry.

References

Albanesius, C. (2009): "Murdoch, Huffington Squabble Over Online News," http://www.pcmag.com/article2/0,2817,2356538,00.asp.

Athey, S., E. Calvano, and J. Gans (2011): "The Impact of the Internet on Advertising Markets for News Media," Discussion paper, Harvard University.

Chiou, L., and C. Tucker (2011): "Copyright, Digitization, and Aggregation," Discussion paper, Massachusetts Institute of Technology.

FTC (2009a): "From Town Crier to Bloggers: How Will Journalism Survive the Internet Age?," http://www.ftc.gov/opp/workshops/news/transcripts/091201transcript.pdf.

——— (2009b): "Rupert Murdoch Before the Federal Trade Commission's Workshop: From Town Crier to Bloggers: How Will Journalism Survive the Internet Age?," `http://www.ftc.gov/opp/workshops/news/docs/murdoch.PDF`.

Schneider, J. (2012): "Top 15 Most Popular News Website," `http://www.ebizmba.com/articles/news-websites`.

A Tables and Graphs

Table 1: Top French News Outlets

Outlet	%Users	%Brows.	%News	PVs per user	%GNHP Dir	% GNHP Ov
lequipe.fr	10.45	.2189	15.73	469.7	.0913	.1454
boursorama.com	4.036	.1107	7.956	237.1	.0142	.0198
eurosport.fr	5.128	.0636	4.565	136.5	.1608	.2093
orange.fr	10.98	.0543	3.909	116.1	.003	.0079
francefootball.fr	1.689	.0373	2.673	79.31	.087	.1113
footmercato.net	2.584	.0368	2.637	77.63	.4978	.5823
lephoceen.fr	.9448	.0238	1.709	50.78	.0552	.0768
leparisien.fr	7.004	.0233	1.675	49.67	3.694	5.124
lefigaro.fr	6.872	.0228	1.642	48.77	3.275	4.416
football365.fr	2.94	.0198	1.425	42.56	.2852	.3262
purepeople.com	6.495	.0192	1.379	40.82	.7704	.8652
ouest-france.fr	3.524	.0176	1.266	37.58	1.298	1.849
lemonde.fr	5.397	.0175	1.261	37.37	3.339	4
tf1.fr	4.194	.0125	.8999	26.51	1.363	1.605
planet.fr	1.523	.0122	.8753	25.79	.0077	.0112

Table 2: Top French Local News Outlets

Outlet	%Users	%Brows.	%News	%Local	PVs per user	%GNHP Dir	% GNHP Ov
ouest-france.fr	3.524	.0176	1.266	18.55	37.58	1.298	1.849
lanouvellerepublique.fr	.6554	.0093	.6698	9.858	19.97	.0626	.0852
lavoixdunord.fr	2.185	.0087	.6248	9.157	18.54	.5823	.6769
laprovence.com	1.4	.0085	.6119	8.979	18.18	.1854	.2506
leprogres.fr	1.258	.0076	.5446	7.976	16.15	.3652	.4101
midilibre.com	1.062	.006	.4282	6.261	12.67	.3794	.4212
letelegramme.com	1.764	.0045	.3207	4.694	9.517	2.616	3.081
dna.fr	.8395	.0037	.2661	3.92	7.926	.3166	.3867
lunion.presse.fr	.6888	.0037	.2649	3.887	7.854	.2268	.2611
nicematin.com	.6756	.0031	.2208	3.239	6.564	.088	.0993
republicain-lorrain.fr	.4341	.0026	.1879	2.754	5.576	.083	.0966
sudouest.com	.6468	.0026	.1858	2.721	5.674	.4109	.4515
estrepublicain.fr	.3172	.002	.1424	2.092	4.236	.0079	.0238
lalsace.fr	.4274	.0016	.1119	1.639	3.317	.5654	.611
lamontagne.fr	.3001	.0015	.1066	1.561	3.158	.1821	.2176

%Users: Outlet unique users as a share of all consistent users; %Browsing: Outlet page views as a share of all browsing page views; %News: Outlet PV's as a share of all news PV's, %Local: Outlet PV's as a share of all local news PV's; PVs per user: Page Views per all consistent users; %GNHP Dir: Share of page views directly referred to from Google News Homepage; % GNHP Ov: Share of page views in sessions originating from GNHP

Table 3: User News Concentration

	Local		Non-Local	
	Mean	Std. Dev.	Mean	Std. Dev.
#Outlets	1.67	1.33	10.1	12.8
Top 1	87.8%	20%	55.0%	28%
Top 3	99.0%	4.7%	80.0%	21%
Top 5	99.8%	1.9%	88.3%	16%

User-level average coverage rate of top N outlets of total news page views

Table 4: Matching Dimensions Summary

	Median	Mean	Std. Dev.	Min	Max
Browsing Days Percent Diff.	4.480	8.338	10.71	0	50
Browsing Page Views Percent Diff.	7.012	12.31	13.04	0	50
Google News Homepage Percent Diff.	0	11.22	15.73	0	50
Local News Page Views Percent Diff.	0	6.358	13.88	0	50
Short-term Local Views Percent Diff.	0	3.010	10.43	0	50
Distance (km)	25.07	98.89	122.3	0	399.8
Total Proximity Score	0.0658	0.162	0.204	0	1.185

Table 5: User Variables Summary

	Median	Mean	Std. Dev.	Min	Max
Consistent Users					
Toolbar Usage Days	8.129	8.282	2.809	1.500	15
Total Browsing Page Views	584.8	875.2	853.3	0	3815.7
GNHP Views	0	0.146	1.049	0	33
Local Outlets	0	0.137	0.470	0	4
Non-Local Outlets	1.167	2.730	4.883	0	48
Local News Page Views	0	0.422	1.683	0	13
Non-Local News Page Views	2.365	9.648	23.75	0	174
Local News Dwell Time	0	31.42	137.0	0	1116
Non-Local News Dwell Time	158.4	778.8	2024.4	0	15278
GNHP Referrals to Local News Sites	0	0.00327	0.0391	0	1
GNHP Referrals to Non-Local News Sites	0	0.0512	0.537	0	19
Direct to Local Sites	0	0.0535	0.255	0	2
Direct to Non-Local Sites	0.288	2.164	6.357	0	42
Selected Control Group					
Toolbar Usage Days	10.79	10.49	2.224	3.613	15
Total Browsing Page Views	838.3	1115.2	899.6	0	3815.7
GNHP Views	0.303	4.195	7.860	0	33
Local Outlets	0.129	0.499	0.924	0	4
Non-Local Outlets	3.935	8.824	11.45	0	48
Local News Page Views	0.190	1.316	2.920	0	13
Non-Local News Page Views	9.032	26.88	41.51	0	174
Local News Dwell Time	9.100	105.6	245.6	0	1116
Non-Local News Dwell Time	703.2	2288.1	3630.7	0	15278
GNHP Referrals to Local News Sites	0	0.0862	0.212	0	1
GNHP Referrals to Non-Local News Sites	0.0502	1.943	4.249	0	19
Direct to Local Sites	0	0.149	0.423	0	2
Direct to Non-Local Sites	1.063	5.188	9.980	0	42
Treatment Users					
Toolbar Usage Days	11.20	10.83	2.182	3.613	15
Total Browsing Page Views	808.0	1112.8	918.4	0	3815.7
GNHP Views	2.912	8.518	10.74	0	33
Local Outlets	0.418	0.924	1.202	0	4
Non-Local Outlets	7.248	12.56	13.36	0	48
Local News Page Views	0.776	2.589	3.882	0	13
Non-Local News Page Views	17.42	37.28	46.87	0	174
Local News Dwell Time	59.45	216.8	332.0	0	1116
Non-Local News Dwell Time	1483.4	3320.3	4228.8	0	15278
GNHP Referrals to Local News Sites	0.0452	0.233	0.344	0	1
GNHP Referrals to Non-Local News Sites	0.758	3.989	5.947	0	19
Direct to Local Sites	0	0.295	0.563	0	2
Direct to Non-Local Sites	2.168	6.815	10.86	0	42

Variables are scaled to a per 14 day level to mirror the regression variables.

Table 6: Regression Variables Summary

	Median	Mean	Std. Dev.	Min	Max
Selected Control Group					
Pre GNHP Views	2	7.180	10.26	0	33
Post GNHP Views	0	5.856	9.861	0	32
Pre Local News Page Views	0	1.767	3.634	0	13
Post Local News Page Views	0	1.559	3.531	0	13
Pre Local News Dwell Time	0	137.0	308.0	0	1116
Post Local News Dwell Time	0	122.7	299.5	0	1115
Pre Non-Local News Page Views	12	34.80	49.27	0	174
Post Non-Local News Page Views	10	31.84	47.22	0	167
Pre Non-Local News Dwell Time	926	2978.2	4360.4	0	15278
Post Non-Local News Dwell Time	705	2726.9	4175.6	0	14623
Pre Toolbar Usage Page Views	917.8	1232.3	1012.3	0	3815.7
Post Toolbar Usage Page Views	852	1188.8	1032.6	0	3847.4
Pre Toolbar Usage Days	12.71	11.79	2.713	2	15.13
Post Toolbar Usage Days	12	11.49	2.839	2	15
Pre Local Outlets Per Day	0	0.733	1.242	0	4
Post Local Outlets Per Day	0	0.635	1.204	0	4
Pre Non-Local Outlets Per Day	6	12.08	14.28	0	48
Post Non-Local Outlets Per Day	5	10.93	13.73	0	46
Pre Direct To Local News Sites	0	0.170	0.518	0	2
Post Direct To Local News Sites	0	0.156	0.501	0	2
Pre Direct To Non-Local News Sites	1	6.179	11.38	0	42
Post Direct To Non-Local News Sites	1	5.726	10.78	0	40
Pre GNHP Referrals To Local News Sites	0	0.128	0.334	0	1
Post GNHP Referrals To Local News Sites	0	0.100	0.300	0	1
Pre GNHP Referrals To Non-Local News Sites	0	3.258	5.705	0	19
Post GNHP Referrals To Non-Local News Sites	0	2.653	5.244	0	18
Treatment Users					
Pre GNHP Views	2	7.843	11.03	0	33
Post GNHP Views	3	8.833	11.22	0	32
Pre Local News Page Views	0	1.966	3.883	0	13
Post Local News Page Views	0	2.411	4.182	0	13
Pre Local News Dwell Time	0	156.9	327.9	0	1116
Post Local News Dwell Time	0	192.9	353.3	0	1115
Pre Non-Local News Page Views	16	37.48	49.52	0	174
Post Non-Local News Page Views	18	38.24	48.23	0	167
Pre Non-Local News Dwell Time	1236	3287.0	4444.6	0	15278
Post Non-Local News Dwell Time	1425	3409.6	4441.3	0	14623
Pre Toolbar Usage Page Views	912.5	1231.7	1029.0	0	3815.7
Post Toolbar Usage Page Views	887	1209.9	1019.0	0	3847.4
Pre Toolbar Usage Days	13	11.80	2.867	2	15.13
Post Toolbar Usage Days	13	11.87	2.717	3	15
Pre Local Outlets Per Day	0	0.779	1.294	0	4
Post Local Outlets Per Day	0	0.926	1.370	0	4
Pre Non-Local Outlets Per Day	7	12.69	14.20	0	48
Post Non-Local Outlets Per Day	8	13.33	14.07	0	46
Pre Direct To Local News Sites	0	0.191	0.537	0	2
Post Direct To Local News Sites	0	0.226	0.580	0	2
Pre Direct To Non-Local News Sites	2	6.787	11.50	0	42
Post Direct To Non-Local News Sites	2	6.518	10.90	0	40
Pre GNHP Referrals To Local News Sites	0	0.151	0.358	0	1
Post GNHP Referrals To Local News Sites	0	0.203	0.403	0	1
Pre GNHP Referrals To Non-Local News Sites	0	3.714	6.168	0	19
Post GNHP Referrals To Non-Local News Sites	0	4.066	6.178	0	18

Table 7: Treatment Effect (local news)

		(1)	(2)	(3)	(4)
GNHP Change	Treatment	0.504***	0.484***		
		(0.0204)	(0.0205)		
Local News	Treatment	0.255***	0.212***	0.122***	0.0239
		(0.0188)	(0.0188)	(0.0182)	(0.0228)
	Treatment x Post GNHP				0.0690***
					(0.0136)
Local News Indicator	Treatment	0.962***	0.815***	0.433***	0.359**
		(0.0597)	(0.0676)	(0.0697)	(0.121)
	Treatment x Post GNHP				0.0401
					(0.0541)
Local News Time	Treatment	0.856***	0.676***	0.368***	0.0138
		(0.0587)	(0.0582)	(0.0568)	(0.0716)
	Treatment x GNHP time				0.0879***
					(0.0161)
Direct Navigation Local	Treatment	0.0471***	0.0378***	0.0262***	0.00938
		(0.00702)	(0.00701)	(0.00698)	(0.00879)
	Treatment x Post GNHP				0.0118*
					(0.00533)
GNHP Refers Local	Treatment	0.130***	0.119***	0.0712***	-0.0270***
		(0.00978)	(0.00961)	(0.00914)	(0.00611)
	Treatment x Post GNHP				0.0690***
					(0.00834)
Local Outlets	Treatment	0.162***	0.128***	0.0593***	-0.00732
		(0.0120)	(0.0119)	(0.0114)	(0.0132)
	Treatment x Post GNHP				0.0467***
					(0.00867)
New Local News Pageviews	Treatment	0.113***	0.0962***	0.0611***	-0.0318
		(0.0139)	(0.0139)	(0.0138)	(0.0173)
	Treatment x Post GNHP				0.0651***
					(0.0104)
Old Local News Pageviews	Treatment	0.0364**	0.0506***	0.0269*	-0.0250*
		(0.0112)	(0.0111)	(0.0108)	(0.0125)
	Treatment x Post GNHP				0.0364***
					(0.00853)
Local News Index Ratio	Treatment	-0.193***	-0.158***	-0.0831***	-0.0132
		(0.0147)	(0.0147)	(0.0142)	(0.0158)
	Treatment x Post GNHP				-0.0490***
					(0.0110)
Local News Index Page Views	Treatment	0.0814***	0.0657***	0.0452***	-0.00819
		(0.0135)	(0.0136)	(0.0135)	(0.0178)
	Treatment x Post GNHP				0.0374***
					(0.0104)
Top Local Outlet Pageviews	Treatment	0.0358***	0.0370***	0.0161	-0.0269*
		(0.0108)	(0.0104)	(0.0102)	(0.0118)
	Treatment x Post GNHP				0.0301***
					(0.00813)
Page Views per Local Outlet	Treatment	0.206***	0.165***	0.0962***	0.0289
		(0.0152)	(0.0153)	(0.0148)	(0.0191)
	Treatment x Post GNHP				0.0471***
					(0.0108)

Table 8: Treatment Effect (non-local news)

		(1)	(2)	(3)	(4)
Non-Local News	Treatment	0.306***	0.247***	-0.00735	0.0450
		(0.0258)	(0.0258)	(0.0244)	(0.0403)
	Treatment x Post GNHP				-0.0367*
					(0.0173)
Non-Local News Indicator	Treatment	0.679***	0.542***	0.151	0.173
		(0.0697)	(0.0761)	(0.0812)	(0.0934)
	Treatment x Post GNHP				-0.0504
					(0.107)
Non-Local News Time	Treatment	0.699***	0.556***	0.0482	0.226*
		(0.0547)	(0.0543)	(0.0526)	(0.102)
	Treatment x GNHP time				-0.0440**
					(0.0161)
Direct Navigation Non-Local	Treatment	0.115***	0.0789***	-0.0462*	0.00515
		(0.0207)	(0.0209)	(0.0208)	(0.0291)
	Treatment x Post GNHP				-0.0360*
					(0.0156)
GNHP Refers Non-Local	Treatment	0.305***	0.291***	0.00833	-0.0588***
		(0.0169)	(0.0170)	(0.0119)	(0.0104)
	Treatment x Post GNHP				0.0471***
					(0.00844)
Non-Local Outlets	Treatment	0.256***	0.214***	-0.00856	0.0132
		(0.0183)	(0.0184)	(0.0166)	(0.0272)
	Treatment x Post GNHP				-0.0153
					(0.0121)
New Non-Local News Pageviews	Treatment	0.464***	0.302***	0.0891***	0.126***
		(0.0230)	(0.0229)	(0.0212)	(0.0328)
	Treatment x Post GNHP				-0.0259
					(0.0150)
Old Non-Local News Pageviews	Treatment	0.0765**	0.0271	-0.177***	-0.175***
		(0.0253)	(0.0255)	(0.0256)	(0.0385)
	Treatment x Post GNHP				-0.00121
					(0.0179)
Non-Local News Index Ratio	Treatment	-0.210***	-0.178***	-0.0436*	-0.0778**
		(0.0196)	(0.0198)	(0.0190)	(0.0282)
	Treatment x Post GNHP				0.0240
					(0.0143)
Non-Local News Index Page Views	Treatment	0.154***	0.105***	-0.0251	-0.00801
		(0.0226)	(0.0227)	(0.0226)	(0.0327)
	Treatment x Post GNHP				-0.0120
					(0.0165)
Page Views per Non-Local Outlet	Treatment	0.100***	0.0644***	0.00363	0.0554*
		(0.0143)	(0.0144)	(0.0145)	(0.0243)
	Treatment x Post GNHP				-0.0362***
					(0.0101)

Table 9: Long-Term Treatment Effect (local news)

			(1)	(2)	(3)	(4)
Local News	1-2 Week	Treatment	0.242***	0.220***	0.139***	0.0719
			(0.0328)	(0.0319)	(0.0313)	(0.0401)
		Treat x Post GNHP				0.0500*
						(0.0238)
Local News	3-4 Week	Treatment	0.223***	0.203***	0.127***	0.0318
			(0.0325)	(0.0318)	(0.0305)	(0.0384)
		Treat x Post GNHP				0.0696**
						(0.0231)
Local News	5-6 Week	Treatment	0.173***	0.153***	0.0753*	0.0525
			(0.0318)	(0.0312)	(0.0301)	(0.0371)
		Treat x Post GNHP				0.0166
						(0.0235)
Local News	7-8 Week	Treatment	0.140***	0.123***	0.0430	0.0413
			(0.0318)	(0.0312)	(0.0300)	(0.0373)
		Treat x Post GNHP				0.00127
						(0.0236)
Local News Indicator	1-2 Week	Treatment	0.936***	0.807***	0.491***	0.468*
			(0.104)	(0.115)	(0.119)	(0.207)
		Treat x Post GNHP				0.0127
						(0.0927)
Local News Indicator	3-4 Week	Treatment	0.894***	0.750***	0.484***	0.260
			(0.105)	(0.115)	(0.116)	(0.215)
		Treat x Post GNHP				0.123
						(0.0924)
Local News Indicator	5-6 Week	Treatment	0.806***	0.657***	0.390**	0.603**
			(0.109)	(0.122)	(0.121)	(0.185)
		Treat x Post GNHP				-0.122
						(0.0872)
Local News Indicator	7-8 Week	Treatment	0.592***	0.414**	0.109	0.295
			(0.115)	(0.129)	(0.128)	(0.208)
		Treat x Post GNHP				-0.102
						(0.0944)
Local News Outlets	1-2 Week	Treatment	0.161***	0.140***	0.0783***	0.0263
			(0.0206)	(0.0200)	(0.0192)	(0.0223)
		Treat x Post GNHP				0.0386*
						(0.0151)
Local News Outlets	3-4 Week	Treatment	0.157***	0.137***	0.0788***	0.00410
			(0.0208)	(0.0202)	(0.0191)	(0.0224)
		Treat x Post GNHP				0.0546***
						(0.0146)
Local News Outlets	5-6 Week	Treatment	0.136***	0.115***	0.0559**	0.0355
			(0.0213)	(0.0208)	(0.0198)	(0.0236)
		Treat x Post GNHP				0.0149
						(0.0154)
Local News Outlets	7-8 Week	Treatment	0.0929***	0.0744***	0.0126	0.0111
			(0.0206)	(0.0202)	(0.0191)	(0.0222)
		Treat x Post GNHP				0.00107
						(0.0150)

Table 10: Long-Term Treatment Effect (non-local news)

			(1)	(2)	(3)	(4)
Non-Local News	1-2 Week	Treatment	0.203***	0.184***	-0.0410	0.0152
			(0.0450)	(0.0449)	(0.0430)	(0.0697)
		Treat x Post GNHP				-0.0417
						(0.0306)
Non-Local News	3-4 Week	Treatment	0.199***	0.180***	-0.0297	0.0375
			(0.0488)	(0.0483)	(0.0460)	(0.0724)
		Treat x Post GNHP				-0.0491
						(0.0328)
Non-Local News	5-6 Week	Treatment	0.207***	0.186***	-0.0318	0.0679
			(0.0479)	(0.0480)	(0.0448)	(0.0679)
		Treat x Post GNHP				-0.0728*
						(0.0316)
Non-Local News	7-8 Week	Treatment	0.237***	0.219***	-0.0180	0.101
			(0.0506)	(0.0505)	(0.0467)	(0.0710)
		Treat x Post GNHP				-0.0872*
						(0.0338)
Non-Local News Indicator	1-2 Week	Treatment	0.466***	0.382**	-0.0188	0.0432
			(0.117)	(0.129)	(0.139)	(0.159)
		Treat x Post GNHP				-0.146
						(0.181)
Non-Local News Indicator	3-4 Week	Treatment	0.426***	0.327**	0.0137	0.120
			(0.111)	(0.120)	(0.130)	(0.156)
		Treat x Post GNHP				-0.165
						(0.132)
Non-Local News Indicator	5-6 Week	Treatment	0.493***	0.414***	0.0741	0.233
			(0.109)	(0.118)	(0.127)	(0.150)
		Treat x Post GNHP				-0.238*
						(0.121)
Non-Local News Indicator	7-8 Week	Treatment	0.533***	0.455***	0.0701	0.292
			(0.106)	(0.117)	(0.129)	(0.152)
		Treat x Post GNHP				-0.316**
						(0.120)
Non-Local News Outlets	1-2 Week	Treatment	0.182***	0.167***	-0.0317	-0.00457
			(0.0316)	(0.0313)	(0.0287)	(0.0457)
		Treat x Post GNHP				-0.0202
						(0.0212)
Non-Local News Outlets	3-4 Week	Treatment	0.170***	0.157***	-0.0280	0.0212
			(0.0344)	(0.0339)	(0.0310)	(0.0481)
		Treat x Post GNHP				-0.0360
						(0.0228)
Non-Local News Outlets	5-6 Week	Treatment	0.182***	0.167***	-0.0254	0.0477
			(0.0346)	(0.0345)	(0.0310)	(0.0460)
		Treat x Post GNHP				-0.0533*
						(0.0226)
Non-Local News Outlets	7-8 Week	Treatment	0.199***	0.186***	-0.0212	0.0683
			(0.0367)	(0.0365)	(0.0322)	(0.0476)
		Treat x Post GNHP				-0.0654**
						(0.0239)

Table 11: Robustness Checks (local news)

			(1)	(2)	(3)	(4)
Local News	Exclude Paris	Treatment	0.260***	0.243***	0.148***	0.0925**
			(0.0234)	(0.0231)	(0.0224)	(0.0281)
		Treat x GNHP views				0.0404*
						(0.0174)
	Continuous Data	Treatment	0.226***	0.216***	0.137***	0.0497
			(0.0265)	(0.0263)	(0.0255)	(0.0315)
		Treat x GNHP views				0.0679***
						(0.0202)
	Early Adopters	Treatment	0.265***	0.257***	0.171***	0.111*
			(0.0423)	(0.0424)	(0.0401)	(0.0502)
		Treat x GNHP views				0.0360
						(0.0268)
Local News Indicator	Exclude Paris	Treatment	0.992***	0.842***	0.624***	0.656***
			(0.0728)	(0.0807)	(0.0764)	(0.117)
		Treat x GNHP views				-0.0195
						(0.0600)
	Continuous Data	Treatment	0.908***	0.842***	0.533***	0.505**
			(0.0851)	(0.0965)	(0.0993)	(0.166)
		Treat x GNHP views				0.0166
						(0.0798)
	Early Adopters	Treatment	0.829***	0.779***	0.414**	0.615*
			(0.124)	(0.144)	(0.150)	(0.248)
		Treat x GNHP views				-0.0973
						(0.103)
Local News Outlets	Exclude Paris	Treatment	0.164***	0.148***	0.0755***	0.0348*
			(0.0146)	(0.0143)	(0.0138)	(0.0162)
		Treat x GNHP views				0.0404*
						(0.0174)
	Continuous Data	Treatment	0.144***	0.133***	0.0726***	0.0149
			(0.0170)	(0.0168)	(0.0160)	(0.0182)
		Treat x GNHP views				0.0447***
						(0.0130)
	Early Adopters	Treatment	0.159***	0.148***	0.0795**	0.0476
			(0.0261)	(0.0261)	(0.0244)	(0.0297)
		Treat x GNHP views				0.0190
						(0.0173)

Table 12: Robustness Checks (non-local news)

			(1)	(2)	(3)	(4)
Non-Loc News	Exclude Paris	Treatment	0.302***	0.277***	0.0205	0.0968*
			(0.0322)	(0.0321)	(0.0300)	(0.0490)
		Treat x GNHP views				-0.0560*
						(0.0217)
	Continuous Data	Treatment	0.243***	0.222***	0.00975	0.0529
			(0.0367)	(0.0366)	(0.0348)	(0.0540)
		Treat x GNHP views				-0.0334
						(0.0243)
	Early Adopters	Treatment	0.266***	0.260***	0.0300	0.158
			(0.0570)	(0.0567)	(0.0537)	(0.0977)
		Treat x GNHP views				-0.0764*
						(0.0367)
Non-Loc News Indicator	Exclude Paris	Treatment	0.708***	0.676***	0.291**	0.336**
			(0.0844)	(0.0902)	(0.0971)	(0.111)
		Treat x GNHP views				-0.109
						(0.131)
	Continuous Data	Treatment	0.582***	0.564***	0.244*	0.296*
			(0.0929)	(0.0993)	(0.106)	(0.120)
		Treat x GNHP views				-0.129
						(0.140)
	Early Adopters	Treatment	0.646***	0.688***	0.283	0.279
			(0.174)	(0.184)	(0.188)	(0.215)
		Treat x GNHP views				0.00866
						(0.250)
Non-Loc News Outlets	Exclude Paris	Treatment	0.261***	0.244***	0.0203	0.0621
			(0.0227)	(0.0227)	(0.0205)	(0.0332)
		Treat x GNHP views				-0.0307*
						(0.0153)
	Continuous Data	Treatment	0.207***	0.193***	0.00739	0.0150
			(0.0259)	(0.0258)	(0.0234)	(0.0363)
		Treat x GNHP views				-0.00590
						(0.0168)
	Early Adopters	Treatment	0.212***	0.210***	0.00583	0.0926
			(0.0399)	(0.0397)	(0.0349)	(0.0631)
		Treat x GNHP views				-0.0517*
						(0.0245)

Table 13: Treatment Effect: Local News Page Views

	(1)	(2)	(3)	(4)
Treatment	0.255***	0.212***	0.122***	0.0239
	(0.0188)	(0.0188)	(0.0182)	(0.0228)
Log Pre Local News Page Views	0.527***	0.485***	0.482***	0.481***
	(0.0172)	(0.0189)	(0.0187)	(0.0187)
Log Pre Google News Homepage Visits		0.0523***	-0.0932***	-0.0923***
		(0.00604)	(0.00745)	(0.00747)
Log Pre Toolbar Usage Page Views		0.00708	0.00958*	0.00969*
		(0.00476)	(0.00469)	(0.00469)
Log Pre Toolbar Usage Days		0.0355*	0.0574***	0.0569***
		(0.0169)	(0.0165)	(0.0165)
Log Pre National Searches		0.0477***	0.0338**	0.0356**
		(0.0110)	(0.0109)	(0.0109)
Log Pre Local Searches		0.542***	0.542***	0.545***
		(0.0406)	(0.0408)	(0.0405)
Log Pre Other Searches		0.00115	-0.0117	-0.0124
		(0.0142)	(0.0139)	(0.0139)
Log Post Google News Homepage Views			0.186***	0.180***
			(0.00624)	(0.00629)
Treatment x Post GNHP				0.0690***
				(0.0136)
Constant	0.132***	-0.0559	-0.112**	-0.108**
	(0.00477)	(0.0361)	(0.0351)	(0.0351)

Figure 4: Log page views (local news)

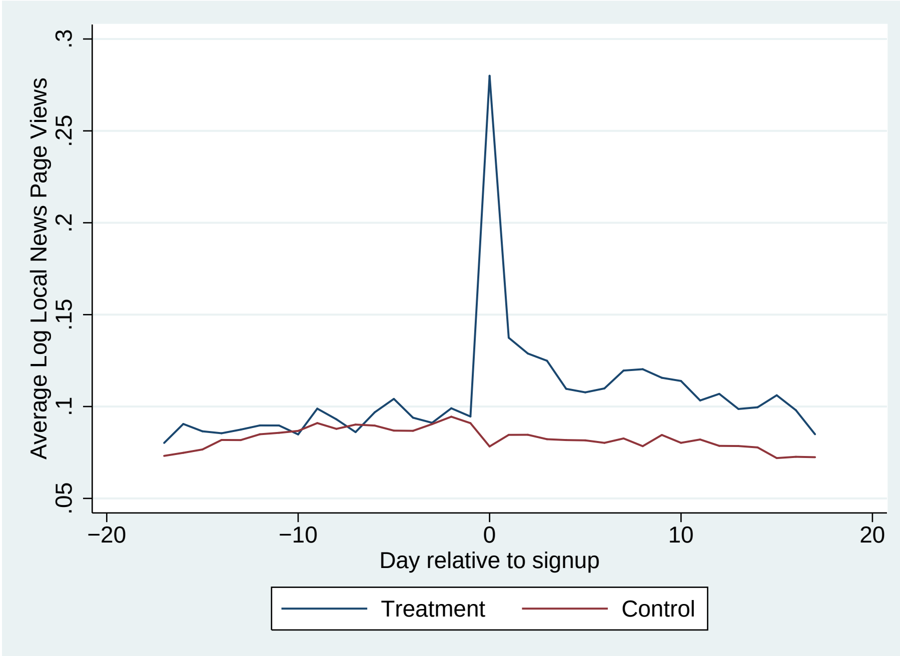

Figure 5:

Figure 6: Log page views (non-local news)

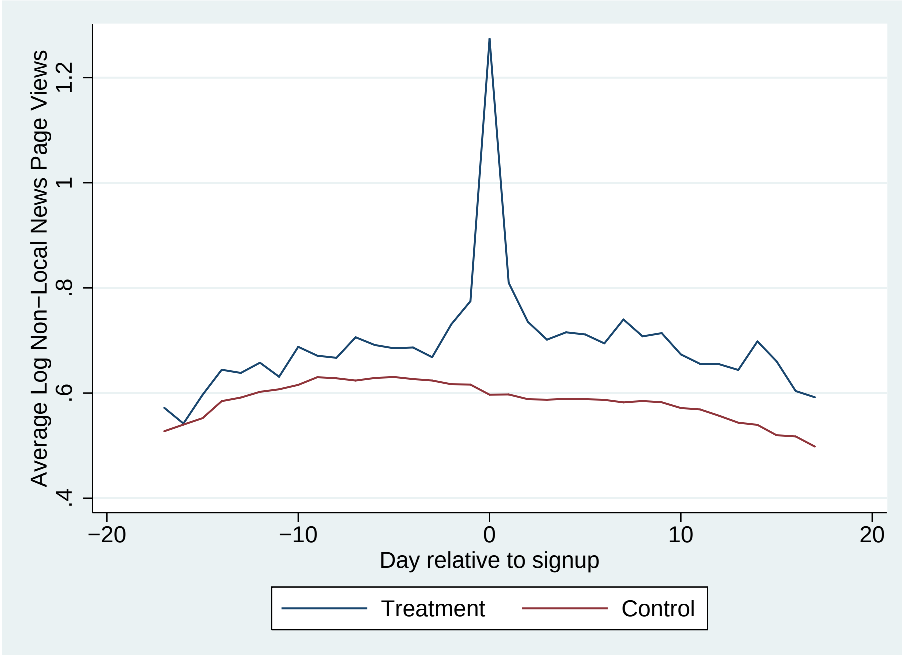

Figure 7:

B Data Appendix

B.1 French News List Construction

We program a scraper to extract the URL's of French News Outlets using Wikipedia category information. The program starts with a set of category page URL's and extracts the HTML code associated with these pages. On each category page, there is a list of subcategories and a list of Wikipedia concept pages associated with the category. We recursively open subcategory pages, and extract all the associated concept page URL's. Next, we search the HTML of each concept page for the official website URL. We clean these URL's to domain form (e.g. lemonde.fr) and add it to our list.

The categories that we start with are:

Presse quotidienne nationale française, Presse écrite française par région, Presse écrite française par département, Presse quotidienne régionale française, Presse mensuelle régionale française, Presse hebdomadaire française locale, and Presse bimestrielle française locale.

We mark the URL's under the regional categories as local. To confirm that the local definition is appropriate for each outlet, we check the geographical distribution of its user base using our browsing data. For each outlet, we compute the Herfindahl index for page views normalized by population at the department level. We drop any outlets in our list that have a Herfindahl index less than 0.1. This filtered out a few small news sites that were hosted on shared large domains.

As another check of our list, we also visited the French GNHP and obtained the local news section for several large French cities. The URL's generated from Wikipedia covered all of the sites that had stories in the local news section for each city.

Finally, we went through the list of the most visited domains in our news list, and added extra URL filters to avoid counting non-news URL's as news. For example, orange.fr hosts its own news, but only pages beginning with actu.orange.fr are news articles.

B.2 Index Page Identification

To create the index page view variable, we first have to determine which URL's are index pages as opposed to article pages. Websites have all sorts of different URL formats for its index pages, so we took a data-driven approach rather than a pattern matching one.

For each domain and month, we look at URL's that are visited in all of the days in that month. Then, we sort these URL's by the infimum over all the days in terms of page views. We then keep the top 10 of these most consistently visited URL's for each domain and month and label these as index pages. The reasoning for looking over all the months is to capture any changes in index page naming conventions by these news websites.

While the resulting list might not cover all of the index pages, it should give us a fairly good representation of the most visited index pages. We manually checked this list for several domains and found it to be very accurate.