Digitization and Aggregation

Lesley Chiou* and Catherine Tucker‡

November 21, 2013

Abstract

The digitization of content has led to the growth of platforms that draw news and information from multiple sources. Policy makers are concerned that these new platforms threaten incentives for the production of original content. As a result, policymakers in Europe are contemplating regulations that would force such aggregators to pay each time they aggregate content. To understand the possible consequences and underlying rationale of such laws, we explore whether aggregation of content by a single platform encourages users to "skim" or to investigate content in depth, and by extension whether users use aggregators to substitute or complement original content. We exploit a contract dispute that led a major aggregator to remove information from a content provider. We find that after the removal, users were less likely to investigate additional content in depth, particularly sources that were horizontally or vertically differentiated.

JEL classification: L86
keywords: digitization, information, consumer search, Internet

*Economics Department, Occidental College, CA

†MIT Sloan School of Management, MIT, Cambridge, MA and National Bureau of Economic Research.

‡An earlier version of this paper was circulated under the title "News and Online Aggregators." We thank Robert Seamans useful comments. We thank Christopher Hafer of Experian Hitwise. We also thank Cassandra Crosby and Sara Mcknight for excellent research assistance. Financial support from the NBER Innovation Policy Group and NET Institute (www.NETinst.org) is gratefully acknowledged.

1 Introduction

The digitization of content has led to the rise of online platforms who bring consumers into contact with an intentionally diverse set of information sources. The long-run implications for content markets of these aggregator platforms depends on whether consumers use these sources as substitute or complements to each other. Most empirical issues in platforms depends crucially on whether different participants stand in a complementary relationship to one another or not. In practice, this open question is an empirical one, as revealed by how users behave.

Digitization and the growth of the Internet have led to an acceleration in platforms and consumer search. Consumers now have a nearly infinite, searchable, and reproducible storehouse of content that they can access relatively quickly and cheaply online. From 26 million pages in 1998, the Internet includes more than 1 trillion webpages today (Alpert and Hajaj, 2008). In response to the explosion of information, online platforms or aggregators such as Google News, Everyblock, and Gawker gather and consolidate information from multiple sources and display it on a single site.

In particular, the importance of platforms can be seen most clearly in media industries in the wake of the digital age. The media industry provides an interesting opportunity to examine the role of platforms. Consumer behavior may drive incentives to produce original content and affect the survival of newspapers as well as the success of advertising in the industry. These factors have far-reaching implications for the growth or demise of media and its industry structure and concentration.

In the media industry, how consumers use platforms has been an extremely controversial issue. Producers of content fear that consumers may use these extracts of content as a substitute for accessing and reading the full content. Aggregators argue that they encourage users to seek additional content by featuring multiple sources. The conflict reflects underlying

several questions. Are consumers using aggregators to reduce search costs and terminate their search for content that they would already seek, or are consumers using aggregators to seek new content that they would not otherwise obtain? What relationship exists between different content sources on a platform?

The theoretical literature has modeled these two countervailing forces of "traffic" and "scanning." These effects have been captured in theoretical models of platforms and aggregators (George and Hogendorn, 2012). For instance, Jeon and Esfahani (2012) describes "market expansion" and "business-stealing" effects of aggregators while Rutt (2011) examines a model with two types of consumers—those that are "loyal" to original content sites and those that are "searchers" who use aggregators. More generally, the effects can be described as to what extent content on platforms as well as aggregators are "complements" and "substitutes" (Athey et al., 2011). These models all reflect the possibility that consumers may use aggregators to go more in depth ("traffic effect") or in lieu of content sites ("scanning effect.")

The predictions of the models depend starkly on which effect dominates. The results have implications for pricing, the choice of quality in content, the investment in primary content, and market concentration. Despite the growing theoretical literature, empirical evidence is absent. We tease apart these empirical effects by exploiting a natural experiment in the provision of content on a major news-feed aggregator, Google News. We exploit a contract dispute between Google News and a content provider as a discontinuous shift in the provision of copyrighted content by an aggregator. In January 2010, after a breakdown in licensing negotiations, Google removed all news articles by The Associated Press from its news aggregator (Haddad, 2010). We compare users' website visits before and after this policy change relative to traffic from Yahoo! News, which continued to provide Associated Press content during this period. Our results indicate that after Associated Press content was removed from Google News, fewer users subsequently visited other news sites after

navigating to Google News relative to users who had used Yahoo! News. We check the robustness of the result in a variety of ways.

The readership of Google News is too small to permit estimates of how aggregation affects independent visits to the content providers' websites—Sandoval (2009), Arrington (2010), and Athey and Mobius (2012) discuss this case. Instead, we measure how a platform's expansion or contraction of content affects navigation by users from that platform. Our results suggest that users of aggregators visit content websites after visiting an aggregator. In other words, users do not view an aggregator as a perfect substitute for content. When users encounter content summarized by an aggregator, they are more likely to be provoked to seek additional sources and read further rather than merely being satisfied with a summary.

We also examine how the policy change affected different types of information content. Our results suggest that websites with either a very national or very local audience suffered the steepest decline in downstream visits after the removal of other content. We argue that this is evidence that aggregation guides users to content that is either vertically differentiated, such as nationally recognized sites with acclaimed standards of quality, or horizontally differentiated, such as local sites that would not otherwise find a broad audience. Our results suggest that aggregation appears to inspire people to seek new content (of a more unusual and high quality).

Our analysis is also related to prior work that describes how different technologies from the Information and Communications Technology (ICT) Revolution have affected search costs and generated spillovers. Shapiro and Varian (1999) present a general model of reduced search costs online. Bakos (1997) examines technologies within the electronic marketplaces, and Ghose et al. (2011) study the mobile Internet. Greenstein (2011) examine spillovers from the adoption of broadband technology. The novelty of our study is that we are the first to explore how an ICT technology affects the set of information gathered by consumers.

Our results also illuminate the implications of intellectual property and copyright on-

line. Many firms rely on intellectual property and copyright to protect their intellectual assets. However, the digital revolution has challenged various aspects of copyright protection (Greenstein et al., 2011). Online aggregators assert that their practice is protected by copyright law because they only display small extracts of information and often this information is factual (Isbell, 2010). To understand the consequences of digitization for copyright, it is therefore important to study how consumers use online aggregators to acquire content. Thus far, most of the literature has centered on digitization and piracy within the music industry (Rob and Waldfogel, 2006; Oberholzer-Gee and Strumpf, 2007; Danaher et al., 2010) or on the use of trademarks (Bechtold, 2011; Chiou and Tucker, 2011). We focus on digitization and the reproduction of content for information more generally. Our results suggests that producers of primary content may actually benefit from relaxing their restrictions on copyright and by allowing others to disseminate their content, particularly if it is either a niche or a high-quality offering.

2 Data and Institutional Setting

2.1 Contractual Dispute between Google and The Associated Press

Two prongs of copyright law make aggregation potentially permissible. The first is the notion that aggregators are simply collecting "facts" rather than original works of creative expression. However, case law suggests that such arguments are unlikely to prevail. In the decision of *Feist Publications, Inc. v. Rural Telephone Service Co., Inc.*, the Supreme Court noted that any spark of creativity, "no matter how crude, humble or obvious it might be" qualified an original work for copyright protection.[1] The second is the notion of "fair use." Since aggregation services are commercial services, the key question centers on the "effect of the use upon the potential market for or value of copyrighted work." Current empirical knowledge on this topic is based upon a survey of 2,787 consumers of whom 44 percent

[1] 499 U.S. 340, 345 (1991).

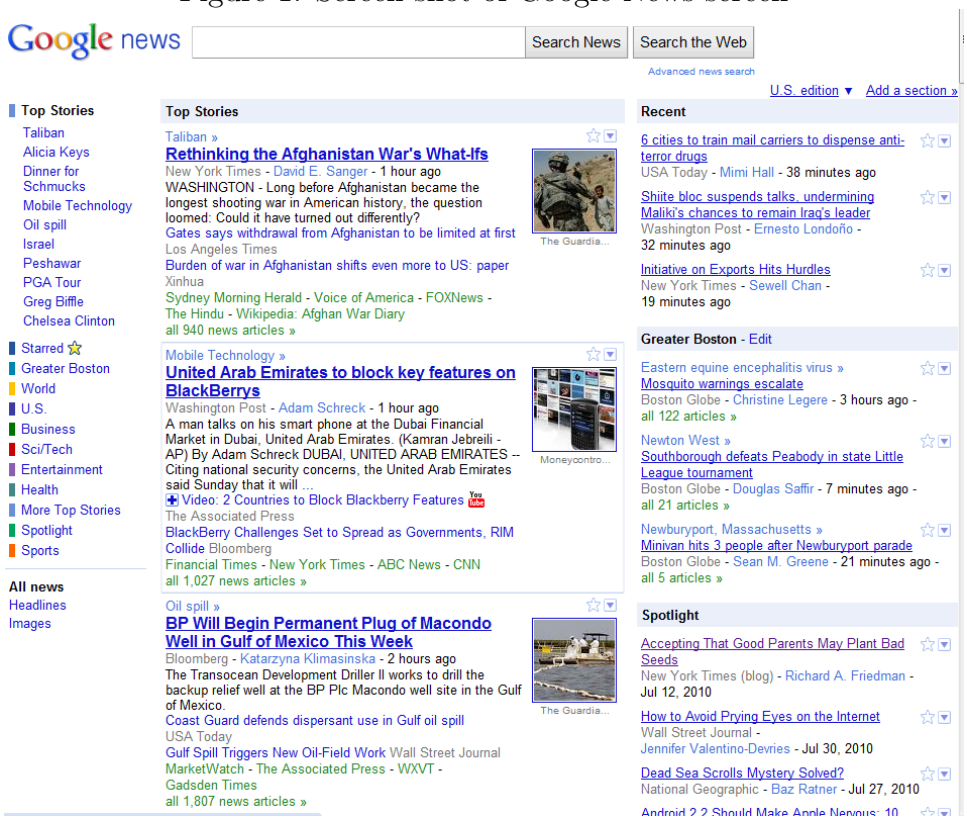
claimed to never click on a link when using Google News (Doctor, 2009). Our study aims to address these questions by providing empirical analysis on usage that is based on actual data browsing patterns alongside exogenous shifts in the availability of copyrighted content.

Google News is ranked as the fifth most visited news website by Hitwise. Receiving 2.90% of all news site visits, it is the second most popular news aggregator service after Yahoo! News, which received 7.09% of all news site visits. Founded on April 2002, Google News electronically aggregates different news sources based upon a proprietary algorithm. As of December 2009, Google News claimed that it received news content from 25,000 publishers across the world and that it sent one billion clicks to these publishers every month (Cohen, 2009). Figure 1 provides a screenshot of Google News. Google News has two noticeable features that distinguishes it from traditional news sites. First, a variety of sources are listed for each story. Second, the order of news is electronically determined based upon users' preferences, the recency of the story, and the interest it has received from other users.

The Associated Press (AP), founded in 1846, is one of the largest news agencies in the world. Since the demise of United Press International, it is the only national news service in the US, and its major competitors are Reuters (based in the United Kingdom) and Agence-France Presse (based in France). The Associated Press is a cooperative owned by various newspapers and radio and television stations in the United States. These stakeholders both contribute stories to The Associated Press and use material written by AP staff journalists. During the past decade, The Associated Press has been at the forefront of efforts by copyright holders to circumscribe "fair use" for digital content and to protect copyholders' rights. For example, in June 2008, The Associated Press invoked the Digital Millennium Copyright Act and insisted that various bloggers remove AP content (Ardia, 2008).

Since both The Associated Press and Google News are key players in the distribution of news online, it is not surprising they have forged a partnership. This licensing agreement also protect Google News from suit for copyright infringement given the current level

Figure 1: Screen shot of Google News screen



Notes: On June, 30 2010, the formatting of Google News changed somewhat and reduced the ability of users to customize the placement of the columns containing news. Therefore the screenshot above, which was produced after this formatting change, may be slightly different from what users viewed during the period that we study.

of uncertainty over the implications of current copyright law for news aggregators. Table 1 summarizes the major events of their relationship. We study a discontinuity in this relationship, which was engendered by negotiations surrounding the contract renewal at the end of January 2010. As part of their existing contract, Google and The Associated Press agreed that AP content could be hosted by Google for a period of 30 days. Therefore, if the contract ended in January 2010 and was not renewed, Google would stop posting new Associated Press content 30 days prior to the end of the contract. Presumably to make this "clean break" a credible outside option, Google did indeed stop posting content for seven weeks during these contract negotiations (Krazit, 2010). We should emphasize that our dis-

cussion is necessarily based upon the observations of industry outsiders, since both Google and The Associated Press signed binding non-disclosure agreements, which prevented them from ever commenting on the course or outcome of negotiations (Sullivan, 2010).

The removal of The Associated Press content represents a useful natural experiment. Since the removal of content was provoked by the intricacies of contract negotiation, its timing can be thought of as reasonably exogenous, as it was determined by the expiration of the contract rather than any considerations of the popularity (or lack thereof) for The Associated Press content at that time. As detailed in Table 1, the dispute with The Associated Press led Google to remove content by The Associated Press starting on December 23, 2009 until sometime in February 2010. Fortunately for our purposes, Yahoo! News continued to host The Associated Press content without interruption during this time, which enables us to use the behavior of Yahoo! News users as a control in our regressions. We compare which websites consumers navigated to after visiting a news aggregator (either Google News or Yahoo! News) before and after the removal of content on Google.

Table 1: Timeline of negotations between Google and The Associated Press

Date	**Event**
August 2006	Google and The Associated Press first sign contract to enable The Associated Press content to appear on Google News for 30 day window.
December 24, 2009	The Associated Press content no longer appears on Google. Industry press speculates that this is in preparation for the expiration of contract between The Associated Press and Google in one month's time.
End January 2010	The Associated Press and Google contract set to expire.
February 2010	The Associated Press content returns to Google News.

It is not clear whether the removal of content will lead aggregator users to seek more or less news after visiting the aggregator. In essence, this depends on whether consumers view

news aggregators as a complement or substitute to original news sources. Do consumers use news aggregators to identify news stories that they then pursue in greater depth, or do they simply stop after reading the first news item? For instance, The Associated Press ran a news story about economic depression in Michigan in August 2010. The screenshot of how the story appeared on Google News is depicted in Figure 2. The links related to The Associated Press story that appear at the bottom of a typical story are also depicted in Figure 2. After reading The Associated Press summary of the story, readers are free to explore the issue further in local newspapers such as the Detroit News and Lansing State Journal. We ask whether the presence of The Associated Press content on Google News makes it more or less likely that a news consumer would then trouble to visit Detroit News or the Lansing State Journal, both of which are members of The Associated Press Network.

Our analysis focuses on the period immediately prior to and during the removal of The Associated Press articles from Google News for two reasons. First, it is not immediately clear at which point in February that Google News and The Associated Press resumed their relationship and reached a new agreement. Second, it is not apparent whether the reinstatement during this time consisted of the older, missing content or new content or whether Google changed the presentation of Associated Press articles afterwards. For example, it would be problematic if Google decided to highlight The Associated Press content after the contract negotiations were concluded, perhaps as a "sweetener" to the deal. For these reasons, we focus on visits to news sites during the months of December 2009 and January 2010.

2.2 Data Description

Our data derive from Experian Hitwise. Hitwise "develops proprietary software that Internet Service Providers (ISPs) use to analyze website logs created on their network." Once the ISP aggregates the anonymous data, the data are provided to Hitwise. According to their website, Hitwise collects these usage data from a "geographically diverse range of ISP networks and

Michigan voters search for economic savior

By KATHY BARKS HOFFMAN (AP) – 1 day ago

LANSING, Mich. — Michigan voters frustrated over lost jobs, home foreclosures and budget deficits will vote in Tuesday's primary election for leaders they hope can move the state out of its economic morass.

With seven men running for governor and nearly two dozen candidates running for three open congressional seats, the hardest task may be sorting through the barrage of names, campaign ads and economic rhetoric.

The candidates and voters agree that Michigan is at a crossroads. After a decade of malaise that has left the state with the nation's second-highest unemployment rate and one in every four residents relying on unemployment insurance, Medicaid, cash assistance or food stamps, creating more jobs is the overwhelming priority and topic of debate.

The gubernatorial candidates are competing to succeed outgoing Democratic Gov. Jennifer Granholm, who can't run again because of term limits and whose popularity sank with her struggles to revive the economy.

All seven gubernatorial candidates say they plan to cut business taxes to attract employers. Most of the five Republicans also say they would slash state regulations and cut state spending. One, Oakland County Sheriff Mike Bouchard, proposes getting rid of laws forcing workers to join unions to get certain jobs.

Among the Democrats, Lansing Mayor Virg Bernero is visiting factory gates and union halls to pledge he'll stand up for middle-class workers and jobs. His opponent, Andy Dillon, a business turnaround specialist who's now the House speaker, promises to bring in more alternative energy jobs to replace lost manufacturing work.

With platforms that are similar, the Republicans are using their job credentials to assure voters they would be the best at managing the economy.

National GOP interest in unseating freshmen Democratic Reps. Mark Schauer in mid-Michigan's 7th District and Gary Peters in the Detroit suburbs in Oakland County has Republicans vying in both districts for the chance at a November matchup.

Copyright © 2010 The Associated Press. All rights reserved.

Related articles

Gov candidates woo undecided voters during final weekend before primary
The Detroit News - 7 hours ago

Race to get names on November ballot in Michigan governor's race is wide open
Lansing State Journal - 18 hours ago

Editorial: Focus on jobs, not abortion in governor race
The Detroit News - 21 hours ago

More coverage (1) »

Associated Press

President Barack Obama addresses employees at the Chrysler's Jefferson North Assembly Plant in Detroit, Friday, July 30, 2010. (AP Photo/Carlos Osorio)

Map

Figure 2: Example screenshot of The Associated Press article hosted on Google News
Notes: Google News, August 1, 2010. Text of article has been slightly edited to fit on page.

opt-in panels, representing all types of Internet usage, including home, work, education and public access." Currently, Hitwise has usage data from a sample of 25 million people worldwide. We include further details on Hitwise's data collection in the Appendix.

Hitwise provides aggregate information on the sites that users visit immediately after navigating to Google News or Yahoo! News. We use weekly data on the top 1500 sites navigated by consumers after visiting Google News or Yahoo! News during the week ending

December 5, 2009 to the week ending January 30, 2010. Hitwise reports the fraction of total traffic that arrives at these "downstream" sites immediately after a visit to Google News and Yahoo! News. We constructed a panel of the percentage of weekly visits a downstream website received from either Google News or Yahoo! News. For instance, we observe the weekly share of visits that nytimes.com receives out of all visits to websites by users immediately after using Google News. In our sample, twenty-six percent of websites received incoming traffic from both Google and Yahoo! News. The remainder of websites were only visited after navigating to one particular aggregator. This pattern may reflect internal complementarities for these companies. For instance, someone using Google News is unlikely to navigate to Yahoo! Mail, and similarly, someone using Yahoo! News is unlikely to navigate to Gmail.

We categorized the websites into two main classes: "news" (e.g., newyorktimes.com, bostonherald.com) and "non-news" (e.g., Yahoo! Mail, myspace.com). As we are interested in traffic to websites of primary news sources, we exclude weather sites and the top aggregators—Yahoo! News, Google News, AOL News, Bing News, Ask News, and Huffington Post—from the "news" category. In addition, we use Hitwise's identification of non-US domains to exclude international sites (e.g., bbc.com/news, hindustantimes.com) from the "news" category, since we do not expect the removal of The Associated Press content to affect international sites that tend to either generate their own content or rely on non-American news agencies for their content. We use data on international sites in our robustness checks. Given the set of "news" sites, we refer to all other sites within our sample as "non-news."

Table 2 reports the summary statistics for our data. News sites represent 20 percent of all sites where we observe subsequent visits within our sample, and non-news sites account for 80 percent. Aggregator, international, and weather sites account for a smaller fraction of sites compared to news sites.

Table 3 displays the top 50 news websites in our dataset and the average percentage of

Table 2: Summary statistics for downstream websites from Google News and Yahoo! News

	Mean	Std Dev	Min	Max	Observations
% visits	0.016	0.19	0	18.3	98730
Google News	0.50	0.50	0	1	98730
Yahoo! News	0.50	0.50	0	1	98730
PeriodDispute	0.67	0.47	0	1	98730
News Site	0.20	0.40	0	1	98730
Non-news Site	0.80	0.40	0	1	98730
Aggregator Site	0.00091	0.030	0	1	98730
International Site	0.048	0.21	0	1	98730
Weather Site	0.0067	0.081	0	1	98730
Observations	98730				

Notes: This table reports statistics for websites visited immediately after Google News and Yahoo! News during December 2009 and January 2010. The variable *%visits* refers to the percentage of visits from each search engine that navigated to a particular site. The dispute between The Associated Press and Google News occurred after December 23, 2009. The variable *PeriodDispute* is an indicator variable for whether the week occurred during the period of the dispute. News sites refer to news and media sites as defined by Hitwise, excluding weather sites, international news sites, and news aggregators from the top 5 search engines.

downstream visits they received from either Google News or Yahoo! News. Downstream visits refer to the number of visits to a website immediately after navigating to the news aggregator. Table 4 displays the top 50 non-news websites in our dataset, excluding international news sites, and the average percentage of downstream visits they receive. As shown in Table 4, the top non-news websites reflect the top website brands on the Internet.

To verify that Yahoo! News could be considered an appropriate control group for Google News, we checked that the users shared similar observable demographics. Hitwise reports the fraction of users within each demographic category for a particular site. As seen in Table A-1 in the appendix, the users of Yahoo! News and Google News do indeed look reasonably similar; they are skewed towards being older, predominantly male, and wealthier than the general U.S. population. For comparison, we also report demographics for users of the New York Times website. The users of the New York Times site are similar, though significantly older, than the average users of a news aggregator. Table A-1 also provides suggestive

Table 3: Top 50 news websites visited after Google News and Yahoo! News

	Avg Visit Pct
abcnews.com	2.11
associatedcontent.com	0.11
bleacherreport.com	0.17
bloomberg.com	0.51
boston.com	0.24
bostonherald.com	0.19
businessweek.com	0.15
cbsnews.com	0.19
celebrity-gossip.net	0.063
chron.com	0.13
cnn.com	1.85
csmonitor.com	0.15
dallasnews.com	0.11
drudgereport.com	0.64
edition.cnn.com	0.20
examiner.com	0.65
foxnews.com	1.13
freep.com	0.13
gather.com	0.34
latimes.com	0.48
mcclatchydc.com	0.095
mercurynews.com	0.44
miamiherald.com	0.15
msnbc.com	0.83
news.com	0.12
nj.com	0.11
npr.org	0.16
nydailynews.com	1.59
nypost.com	0.26
nytimes.com	2.88
pcworld.com	0.18
people.com	0.39
philly.com	0.15
politico.com	0.53
radaronline.com	0.060
reuters.com	0.69
seattlep-i.nwsource.com	0.11
seattletimes.nwsource.com	0.11
sfgate.com	0.17
sportsillustrated.cnn.com	0.10
thedailybeast.com	0.17
theweek.com	0.14
time.com	1.16
upi.com	0.093
usatoday.com	0.72
usmagazine.com	0.23
usnews.com	0.082
voanews.com	0.13
washingtonpost.com	1.74
wired.com	0.083
wsj.com	0.86

Table 4: Top 50 Non-news websites visited after Google News and Yahoo! News

	Avg Visit Pct
address.yahoo.com	0.12
amazon.com	0.59
aol.com	0.46
aralifestyle.com	0.14
ask.com	0.19
autoinsurance.lowermybills.com	0.091
bankofamerica.com	0.18
bing.com	0.62
blogsearch.google.com	0.77
buzz.yahoo.com	0.21
chase.com	0.14
cosmos.bcst.yahoo.com	0.95
ebay.com	1.00
education.yahoo.net	0.34
espn.com	0.56
facebook.com	6.23
fastflip.googlelabs.com	3.60
finance.google.com	0.36
finance.yahoo.com	0.60
games.yahoo.com	0.099
gmail.com	1.55
google.com	11.6
howlifeworks.com	1.04
huffingtonpost.com	0.96
images.google.com	0.50
latimesblogs.latimes.com	0.16
livescience.com	0.38
mail.live.com	1.28
mail.yahoo.com	9.94
maps.google.com	0.23
members.yahoo.com	0.29
movies.yahoo.com	0.13
msn.com	1.03
my.yahoo.com	0.67
myspace.com	1.54
news.google.com	0.24
omg.yahoo.com	0.32
rivals.com	0.10
search.yahoo.com	2.20
shine.yahoo.com	0.13
space.com	0.15
sports.yahoo.com	0.26
tmz.aol.com	0.20
tv.yahoo.com	0.12
video.google.com	0.27
weather.com	0.67
weather.yahoo.com	0.39
wikipedia.org	0.50
yahoo.com	7.20
youtube.com	2.47

evidence of why the debate over ad revenues from news content is so contentious. These readers are a remarkably attractive demographic group from an advertiser's perspective.

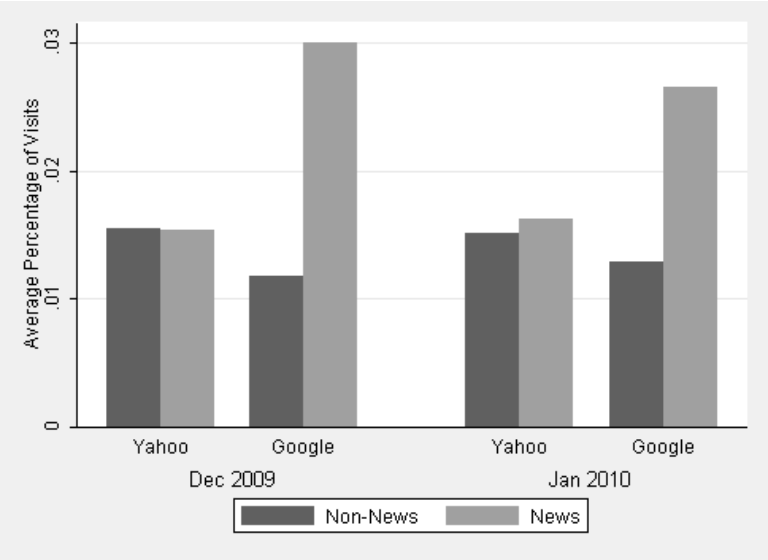
3 Analysis

3.1 Downstream Traffic after Visiting an Aggregator

We examine how digital tools that permit content aggregation affect their users' search for information. Theoretically, the effect can go in either direction. On one hand, the removal of content may raise the costs of information acquisition, and users may be less likely to subsequently pursue further information. On the other hand, if users rely solely on the abbreviated descriptions of the article without pursuing the original content or if they instead substitute towards other content on the aggregator, then the content removal will not affect users' subsequent search (Kaplan, 2010). Furthermore, it is not obvious how different types of content may encourage users to seek further information.

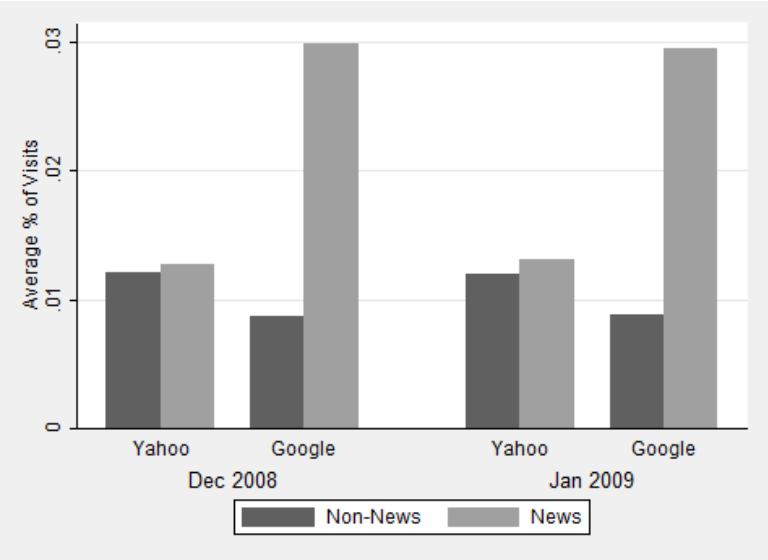
We start with an overall analysis of aggregate behavior before examining the heterogenous effects on different websites. Figure 3 illustrates the aggregate mean percentage of downstream traffic for users that visited Google News and Yahoo! News during our period. As seen in the graph, little change occurs in downstream site navigation for Yahoo! However, news sites experience a decline in visits from Google News after the removal of The Associated Press relative to the change in traffic from Yahoo! News. To investigate whether this pattern could be due to underlying seasonality in news consumption, we examine the change in visits in the prior year during the same calendar months. As expected, Figure 4 illustrates that no such change in visits occurred between December 2008 and January 2009.

Figure 3: Downstream sites visited after Google News and Yahoo! News



Notes: This figure shows the average percentage of visits to news and non-news sites after users visited Google News and Yahoo! News before and after the removal of The Associated Press from Google News in December 2009 and January 2010.

Figure 4: Downstream sites visited after Google News and Yahoo! News in prior year (December 2008 and January 2009)



Notes: This figure shows the average percentage of visits to news and non-news sites after users visited Google News and Yahoo! News in December 2008 and January 2009 for the year prior to the removal of The Associated Press from Google News.

To formalize the insights provided by Figure 3, we run a difference-in-differences regression for the policy change and estimate the following regression for the percentage of visits to website i after visiting news aggregator j in week t:

$$\begin{aligned} \%visits_{ijt} &= \beta_0 + \beta_1 News_i \times Google_j \times PeriodDispute_t + \beta_2 News_i \times PeriodDispute_t \\ &+ \beta_3 News_i \times Google_j + \beta_4 Google_j + \alpha_i + week_t + \epsilon_{ijt} \end{aligned}$$

where $News$ is an indicator variable equal to 1 if the website is a news site, $Google$ is an indicator variable equal to 1 if the traffic originated after viewing Google News, and $PeriodDispute$ is an indicator variable equal to 1 for the weeks after the removal of The Associated Press from Google News. The controls α are downstream-website fixed effects. The vector $week_t$ contains weekly fixed effects to capture national variation in the volume and interest generated by news stories in that week. The coefficient β_1 on the interaction term $News \times Google \times PeriodDispute$ captures the effect of The Associated Press removal on visits to news sites compared to non-news sites from Google News with the corresponding change in news and non-news sites on Yahoo! as a control. We estimate this specification using ordinary least squares and cluster our standard errors at the website level to avoid the downward bias reported by Bertrand et al. (2004).

Table 5 reports the results in column (1) for our full specification as described by equation (1). The negative coefficient on $News * Google * PeriodDispute$ implies that during the dispute with The Associated Press, Google News users were less likely to visit news websites after visiting Google News. This suggests that the presence of The Associated Press articles in Google News prompted users to seek further information at news sites. More generally, our results suggest that news aggregators may complement the news sources that they feature by directing traffic to these news sites.

News sites on Google experience a 0.006 percentage point decrease in visits after the

removal of The Associated Press articles. Compared to the mean percentage share of 0.029 percent before the policy change, this drop represents an approximately 20 percent decrease in traffic to news sites after the removal of The Associated Press articles from Google. If the claim in Cohen (2009) is true that Google sends a billion visits each month to its partner news providers, then this percentage translates into a very large change in the number of visits that news websites receive. While we do not know precisely the international breakdown, our data from Hitwise suggest that 40 percent of all visits before the policy change went to news media websites hosted in the US for the subset of users who use Google News. Therefore, this 20 percent decrease could imply a 80 million decrease in visits each month from Google News users each month to news media websites hosted in the US.

Table 5: Downstream traffic from Google and Yahoo! News before and after the policy change

	(1)	(2)	(3)	(4)
PeriodDispute × Google × News	-0.00600**	-0.00599**	-0.00623**	-0.00622**
	(0.00289)	(0.00289)	(0.00305)	(0.00305)
PeriodDispute × Google	0.00159	0.00158	0.00182	0.00181
	(0.00233)	(0.00234)	(0.00253)	(0.00253)
PeriodDispute	-0.000411	-0.000396	-0.000497	-0.000481
	(0.00110)	(0.00110)	(0.00115)	(0.00115)
Google	-0.0119*	-0.0117*	-0.0156**	-0.0155**
	(0.00638)	(0.00640)	(0.00696)	(0.00698)
PeriodDispute × News	0.00135	0.00134	0.00137	0.00136
	(0.00102)	(0.00102)	(0.00104)	(0.00104)
News × Google	0.0326***	0.0324***	0.0363***	0.0361***
	(0.00775)	(0.00777)	(0.00823)	(0.00825)
Website Fixed Effects	Yes	Yes	Yes	Yes
Week Fixed Effects	Yes	Yes	Yes	Yes
Observations	98730	98640	93951	93861
R-Squared	0.581	0.580	0.581	0.580

Notes: Robust standard errors clustered at website level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$. The dependent variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. The policy change is the removal of hosted articles by The Associated Press from Google News. Column (1) presents the initial analysis on all websites. Columns (2)-(4) provide robustness checks with alternative definitions of the control group that exclude the top aggregators (column (2)), international sites (column (3)), and both top aggregators and international sites (column (4)).

3.2 Robustness Checks

We conducted various robustness checks. In Table 5, columns (2)-(4) check robustness of the results to alternative definitions of the control group. As described previously, users navigated to a variety of "non-news" sites after visiting a news aggregator. In columns (2) and (3), our robustness checks omit the top news aggregators and international websites as part of the control group. These alternative definitions of the control group could be warranted if the removal of The Associated Press content also affected navigation to these sites directly (e.g., if The Associated Press content had previously encouraged people to visit international websites) or if the removal of The Associated Press content on Google altered people's perceptions of news aggregators. In column (4), we check robustness to removing both aggregators and international sites from our control group. In general, the results are robust in sign and similar in magnitude.

In Table A-2 of the Appendix, we check the robustness of our results to alternative specifications. We apply a Tobit regression to account for sites that receive zero visits in a given week and also a semi-log regression.[2] Both regressions have similar signs for the coefficients of interest; news sites receive less traffic from Google after the policy change.

We also verified that no global changes occurred in the usage of Yahoo! News and Google News during the period we study. Of particular concern is that the omission of The Associated Press content led people to perhaps leave Google News and explore alternative news aggregators. When we checked the Hitwise data, we found no evidence of such changes in behavior. Indeed, throughout the period we study, Google News remained solidly ranked as fifth for unique visits among news websites while Yahoo! remained ranked as first. Moreover, no change occurred in alternative metrics such as "average visit time" or the number of pages navigated within a website.[3]

[2]For the semi-log regression, we use log(%visits+0.01) as the dependent variable, since some sites receive no visits during a given week.

[3]The share of page views among all news and media sites were 7% and 3% for Yahoo! News and Google

4 Locally Concentrated vs. Nationally Diffuse Sites

In the prior section, we found that users employ technological advances, such as aggregation, to seek further, more specific information. Given the expansion in users' information set, we next consider what information do users seek and which types of content benefit. Depending upon their content, sites may be horizontally differentiated with a very local audience or vertically differentiated with a national audience and acclaimed standards of quality.

Given our finding that overall traffic to news sites from Google News declined after the removal of Associated Press articles, we explore which sites were most affected by the removal of the news content from the aggregator and consequently which sites benefit the most from aggregation. Specifically, we examine whether the extent of the decline varied by the site's level of differentiation. News sites can be local in news coverage with a readership that is regionally concentrated, or sites can be national and diffuse in reach. Tastes for local news sites vary horizontally, depending upon the consumer's interest in regional news while tastes for national news can be vertically differentiated with readers seeking sites, such as The New York Times, with acclaimed standards of quality.

To capture the degree of concentration and diffusion of a news site, we collect monthly data from Hitwise on the fraction of visitors to a given site that originate from each state. Our sample consists of state-level data for 1211 sites for the four weeks ending December 26, 2009 (prior to the dispute).[4] We first calculate the concentration ratio, which we define as the largest share from a state. For instance, if the largest share (47 percent) of readers to boston.com reside in Massachusetts, then the concentration ratio for the site is 0.47. A significant amount of variation in concentration of readership exists in our sample. The average concentration ratio for a site in our sample is 58 percent with a minimum of 6

News. The average visit time for Google was 22 to 23 seconds, and the average visit time for Yahoo was 5 seconds.

[4]Due to minimum reporting standards, some sites did not have state-level data available.

percent to a maximum of 98 percent.

We run a regression similar to equation (1) where we include additional interactions between this measure of concentration and the square of the measure. The specification allows for the policy change to have a quadratic relationship with a site's degree of concentration. As seen in column (1) of Table 6, our results indicate that visits to sites decrease the most for sites with either very low or very high levels of concentration.[5]

Our initial measure of concentration captures readership in the "largest share" state. To reflect the relative degree of concentration across all states, we computed an alternative measure, the Herfindahl-Hirschman Index (HHI), as the sum of the squared shares of readers from each state. The measure lies between 0 and 1, and sites with larger values of the HHI will tend to have readers that are more geographically concentrated. The advantage of the HHI is that it captures the distribution of readers across all states (and not just the largest state), and the HHI places more weight on states with large reader shares. HHI values also range from very low levels of concentration (0.04) to high levels of concentration (0.97), and the average site in our sample has a HHI of 0.43. Column (2) of Table 6 uses HHI as a measure of concentration. The results are qualitatively similar across the different measures and again suggest that the dispute harmed sites with either very concentrated or very diffuse readerships.

Our results are consistent with news aggregators reducing consumers' search costs and allowing readers to easily find sites that specialize in local news. Local news sites may not otherwise find an audience outside of their local region. Our results have an important public policy implication as policymakers enact legislation to encourage the growth of local media, which is viewed as necessary to encourage civic engagement among the public.

[5] We checked that the inflection of the quadratic relationship lies between 0 and 1, which are the relevant values of our concentration ratio. If β_1 and β_2 are the coefficients for $News \times Google \times Concentration$ and $News \times Google \times Concentration^2$, then the effect of policy change varies with respect to concentration according to $\beta_1 + 2\beta_2 Concentration$.

Our findings also suggest that aggregators encourage visits to vertically differentiated sites such as national newspapers with acclaimed standards of quality. As Gentzkow and Shapiro (2011) note, news is vertically differentiated with a small number of sites capturing a large fraction of readers. We examine two pieces of evidence that suggest that these sites with diffuse readership are of higher "quality." First, the sites with the most diffuse readership account for a disproportionate number of visits. For instance, 25 percent of the most diffuse sites account for over half of all visits to news sites. Second, we obtain a list of Pulitzer Prize winners and finalists and confirm that a disproportionate number fall among the most diffuse sites. Figure 5 graphs the proportion of finalists and winners by concentration as measured by HHI. For the 13 categories of news reporting in 2009 and 2010, a vast majority of winners and finalists fall within the highest levels of diffusion.[6]

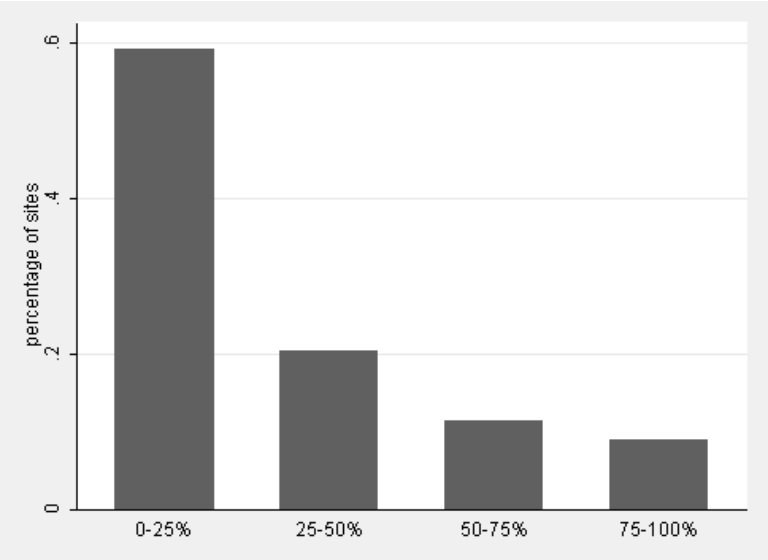
[6]We obtained the list of Pulitzer Prize winners and finalists from the official website www.pulitzer.org. The categories include breaking news reporting, breaking news photography, commentary, correspondence, criticism, editorial cartooning, editorial writing, explanatory reporting, feature photography, feature writing, international reporting, local reporting, and national reporting.

Table 6: The dispute harmed sites with either very diffuse or very concentrated readership

	(1) Concentration ratio	(2) HHI
PeriodDispute × Google × Concentration	0.0860**	0.0857**
	(0.0432)	(0.0371)
PeriodDispute × Google × Concentration^2	-0.0586*	-0.0743**
	(0.0326)	(0.0330)
PeriodDispute × Google × News	-0.0327**	-0.0245**
	(0.0140)	(0.00979)
PeriodDispute × Google	0.00159	0.00159
	(0.00233)	(0.00233)
PeriodDispute	-0.000417	-0.000417
	(0.00112)	(0.00112)
Google	-0.0119*	-0.0119*
	(0.00638)	(0.00638)
PeriodDispute × News	0.00715	0.00552
	(0.00504)	(0.00439)
News × Google	0.144***	0.118***
	(0.0307)	(0.0238)
PeriodDispute × Concentration	-0.0197	-0.0201
	(0.0124)	(0.0156)
Google × Concentration	-0.301***	-0.356***
	(0.0867)	(0.0828)
PeriodDispute × Concentration^2	0.0143*	0.0185
	(0.00809)	(0.0130)
Google × Concentration^2	0.165**	0.277***
	(0.0647)	(0.0700)
Website Fixed Effects	Yes	Yes
Week Fixed Effects	Yes	Yes
Observations	96066	96066
R-Squared	0.582	0.582

Notes: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The dependent variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. The policy change is the removal of articles by The Associated Press from Google News. Each column contains a measure of geographic concentration for a site. Column (1) uses the concentration ratio, the share of the state with the largest fraction of readers. Column (2) uses the Herfindahl-Hirschman index (HHI), which is the sum of the squared shares of readers from each state. Sites with higher values of the concentration ratio or HHI have a more geographically concentrated readership.

Figure 5: Pulitzer prize winners and finalists by site's geographic concentration



Notes: This figure shows the percentage of sites that were Pulitzer prize winners and finalists across different levels of geographic concentration. The sites are divided into four groups according to the HHI measure described in Section 4. For instance, the category "0-25%" represents the top 25 percent of sites with the most diffuse readership, and the category "75-100%" represents the 25 percent of sites with the least diffuse readership.

5 Conclusion

The digital revolution poses new challenges to our interpretation and application of copyright protection. In particular, the practice of digital aggregation or the collection of extracts of copyrighted work onto a single website has led to both lawsuits and uncertainty over the economic consequences of such practices. To investigate the consequences of exposure to snippets of copyrighted content, we exploit an unusual natural experiment—a breakdown in contract negotiations between The Associated Press and Google—which prompted Google to stop hosting The Associated Press content for 7 weeks. Our unique dataset on Internet users derives from Hitwise, which documents sites that users visit after navigating to an aggregator. We find evidence that when Google News no longer hosted The Associated Press content, Google News users were less likely to visit other news websites after visiting Google News relative to Yahoo! News users who experienced no such removal of The Associated Press content. Our results suggest that this pattern was driven by a reduction in visits to either very regionally concentrated or national websites. Consequently, the relaxation of intellectual property rights may benefit content that is either horizontally differentiated, such as local sites, or vertically differentiated, such as national sites with acclaimed standards of quality.

Our results have implications for policies regarding intellectual property and copyright. As stated by Isbell (2010), "for all of the attention that news aggregators have received, no case in the United States has yet definitively addressed the question of whether their activities are legal." One of the major criteria for fair use, as spelled out by section 107 of the Copyright Act, is to understand "the effect of the use upon the potential market for or value of the copyrighted work." Our results suggest that, at least for the users of news aggregators, that these extracts of copyrighted content do not served a complete substitute and do induce users to navigate further. Our work provides some evidence that the potential economic harm is limited by positive spillovers between the extracts of copyrighted content

and the original works.

This paper also has several implications for our understanding of the "Information Economy." Our results suggest that when digital advances reduce search costs, this promotes a greater search for information rather than simply reducing the time that a person spends on a predefined set of information. We also explore which types of content may benefit from aggregation. A new trend has emerged whereby content providers have started creating "hyperlocal" sites and "microcontent" that focuses on information targeted to a very specific geographic area, sometimes down to the neighborhood or block-level (Miller and Stone, 2009). Even though the set of potential users is "inherently small" for microcontent, our results imply that aggregation of content from hyperlocal sites may encourage consumer traffic to these sites and help expand the user base. Furthermore, speculation often ensues over whether "quality" content will survive in the onslaught of information online (Scheck, 2010). Our results suggest that digital aggregation does benefit high quality sources and that even with the plethora of sources available in the Internet age, users still do seek sources with acclaimed standards of quality.

References

Alpert, J. and N. Hajaj (2008, July 25). We knew the web was big. *Google Blogspot*.

Ardia, D. (2008, June 16). Associated Press Sends DMCA Takedown to Drudge Retort, Backpedals, and Now Seeks to Define Fair Use for Bloggers. *Citizen Media Law Project*.

Arrington, M. (2010, Feb 2). Everybody forgets the readers when they bash news aggregators. *Techcrunch*.

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

Athey, S. and M. Mobius (2012). The Impact of News Aggregators on Internet News Consumption: The Case of Localization. working paper.

Bakos, J. Y. (1997). Reducing buyer search costs: Implications for electronic marketplaces. *Management Science 43*(12), 1676–1692.

Bechtold, S. (2011, January). Google AdWords and European trademark law. *Communications of the ACM 54*, 30–32.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Chiou, L. and C. Tucker (2011). How does the use of trademarks by third-party sellers affect online search? *Mimeo, MIT*.

Cohen, J. (2009, December 2). Same protocol, more options for news publishers. *Posting on Google News's Blog*.

Danaher, B., S. Dhanasobhon, M. D. Smith, and R. Telang (November/December 2010). Converting pirates without cannibalizing purchasers: The impact of digital distribution on physical sales and internet piracy. *Marketing Science 29*(6), 1138–1151.

Doctor, K. (2009). News users 2009. *Outsell Market Intelligence Service: Market Report*.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics*.

George, L. and C. Hogendorn (2012). Aggregators, Search and the Economics of New Media Institutions. *Information Economics and Policy 24*(1), 40–51.

Ghose, A., A. Goldfarb, and S. Han (2011). How is the Mobile Internet Different? Search Costs and Local Activities. *Mimeo, New York University*.

Greenstein, S. (2011, April). The direction of broadband spillovers. *Micro Economics*, 103–104.

Greenstein, S., J. Lerner, and S. Stern (2011). The economics of digitization: An agenda for NSF. *NSF Report*.

Haddad, M. (2010). Associated Press Google News Partnership Ends. Business 2.0 Press, January 12.

Isbell, K. (2010). The Rise of the News Aggregator: Legal Implications and Best Practices. *Berkman Center, Research Publication No. 2010-10*.

Jeon, D. and N. N. Esfahani (2012). News Aggregators and Competition among Newspapers in the Internet. working paper.

Kaplan, D. (2010, Feb 2). Mark Cuban: Google, Content Aggregators Are Vampires; Newspapers Are Zombies. PaidContent.org.

Krazit, T. (2010, Jan 11). Hosted AP content on hold in Google News . *CNET*.

Miller, C. and B. Stone (2009). 'Hyperlocal' Web Sites Deliver News without Newspapers. The New York Times, April 12.

Oberholzer-Gee, F. and K. Strumpf (2007). The effect of file sharing on record sales: An empirical analysis. *Journal of Political Economy 115*, 1–42.

Rob, R. and J. Waldfogel (2006). Piracy on the high C's: Music downloading, sales displacement, and social welfare in a sample of college students. *Journal of Law & Economics 49*(1), 29–62.

Rutt, J. (2011). Aggregators and the News Industry: Charging for Access to Content. working paper.

Sandoval, G. (2009). Google May Lose WSJ, Other News Corp. Sites. CNET News, November 9.

Scheck, O. (2010, March 15). What the Internet Will Mean for Journalism and Journalists: Insights from the Edge. *3 Quarks Daily*.

Shapiro, C. and H. Varian (1999). *Information rules: A strategic guide to the network economy*. Harvard Business School Press, Boston.

Sullivan, D. (2010, January 8). Where is AP In Google News? Apparently In Limbo, As Contract Running Out. *Search Engine Land*.

Table A-1: Demographic description of users

Measure	Yahoo! News	Google News	New York Times
Male	59.95	63.8	61.21
Age 18-24	12.12	13.89	6.17
Age 25-34	18.05	14.72	13.93
Age 35-44	19.03	17.08	12.98
Age 45-54	21.41	22.24	19.45
Age 55+	29.38	32.06	47.47
Income <30k	22.33	20.77	20.76
Income 30-60k	28.82	27.53	26.36
Income 60-100k	24.95	24.6	24.82
Income 100-150k	14.61	17.5	17.29
Income >150k	9.29	9.6	10.77

Source: Hitwise

Notes: This table reports the fraction of users within each demographic category. Statistics are reported for users of Yahoo! News, Google News, and the New York Times website.

Appendix

The following contains excerpts from Experian Hitwise "How We Do It" description on its official website.

> Hitwise has developed proprietary software that Internet Service Providers (ISPs) use to analyze website logs created on their network. This anonymous data is aggregated and provided to Hitwise, where it is analyzed to provide a range of industry standard metrics relating to the viewing of websites including page requests, visits, average visit length, search terms and behaviour.
>
> Hitwise is able to combine this rich ISP data with data from opt-in panel partners and with region specific consumer demographic and lifestyle information.
>
> Hitwise collects aggregate usage data from a geographically diverse range of ISP networks and opt-in panels, representing all types of Internet usage, including home, work, educational and public access. To ensure this data is accurate and representative, it is weighted to universe estimates in each market. Because of the extensive sample size (25 million people worldwide, including 10 million in the US), Hitwise can provide detailed insights into the search terms used to find thousands of sites as well as a range of clickstream reports, analyzing the movement of visitors between sites.
>
> Hitwise only extracts aggregate information. No personal information is seen or captured by Hitwise in according with local and international privacy guidelines. Hitwise's methodology is audited by PricewaterhouseCoopers on an annual basis.

Table A-2: Robustness checks: Downstream traffic to local news sites from Google News and Yahoo! News before and after the policy change

	(1) Tobit	(2) Semi-log
PeriodDispute × Google × News	-0.0240**	-0.0225*
	(0.00951)	(0.0127)
PeriodDispute × Google	0.00391	-0.00753
	(0.00569)	(0.00566)
PeriodDispute	-0.00482	0.00335
	(0.00452)	(0.00517)
Google	0.0249***	-0.0255**
	(0.00891)	(0.0123)
News	0.117***	
	(0.0216)	
PeriodDispute × News	0.0143**	0.0126
	(0.00568)	(0.00882)
News × Google	-0.0127	0.0749***
	(0.0144)	(0.0233)
Website Fixed Effects	No	Yes
Week Fixed Effects	Yes	Yes
Observations	98730	98730
R-Squared		0.688

Notes: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The dependent variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. The policy change is the removal of articles by The Associated Press from Google News.