NBER WORKING PAPER SERIES

CONTENT AGGREGATION BY PLATFORMS: THE CASE OF THE NEWS MEDIA

Lesley Chiou
Catherine Tucker

Working Paper 21404
http://www.nber.org/papers/w21404

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2015

Earlier versions of this paper were circulated under the titles "Copyright, Digitization, and Aggregation" and "News and Online Aggregators." We thank Robert Seamans for useful comments. We thank Christopher Hafer of Experian Hitwise. We also thank Shreya Bhaskaran, Cassandra Crosby, Sara Mcknight, and Anthony Quach for excellent research assistance. Financial support from the NBER Innovation Policy Group, NSF CAREER Award #1053398 and NET Institute (www.NETinst.org) is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed a financial relationship of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w21404.ack

Content Aggregation by Platforms: The Case of the News Media
Lesley Chiou and Catherine Tucker
NBER Working Paper No. 21404
July 2015
JEL No. L63,L82,L86,L88

ABSTRACT

The digitization of content has led to the emergence of platforms that draw information from multiple sources. Policymakers are concerned that these new platforms threaten incentives for the production of original content. As a result, policymakers are contemplating regulations that would force aggregation platforms to pay or require an explicit "opt-in" for content providers. 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" content or to investigate in depth. We study a contract dispute that led a major aggregator to remove information from a major content provider. We find that after the removal, users were less likely to investigate additional, related content in depth, particularly sources that were horizontally or vertically differentiated.

Lesley Chiou
Occidental College
1600 Campus Road
Los Angeles, CA 90041
lchiou@oxy.edu

Catherine Tucker
MIT Sloan School of Management
100 Main Street, E62-533
Cambridge, MA 02142
and NBER
cetucker@mit.edu

1 Introduction

Platforms play an important role as intermediaries. By bringing together two sides of a market, platforms can facilitate transactions and provide information that consumers then use for making decisions. In recent years, the digitization of content has led to the prominence of platforms as aggregators of content in many economically important industries, including media and Internet-based industries (Evans and Schmalensee, 2012).

These new platforms consolidate content from multiple sources into one place, thereby lowering the transactions costs of obtaining content and introducing new information to consumers. Such aggregation platforms represent a shift away from peer-to-peer platforms operating outside of the law towards platforms that aggregate digital content within a legal framework, including Spotify for music, Hulu for movies, and Google News for news content. While an extensive literature has focused on pricing and piracy by platforms (Rob and Waldfogel, 2006; Oberholzer-Gee and Strumpf, 2007; Danaher et al., 2010), little is known about how the quantity and quality of content provided by a platform that tries to operate within legal bounds influences consumer search and the resulting consequences for content providers.

For these reasons, platforms have attracted considerable legal and policy attention. For instance, musicians have sued Spotify, a music aggregator, for its creation of music playlists (Gardner, 2014). A platform aggregator of movie scripts has been sued by Fox for infringing on copyright (Masnick, 2010). The news agency L' Agence France Presse has sued Google News for aggregating its content and infringing copyright (Isbell, 2010). Recent regulation in the European Union has attempted to make content on a platform an "opt-in" decision (Pfanner, 2012; Eddy, 2013) where a content provider has the right to decide whether or not their content appears on the aggregation platform. Therefore, a natural question for a content provider is how opting in to providing content on a platform affects consumer

behavior on the platform. Content producers fear that consumers may use these extracts of content as a substitute for accessing and reading the full content. Platforms argue that aggregation encourages users to seek out additional content because aggregators feature multiple sources. In practice, this open question is an empirical one.

To investigate the potential importance of these issues, we examine how a change in content provided by a platform affects subsequent consumer search for different types of information and its consequences for content providers. We study the case of the news media industry. The news media industry presents an attractive setting for studying these issues, due to the rapid growth of digital news content, multiple content providers who have expressed fears about the viability of their business model in the presence of aggregators, and the prominence of platforms such as Google News. Specifically, we focus on arguably two of the largest players in US news media industry, Google News and The Associated Press (AP). Google News is among the most-read news aggregators, automating the aggregation of news content from 25,000 news sources. The Associated Press is a prominent content provider and was created in 1846 to fund news-gathering activities between its newspaper participants. The Associated Press has received 51 Pulitzer Prizes, reflecting its status in the news industry and investments in journalism.

To overcome a major challenge in tackling these questions, we use a novel approach and exploit a contract dispute as an exogenous shifter of the availability of content on Google News. In January 2010, after a breakdown in licensing negotiations, Google removed all news articles that were syndicated by The Associated Press from its news aggregator (Haddad, 2010). These articles were typically shortened versions of stories that appeared in a select number of The AP-associated newspapers. The contractual dispute provides a useful case study to examine the effects of a content provider "opting-out" of providing content and therefore the role of the aggregation platform in influencing consumer behavior. One attractive feature of studying this dispute is the scale of The Associated Press. Not only

does the removal represent a large shock, but the removal of content most likely did not reflect the wishes of a particular newspaper or any self-selection issues. Since The AP was created historically with deep ties to the news media, The AP can capture a broader effect of content change for news media.

We compare users' website visits before and after this contract dispute relative to traffic from Yahoo! News, which continued to provide The Associated Press content during this period. We link the insights from our empirical results to the theoretical literature on platforms. In theory, on one hand, consumers may use platforms to scan the extracts of content without clicking through to pursue more in-depth material ("scanning effect"). On the other hand, consumers may use platforms to explore new material more deeply ("traffic effect.") Our data allows us to measure both of these effects.

Our results indicate that after The Associated Press content was removed from Google News, fewer users subsequently visited news sites after navigating to Google News relative to users who had used Yahoo! News. The pattern was driven by news websites that were specifically local in content or news websites with national recognition as being high quality. Thus, we find evidence that the traffic effect is large, as aggregators may guide users to new content. We do not find evidence of a scanning effect, as overall traffic to Google News and Yahoo! News remained relatively comparable during our time period. We also explore the institutional relationship between news sites and The AP and find that websites with stronger ties to The AP suffered a drop in traffic after the dispute.

Our results inform the legal and public policies for two reasons. First, our results suggest that the decision to opt-in to an aggregation platform should depend on whether the content provider is considered high-quality or highly unusual. Both these characteristics, appear to encourage users to use the aggregator to explore content more deeply instead of scanning content. Second, one surprising development is that despite Germany publishers lobbying for an opt-in law, none have chosen to opt-out (Lomas, 2013). Our paper provides an explanation

of such behavior—ultimately aggregators may benefit many newspapers, especially high-quality ones, and the purpose of the "opt-in" provision may be to increase bargaining power over payments to news providers rather than an actual desire for copyright holders to opt-out of the aggregator systems.

Our focus is not on how aggregators affect direct navigation to the content providers' websites—Sandoval (2009), Arrington (2010), and Athey and Mobius (2012) discuss this. Instead, we measure how a platform's expansion or contraction of content affects subsequent navigation by users. More broadly, our analysis is also related to prior work that describes how digital technologies have affected search costs and generated spillovers (Shapiro and Varian, 1999; Bakos, 1997; Ghose et al., 2011; Greenstein, 2011). The novelty of our study is that we are the first to explore how digital technology affects the set of information gathered by consumers.

Our results have implications for copyright policy regarding platforms that aggregate digital content. The digital revolution has challenged various aspects of copyright protection (Greenstein et al., 2011) but much of the focus has been on peer-to-peer piracy rather than newer legitimate business models that aggregate specific kinds of content. Online aggregators in media 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). Our empirical distinction between a scanning effect where the aggregator substitutes for original content and a traffic effect where the aggregator is complementary, is useful for analyzing the potential policy implications of such business models. The fact we find evidence of a "traffic effect" even with a relatively large amount of content on an aggregator, is perhaps evidence that the "fair use" exemptions often relied on by such sites are less potentially damaging to the original copyright holder than often thought.

2 Institutional Setting and Data

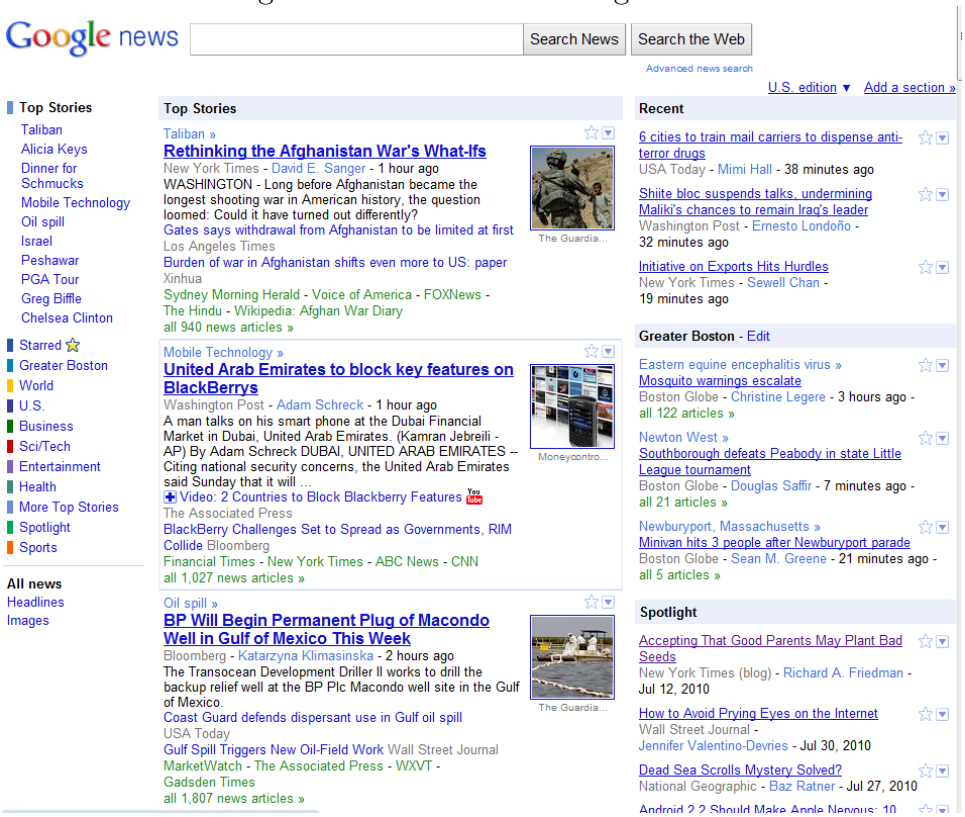
2.1 Contractual Dispute between Google and The Associated Press

Google News is ranked as the fifth most visited news website by Hitwise. Receiving 2.90% of all news site visits, Google News is the second most popular news aggregator service after Yahoo! News, which received 7.09% of all news site visits. 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 distinguish 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, The AP 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 AP is a cooperative owned by various newspapers and radio and television stations in the United States. These stakeholders both contribute stories to The AP and use material written by The AP staff journalists. During the past decade, The AP 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 AP invoked the Digital Millennium Copyright Act and insisted that various bloggers remove The AP content (Ardia, 2008).

The origins of The AP and its business model reveal that The AP's role was to serve as a coordination function in the old media world of physical newspapers, and to allow newspapers to pool content and stories and hence enjoy economies of scale in news reporting.

Figure 1: Screenshot of Google News



Note: 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.

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.

Little evidence exists that The AP has tried to push its own website as an alternative "news-wire" service; instead, The AP website functions mainly as a corporate site which simply lists member newspapers. The AP's reluctance to perform a news-wire role may be due to its origins as a newspaper association; it may be reluctant to pursue anything that competes directly with a newspaper's business model. It is not clear how an organization founded under the traditional model where each newspaper provided full news coverage to individual print subscribers fits into a world where consumers consume news digitally. Table 1, which summarizes the major events of the AP and Google relationship, makes clear that The AP is worried about the implications of the rise of search and aggregation technology for its business model.

Since both The AP and Google News are key players in the distribution of news online, it is not surprising they have forged a partnership. Their licensing agreement also protects Google News from allegations of copyright infringement over The AP content, given the current uncertainty over copyright law for aggregators. We study a discontinuity in this relationship, produced by negotiations surrounding the contract renewal at the end of January 2010. As part of their existing contract, Google and The AP agreed that The 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 content from The Associated Press 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 discussion is necessarily based upon the observations of industry outsiders, since both Google and The AP signed binding non-disclosure agreements, which prevented them from ever commenting on the course or outcome of negotiations (Sullivan, 2010).

The removal of The AP content represents a useful quasi-experiment. Since the removal of content was provoked by the intricacies of contract negotiations, its timing can be thought

of as reasonably exogenous, as the removal was determined by the expiration of the contract rather than any considerations of the popularity (or lack thereof) of The AP content at that time. As detailed in Table 1, Google removed The AP content from December 23, 2009 until sometime in February 2010. Fortunately for our purposes, Yahoo! News continued to host The AP 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 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, do consumers use aggregators to go more in depth into content ("a traffic effect") or in lieu of content sites ("scanning effect")? This depends upon whether consumers view news aggregators as a complement or substitute to original news sources.

For instance, The AP ran a news story about the economic depression in Michigan. 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. These papers are local affiliates of The Associated Press and typically expand in their newspapers on the summary The Associated Press content. 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 preliminary analysis focuses on the period immediately prior to and during the removal of The AP articles from Google News from December 2009 to January 2010. Contract negotiations continued until August 30, 2010 when a long-term contract was signed between

Figure 2: Example screenshot of The Associated Press article hosted on Google News

Michigan voters search for economic savior

Associated Press

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.

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

Map

Michigan
Milwaukee
Chicago
Grand Rapids
Detroit
Toledo
Fort Wayne
Joliet
Ohio
Indiana

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) »

Note: Google News, August 2010. Text of article has been slightly edited to fit on page.

Google News and The AP (Krazit, 2010). We also examine the consequences of the reinstatement of the long-term relationship between Google News and The AP between January 2010 and October 2010.[1] As can be seen in the screenshot in Figure 2, some of The AP

[1]In the interim, on February 2010, a temporary short-term deal was agreed upon which gave rise to the content that can be seen in Figure 2. It is not clear at which point in February the relationship was resumed between Google News and The AP. It is also not apparent whether the short-term deal during this time consisted of the older, missing content or new content or whether Google changed the presentation of The AP articles afterwards. For example, it would be problematic if Google decided to highlight The AP content after the contract negotiations were concluded, perhaps as a "sweetener" while on-going negotiations continued until September. For these reasons, we collect data on the full reinstatement of the long-term relationship between Google News and The AP that resumed.

content was already restored at the end of August, however, it appears that October is the appropriate date of analysis for the reinstatement of The AP material. Since content is added daily and appears for 30 days on Google News, one month following August 30 is the month of October 2010 when all content for the past 30 days was fully reinstated and available.

2.2 Description of Data on Consumer Behavior

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 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 2000 sites navigated to 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. When Hitwise reports data for downstream sites, it tracks web-surfing behavior by recording where people navigate to after visiting a particular site at an aggregate level.

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.

To identify which sites are related to The AP, we categorized the websites into two main classes: "news" (e.g., newyorktimes.com, bostonherald.com) and "non-news" (e.g., Yahoo! Mail, Youtube.com). Our news category consists of a strict definition of sites that fall under Hitwise's categories of print media and broadcast media. We made sure that these news sites reflected The AP network of member news organizations as well as news outlets that subscribe to The AP service and provide The AP content. As we are interested in traffic to websites of primary news sources, we exclude weather sites and the top aggregators—e.g., Yahoo! News, Google News, AOL News, Bing News, Ask News, 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 AP 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. A site received on average 0.016 percent of downstream visits. News sites represent 15 percent of all sites where we observe subsequent visits within our sample, and non-news sites account for 85 percent. Aggregator, international, and weather sites account for a smaller fraction of sites compared to news sites.

Table 3 displays the top 40 news websites in our dataset and the average percentage of 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

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

	Mean	Std Dev	Min	Max
% visits	0.00016	0.0019	0	0.18
Google News	0.50	0.50	0	1
Yahoo! News	0.50	0.50	0	1
APContentRemoval	0.67	0.47	0	1
News Site	0.15	0.36	0	1
Non-news Site	0.85	0.36	0	1
Aggregator Site	0.0013	0.036	0	1
International Site	0.048	0.21	0	1
Weather Site	0.0067	0.081	0	1
Observations	98730			

Note: 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; this variable is measured from 0 to 1. The dispute between The Associated Press and Google News occurred after December 23, 2009. The variable $APContentRemoval$ is an indicator variable for whether the week occurred during the period of the dispute. News sites refer to print media and broadcast media sites as defined by Hitwise, excluding weather sites, international news sites, and top news aggregators.

aggregator. Table 4 displays the top 40 non-news websites in our dataset, excluding international and aggregator 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; the users 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 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 Theoretical Predictions of Scanning and Traffic

Theoretically two countervailing forces exist when users arrive at an aggregator and use it to search for information. Consumers may use aggregators to explore content more in depth ("traffic effect") or in lieu of content sites ("scanning effect.") These effects have been captured in theoretical models of platforms and aggregators. For instance, Jeon and Esfahani (2012) describes "market expansion" and "business-stealing" effects of aggregators. Rutt (2011) examines a model with two types of consumers—those "loyal" to original content sites and "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; George and Hogendorn, 2012).

Under the "scanning effect," consumers scan all articles at an aggregator and either finish

their search or move on to further articles, depending upon whether the article satisfies them. The interpretation of "fair use" and other facets of copyright law shape the scanning effect. If fair use is more permissive, then more content is featured on aggregator, which leads scanning to be more valuable. Consumers may scan more content at a single site. Our study examines a change in content at one aggregator. If consumers use the aggregator for scanning, then the removal of The AP content will reduce the quality of scanning, and subsequently, total visits to Google News may decline as consumers seek other aggregators.

The "traffic effect" suggests that consumers are interested in pursuing content more in depth. When consumers read a headline or excerpt on the aggregators, they will be prompted to click on the links to the content provider for further details of the story. If consumers use aggregators to "dig" for more content, then the removal of The AP content will lead to less traffic to news sites.

We will empirically test for the two effects of scanning and traffic in our analysis below.

3.2 The Scanning Effect: Overall Visits to an Aggregator

First, to test for the scanning effect, we investigate whether the removal of The AP content from Google News led to a shift away from Google News. The idea is that if consumers use aggregators to merely scan headlines and excerpts of articles, then the removal of content from Google News will lower the quality of scanning on Google News. Consumers will shift away from Google News towards other aggregators.

We collect additional data from comScore on total visits to Google News and Yahoo! News. ComScore tracks the online activity of a panel of more than 2 million users based in the US and subsequently aggregates their search patterns for resale to commercial clients. ComScore recruits its panel members through affiliate programs and partnering with third party application providers. ComScore emphasizes and discusses the representativeness of their sample to the general population in their Marketer User Guide. ComScore data has also

been used in several academic studies and noted as a "highly regarded proprietary [source] for information on the size and composition of media audiences" (Gentzkow and Shapiro, 2011; Montgomery et al., 2004; De Los Santos et al., 2012; Chiou and Tucker, 2010).

Table 5 reports the number of monthly visits to each aggregator during our period of study. We do not find evidence of a precipitous drop in monthly visits to Google News relative to Yahoo! News in the wake of the dispute from December 2009 to January 2010.

When we checked the Hitwise data, we found no evidence of 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. The share of page views among all news and media sites were 7% and 3% for Yahoo! News and Google News. The average visit time for Google was 22 to 23 seconds, and the average visit time for Yahoo was 5 seconds.

In theory, removing The AP content reduces the quality of scanning on Google News and could potentially reduce total visits to Google News as it loses in competition to other aggregators. We do not find evidence that competing aggregators act as substitute platforms for one another.

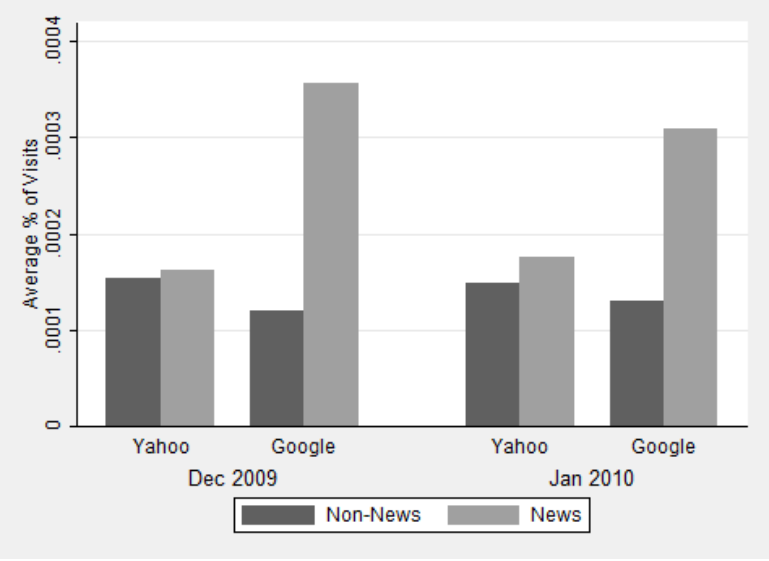
3.3 The Traffic Effect: Downstream Visits after an Aggregator

In this section, we investigate the magnitude of the "traffic effect." We explore whether content on a news aggregator prompts users to seek further news. We examine whether traffic to news sites from Google News falls after the removal of The AP content from Google News. Our preliminary analysis examines visits to news sites after navigating to an aggregator.

Figure 3 illustrates the aggregate mean percentage of downstream traffic to news and non-news sites for users that visited Google News and Yahoo! News during this period. As

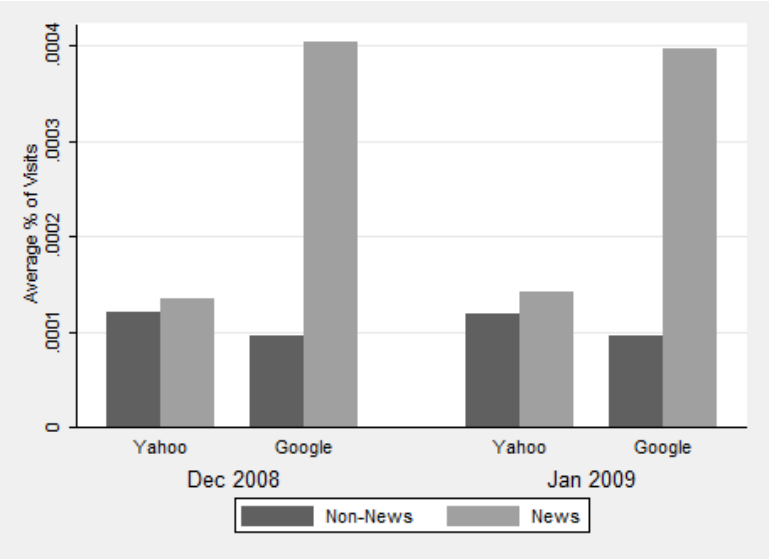
seen in the graph, little change occurs in downstream site navigation for Yahoo! However, news sites experience a decline in visits from Google News during the period of the removal of The AP content, relative to the change in traffic from Yahoo! News. To investigate whether this pattern could be due to underlying seasonality in news consumption, we examine the change in visits in the prior year during the same calendar months. 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 in period before and after The Associated Press content was removed



Note: 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 when no content was removed (December 2008 and January 2009)



Note: 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 content from Google News.

Table 3: Top 40 news websites visited after Google News and Yahoo! News

	Avg Visit Pct
abcnews.com	0.021
bloomberg.com	0.0051
boston.com	0.0048
bostonherald.com	0.0038
businessweek.com	0.0031
cbsnews.com	0.0037
chron.com	0.0027
cnn.com	0.019
csmonitor.com	0.0030
dallasnews.com	0.0021
edition.cnn.com	0.0040
examiner.com	0.013
foxnews.com	0.011
freep.com	0.0025
latimes.com	0.0048
mcclatchydc.com	0.0019
mercurynews.com	0.0044
miamiherald.com	0.0030
msnbc.com	0.0083
nj.com	0.0023
npr.org	0.0032
nydailynews.com	0.016
nypost.com	0.0051
nytimes.com	0.029
pcworld.com	0.0037
people.com	0.0078
philly.com	0.0030
politico.com	0.011
reuters.com	0.014
seattlep-i.nwsource.com	0.0022
seattletimes.nwsource.com	0.0022
sfgate.com	0.0033
theweek.com	0.0029
time.com	0.012
usatoday.com	0.0072
usmagazine.com	0.0046
voanews.com	0.0026
washingtonpost.com	0.017
wsj.com	0.017

Table 4: Top 40 Non-news websites visited after Google News and Yahoo! News

	Avg Visit Pct
amazon.com	0.0059
aol.com	0.0023
aralifestyle.com	0.0014
ask.com	0.0019
bing.com	0.0062
blogsearch.google.com	0.0077
cosmos.bcst.yahoo.com	0.0095
ebay.com	0.0100
education.yahoo.net	0.0034
espn.com	0.0028
facebook.com	0.062
finance.google.com	0.0036
finance.yahoo.com	0.0060
gmail.com	0.015
google.com	0.12
howlifeworks.com	0.010
images.google.com	0.0050
latimesblogs.latimes.com	0.0016
livescience.com	0.0038
mail.live.com	0.013
mail.yahoo.com	0.099
maps.google.com	0.0023
members.yahoo.com	0.0029
movies.yahoo.com	0.0013
msn.com	0.010
my.yahoo.com	0.0067
myspace.com	0.015
omg.yahoo.com	0.0032
search.yahoo.com	0.022
shine.yahoo.com	0.0013
space.com	0.0015
sports.yahoo.com	0.0026
tmz.aol.com	0.0020
tv.yahoo.com	0.0012
video.google.com	0.0027
weather.com	0.0067
weather.yahoo.com	0.0039
wikipedia.org	0.0050
yahoo.com	0.072
youtube.com	0.025

Notes: All visit percentages are measured between 0 and 1. News sites are defined as sites for print and broadcast media. International and aggregator sites are excluded from the table above. Please see text for further details.

Table 5: Total monthly visits to Google News and Yahoo! News

Date	**Google News**	**Yahoo! News**
November 2009	75,667,000	224,160,000
December 2009	78,160,000	267,570,000
January 2010	79,373,000	262,700,000

Source: ComScore

To formalize the insights provided by Figure 3, we run a difference-in-differences regression for the policy change and estimate percentage of visits to website i after visiting news aggregator j in week t. We use a Generalized Linear Model (GLM) framework to estimate the following equation:

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

where $g(.)$ is the generalized function, $E(\%visits)$ is the expected value of the percentage of visits, $News$ 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 $APContentRemoval$ is an indicator variable equal to 1 for the weeks after the removal of The AP content from Google News. The controls α are downstream-website fixed effects.[2] 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 $News \times Google \times APContentRemoval$ captures the effect of The AP content 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.

Our outcome variable is the probability or share for each website as measured relative to a particular aggregator. Our specification captures the probability or percentage of visits to website i after visiting news aggregator j in week t where the percentage is calculated from all outgoing visits from aggregator j. For instance, if V_{jt} is the number of all outgoing visits from search engine j in week t to any site and v_{ijt} is the number of outgoing visits from search engine j in week t to site i, then we observe the percentage of visits where

[2]Due to computational limitations, we include fixed effects for the top 500 websites.

$\%visits_{ijt} = v_{ijt}/V_{jt}$. If consumers were equally likely to visit news sites before and after the policy change, then the measure of the share of outgoing traffic to news sites from each aggregator will remain the same. The shares of each site outgoing from Google News are measured as a fraction or probability conditional on traffic to Google News. In other words, the market shares are calculated separately for Google News and Yahoo! News. Because the dependent variable is measured relative to traffic from each news aggregator separately, the share of outbound traffic to news sites would remain the same even with shifts in total traffic to each news site.

We estimate this specification using a Generalized Linear Model with a fractional response variable (Papke and Wooldridge, 1996). Following Papke and Wooldridge (2008), a GLM with link logit function $g(.)$ and family binomial takes into account that the dependent variable (percentage of visits) lies between 0 and 1. We cluster our standard errors at the website level to avoid the downward bias reported by Bertrand et al. (2004).

Table 6 reports the results in Column (1) for the coefficients in our full specification as described by equation (1). The months December 2009 and January 2010 reflect the period before and after the dispute between The AP and Google News. The negative coefficient on $News \times Google \times APContentRemoval$ 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.

Under the GLM logit link, we interpret our exponentiated coefficients as odds ratios. In our setting, the odds is the probability of visiting a site compared to the probability of not visiting a site, or the share of visits to a site compared to the share of visits to all other sites. The coefficient -0.332 on $APContentRemoval \times Google \times News$ implies that the odds ratio

is 0.72 or 72 percent of its level prior to the policy change, since the exponential of -0.332 is 0.72. In other words, the odds ratio fell by 28 percent. Consequently, the odds of visiting a news site on Google News relative to a non-news site on Google News decreased by 28 percent compared to the odds of visiting a news site on Yahoo! News relative to a non-news site on Yahoo! News.

Table 6: Downstream traffic from Google News and Yahoo! News during removal and reinstatement of content from The Associated Press

	(1) Removal	(2) Reinstatement	(3) Falsification	(4) Falsification
APContentRemoval × Google × News	-0.332*			
	(0.177)			
APContentRestored × Google × News		0.670**	0.298	
		(0.283)	(0.193)	
December × Google × News				-0.0367
				(0.119)
APContentRemoval × Google	0.113			
	(0.162)			
APContentRemoval	-0.0281			
	(0.0691)			
Google	0.0883	0.171	0.109	0.255***
	(0.114)	(0.119)	(0.0773)	(0.0595)
News	0.605***	0.859***	0.752***	0.562***
	(0.0600)	(0.0756)	(0.0768)	(0.0686)
APContentRemoval × News	0.108*			
	(0.0631)			
News × Google	-0.210	-0.553***	-0.331***	-0.405***
	(0.134)	(0.146)	(0.118)	(0.0844)
APContentRestored		0.0257	0.0233	
		(0.109)	(0.0668)	
APContentRestored × Google		-0.103	0.0160	
		(0.239)	(0.129)	
APContentRestored × News		-0.510***	-0.386***	
		(0.150)	(0.128)	
December				-0.00958
				(0.134)
December × Google				-0.00202
				(0.106)
December × News				0.108
				(0.108)
Week Fixed Effects	Yes	Yes	Yes	Yes
Website Fixed Effects	Yes	Yes	Yes	No
Observations	98730	119640	103113	84048

Note: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The outcome variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. In Column (1), the removal of The AP content is the removal of hosted articles by The Associated Press from Google News in January 2010. In Column (2), the restoration of The AP content is the restoration of hosted articles by The Associated Press in Google News in October 2010. In Column (3), the falsification check compares December 2009 and October 2010 when The AP content was available in Google News in both months. In Column (4), the falsification check compares November 2009 and December 2009 to test for a pre-trend.

We also collect additional data that tests the effects after the dispute was resolved. Around August 30, 2010, Google News and The AP formally signed a long-term contract to continue their relationship (Krazit, 2010). Since content is added each day and appears for 30 days on Google News, we collect data for the weeks in October 2010 when all content for the past 30 days was fully reinstated and available. In Table 6, Column (2) compares January 2010 (when no content from The AP was available) to October (when The AP content was fully reinstated). As expected, we observe a positive effect, as visits to news sites are higher after the reinstatement of content through a long-term contract.

So far our analyses in this section of the removal and subsequent reinstatement of The AP content suggest that a "traffic effect" does exist. Consumers do appear to use platforms to seek new and further content. One striking feature of the way that The AP content was featured on Google News is that in general as shown in Figure 2 quite a large amount of news content was shown rather than merely a snippet. This makes the fact we find evidence of a traffic effect rather than merely a scanning effect more striking.

4 Robustness Checks

4.1 The Relationship between The Associated Press and News Organizations

As a sharper test of our theory and as a robustness check, we examine whether the traffic effect was strongest among sites more embedded within The AP ecosystem. To clarify further institutional details of The AP, recall that news organizations may subscribe to content from The AP, which they then disseminate on their site. We collect data on the number of stories featured from The AP on each news site.[3] We create a measure *APstories* that measure the fraction of stories from The AP featured on the homepage of the news site. Our measure is intended to identify which news sites are more likely to be featured alongside the hosted

[3]Specifically, for each news site, we counted the number of articles from The AP that are featured in the top 20 stories on the home page.

article and therefore removed from Google News when The AP hosted content was removed. The measure $APstories$ reflects the strength of the relationship between the site and The AP.

Table 7 reports our estimates of equation (1) with our continuous measure $APstories$ as our variable of interest instead of the indicator variable $News$. Consistent with our results on the traffic effect, we find that sites that are more embedded in The AP ecosystem experienced a drop in traffic relative to Yahoo! News. 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 international and news aggregators 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.

Note that $APstories$ is a continuous measure, so our estimated coefficients represent a percentage change in the odds of visiting vs. not visiting a site. Using the conservative estimate of -0.679, for every one percentage point increase in the fraction of stories from The AP, the odds of visiting a site fall by 0.66 percent.[4]

[4]We also attempt instrumenting for $APstories$ by using the age of the media organization. The estimated coefficients for GLM with instrumental variables is of similar magnitude and sign, though the first stage is weak.

Table 7: The dispute had a larger effect for sites that featured more stories from The Associated Press.

	(1) All	(2) No international	(3) No aggregators	(4) No international & no aggregators
PeriodDispute × Google × APstories	-1.114**	-1.142**	-0.679***	-0.672***
	(0.488)	(0.515)	(0.252)	(0.254)
PeriodDispute × Google	0.0505	0.0592	-0.0735	-0.0751
	(0.126)	(0.135)	(0.0476)	(0.0493)
PeriodDispute	-0.0142	-0.0186	0.0444	0.0425
	(0.0674)	(0.0698)	(0.0354)	(0.0359)
Google	0.0657	0.0453	0.147***	0.133***
	(0.0887)	(0.0950)	(0.0407)	(0.0421)
APstories	1.383***	1.406***	1.383***	1.406***
	(0.193)	(0.193)	(0.193)	(0.194)
PeriodDispute × APstories	0.255	0.256	0.258	0.259
	(0.159)	(0.160)	(0.161)	(0.161)
APstories × Google	-0.253	-0.190	-0.528*	-0.486*
	(0.394)	(0.408)	(0.293)	(0.294)
Week Fixed Effects	Yes	Yes	Yes	Yes
Website Fixed Effects	Yes	Yes	Yes	Yes
Observations	97668	92889	97542	92763

Note: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The outcome variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. The variable *APstories* measures the percentage of stories that were from The Associated Press.

4.2 The Availability of Content from The Associated Press

We also performed an additional falsification check by comparing downstream traffic to Google News during two periods when The AP content was available. We expect no difference in traffic between these two periods, since The AP content was available in both time periods. We collect additional data for December 2009 prior to the dispute and October 2010 when the dispute was resolved; in both months, The AP content was featured on Google News. Table 6 reports the results of the estimation in Column (3). As expected, no distinguishable effect exists between the two time periods, as The AP content was available in both instances for the past 30 days.

4.3 Checking for a Pre-trend

As a final falsification check, we use our sample to test for a pre-trend in the data prior to the dispute. The concern may be that the policy change coincides with a pre-existing trend in the data. We collected data from November and December of 2009, and we re-run our analysis with a dummy variable for the month of December instead of a post variable. As shown in Table 6 Column (4), we do not find evidence of a pre-trend in the months preceding the contract dispute.

Despite these checks, it is important to remember that there are broader limits exist to the generalizability of the experiment we study, since we focus on a particular natural experiment. However, as a single natural experiment, it is reasonably attractive as it captures the dispute between the second largest news aggregator and one of the largest provider of news content in the US. This means that even if our results do not generalize across the industry, the results may themselves be inherently interesting. This is especially the case as news readership can be highly concentrated (Gentzkow and Shapiro, 2011), meaning there is value on focusing on larger players in the industry.

5 Local vs. National Sites

In the previous section, we found that users employ technological advances, such as aggregation, to seek further, more specific information. That is, our results imply that the "traffic effect" dominates the "scanning effect." Given the expansion in users' information set, we next consider what information do users seek and which types of content benefit from aggregation. Depending upon its content, a site 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 downstream traffic to news sites from Google News declined after the removal of The AP content, 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 explore whether the extent of the decline varied by a site's type of content differentiation. News sites may be local in news coverage with a readership that is geographically local, or sites may be national with a broad readership. 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 identify the extent to which a site's content is very local or very national, we examine the top 20 stories featured on the homepage of each news site in our sample, and we identify the fraction of these stories that are local or national news. We classify a story as "local news" if the news is specific to the site's state of headquarters. We classify a story as "national news" if the news focuses on an area outside of the local state headquarters. Note that some stories may not fall into either category such as movie reviews, cultural opinions, etc. To ensure that we properly identify local news stories, we focus on sites affiliated with newspapers or local TV or radio stations in order to exclude travel or leisure sites that

feature local articles, often in the guise of advertising. We classify a news site as "local" if a majority of the stories featured on the site are local news. We classify a news site as "national" if a majority of stories featured on the site are national news.

We estimate a GLM equation similar to equation (1) where we include additional interactions between our measures of local and national news sites. We include the main effect of the policy change on the news category, and we allow an incremental effect if a news site is local or national. As seen in Column (1) of Table 8, our results indicate that visits to sites decrease the most for sites that are either local or national. Columns (2)-(4) check the robustness of the results to alternative definitions of the control group. In general, the results are robust in sign and magnitudes.

Using the odds-ratio interpreation of our results, the odds of visiting a local news site on Google News declined by an extra 18 percent relative to other news sites and compared to Yahoo! News because the exponential of -0.203 is 0.82, thereby one minus 0.82 is 0.18. The odds of visiting a national news site on Google News declined by an extra 26 percent relative to other news sites and compared to Yahoo! News, since one minus the exponential of -0.299 is 0.26.

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.

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 are of higher "quality." First, such sites account for a disproportionate number of visits. For

instance, 25 percent of national 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 national sites. Even though national sites comprise 20 percent of all sites, they account for the majority (60 percent) of winners and finalists for the 13 categories of news reporting in 2009 and 2010.[5]

[5]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 8: The dispute harmed sites that were either local or national

	(1) All	(2) No international	(3) No aggregators	(4) No international & no aggregators
PeriodDispute × Google × Local	-0.203**	-0.203**	-0.203**	-0.203**
	(0.0868)	(0.0868)	(0.0868)	(0.0868)
PeriodDispute × Google × National	-0.299***	-0.299***	-0.299***	-0.299***
	(0.114)	(0.114)	(0.114)	(0.114)
PeriodDispute × Google × News	-0.0870	-0.105	0.0775	0.0779
	(0.171)	(0.186)	(0.0752)	(0.0767)
PeriodDispute	-0.0215	-0.0261	0.0368	0.0347
	(0.0697)	(0.0722)	(0.0397)	(0.0404)
Google	0.0883	0.0622	0.198***	0.184***
	(0.114)	(0.126)	(0.0443)	(0.0461)
PeriodDispute × Google	0.113	0.132	-0.0511	-0.0515
	(0.162)	(0.178)	(0.0516)	(0.0537)
Local	-0.391***	-0.391***	-0.391***	-0.391***
	(0.0861)	(0.0861)	(0.0861)	(0.0861)
PeriodDispute × Local	0.115**	0.115**	0.115**	0.115**
	(0.0467)	(0.0467)	(0.0467)	(0.0467)
Google × Local	0.185	0.185	0.185	0.185
	(0.134)	(0.134)	(0.134)	(0.134)
National	-0.495***	-0.495***	-0.495***	-0.495***
	(0.136)	(0.136)	(0.136)	(0.136)
PeriodDispute × National	0.121*	0.121*	0.121*	0.121*
	(0.0682)	(0.0682)	(0.0682)	(0.0682)
Google × National	0.436**	0.436**	0.436**	0.436**
	(0.192)	(0.192)	(0.192)	(0.192)
News	0.927***	0.943***	0.927***	0.943***
	(0.0782)	(0.0785)	(0.0784)	(0.0788)
PeriodDispute × News	-0.0146	-0.0139	-0.0133	-0.0126
	(0.0541)	(0.0547)	(0.0549)	(0.0555)
News × Google	-0.484***	-0.458***	-0.593***	-0.580***
	(0.157)	(0.165)	(0.116)	(0.117)
Week Fixed Effects	Yes	Yes	Yes	Yes
Website Fixed Effects	Yes	Yes	Yes	Yes
Observations	97704	92925	97578	92799

Note: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The outcome variable is the fraction of traffic to websites after visiting Google News or Yahoo! News. The variable *Local* is a dummy variable that equals one if the majority of stories on the news site were local news; the variable *National* is a dummy variable that equals one if the majority of stories on the news site were national news.

6 Conclusion

In the first wave of digitization of content, peer-to-peer networks disseminated copyrighted content, sometimes illegally. Now the second wave of the digitization of content has led to the rise of platforms which aggregate content in a convenient and unified form for consumers. The new practice of digital aggregation has led to both lawsuits and uncertainty over the economic consequences. To investigate the consequences of these disputes, we examine a breakdown in contract negotiations between The Associated Press (AP) and Google—which prompted Google to stop hosting The AP content. We find when Google News no longer hosted content from The AP, Google News users were less likely to visit other news websites with content related to The AP after visiting Google News relative to Yahoo! News users who experienced no such removal of The AP content. This pattern was driven by a reduction in visits to either local 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 top news websites with acclaimed standards of quality.

This is important both for content producers trying to understand the likely consequences of the rise of these platforms for their business and also for regulators trying to understand whether regulation of such platforms is advisable. In the European Union, for example, legislators have moved to force aggregators to offer an explicit "opt-in" for all their content. Our paper offers an explanation for why content platforms have not "opted-out" of aggregators despite lobbying for laws that require such consent. One possible reason is that ultimately aggregators may benefit platforms, and the purpose of the "opt-in" provision may be to increase bargaining power over payments to news providers.

There are several limitations of this paper. First, as we use data at the aggregate level, we focus on uncovering heterogeneity in responses at the website level rather than the consumer

level. Second, as with any attempt at analyzing a quasi-experiment, limitations may exist both because of the potential endogeneity of actions of agents surrounding the experiment and also its generalizability. Third, we focus on the aggregation of news content which has attracted much press, but may have different search and consumption patterns from other content such as music and movies. Notwithstanding these limitations, we believe this paper provides some useful first evidence about the effects of digital aggregation on the consumption of copyrighted content.

References

Ardia, D. (2008, June 16). Associated Press Sends DMCA Takedown to Drudge Report, 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.

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 (2010). How does pharmaceutical advertising affect consumer 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.

De Los Santos, B., A. Hortacsu, and M. R. Wildenbeest (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review 102*(6), 2955–80.

Eddy, M. (2013). German copyright law targets Google links. *The New York Times*.

Evans, D. and R. Schmalensee (2012). The antitrust analysis of multi-sided platform businesses. *Coase-Sandor Working Paper Series in Law and Economics*.

Gardner, E. (2014, 27 Feb). Spotify settles Ministry of Sound's lawsuit over user playlists. *Hollywood Reporter*.

Gentzkow, M. and J. M. Shapiro (2011). Ideological segregation online and offline. *The Quarterly Journal of Economics 126*, 1799–1839.

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.

Krazit, T. (2010). Google, AP reach deal for google news content. *CNET*.

Lomas, N. (2013, August 13). German publishers can't wean themselves off Google News, despite winning copyright law change. *TechCrunch*.

Masnick, M. (2010, Nov 29). Fox sues woman for $15m because she aggregated tv and movie scripts she found online. *TechDirt*.

Montgomery, A. L., S. Li, K. Srinivasan, and J. C. Liechty (2004). Modeling online browsing and path analysis using clickstream data. *Marketing Science 23*(4), 579–595.

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.

Papke, L. and J. Wooldridge (1996). Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 619–632.

Papke, L. and J. Wooldridge (2008). Panel data methods for fractional response variables with an appliation to test pass rates. *Journal of Econometrics*, 121–133.

Pfanner, E. (2012). A clash across Europe over the value of a click. *The New York Times*.

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.

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*.

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-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

Note: 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.