1 Introduction

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 transaction costs of obtaining content and introducing new information to consumers. While an extensive literature focuses 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 influences consumer search.

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. For identification, we exploit a contract dispute as an exogenous shifter of content on a major new aggregator, Google News. In January 2010, after a breakdown in licensing negotiations, Google removed all news articles that were syndicated by a major content provider, The Associated Press (AP), from its news aggregator (Haddad, 2010). These articles were typically shortened versions of stories that appeared in a select number of newspapers associated with The AP.

Our setting presents an attractive opportunity to study these issues for several reasons. First, content aggregation is a prominent concern in the news media industry 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. Second, our experiment studies a large shock as we focus on arguably two of the largest players in US news media industry, Google News and The Associated Press. 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

increase consumers' consumption of content by 26-30% (Athey and Mobius, 2012; Xu et al., 2014). Back-of-the-envelope calculations suggest that the removal of content potentially led to a decrease of 110 million visits each month from Google News to news media websites hosted in the US. We also explore the institutional relationship between news sites and The AP and nd that websites with stronger ties to The AP su ered a drop in tra c after the dispute.

Our results inform legal and public policy debates. Recent regulation in the European Union attempts 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. 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 of these characteristics appear to encourage users to explore content more deeply instead of scanning content. One surprising development is that despite German 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 bene t 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.

More broadly, our study is related to prior work that describes how digital technologies a ect search costs and generate spillovers (Shapiro and Varian, 1999; Bakos, 1997; Ghose et al., 2011; Greenstein, 2011). The novelty of our study is that we are the rst to explore how access to di erent types of digital content a ects information gathered by consumers. Note that the focus of our study is not how aggregators a ect 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 a ects subsequent navigation by users.

Our results have implications for copyright policy regarding platforms that aggregate digital content. The digital revolution challenges various aspects of copyright protection (Greenstein et al., 2011), but much of the focus has been on peer-to-peer piracy rather than newer legal models of business that aggregate specific types 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 that we find evidence of a traffic effect, even with a relatively large amount of content on an aggregator, is perhaps evidence

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 sta journalists. During the past decade, The AP has been at the forefront of e orts 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 enabled newspapers to pool content and stories in the old media world of physical newspapers and to hence enjoy economies of scale for news reporting in response to the new telegraph technology. 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 compete 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 ts 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 rise of search and aggregation technology for its business model. The Associated Press attempts to grapple with the shift to digital content. As described by (Halberstam, 2007), in 2005, the CEO of the Associated Press stated that \Advertising is following the migrating eyeballs, and new distribution networks are requiring us to rethink how our content reaches consumers."

Since both The AP and Google News are key players in the distribution of news online, it is not surprising they 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, surrounding negotiations of 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 that prevent 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

aggregators to \dig" for more content, then the removal of The AP content will lead to less tra c to news sites.

We empirically test for the two e ects of scanning and tra c in our analyses below.

3.2 Testing for Scanning E ect: Overall Visits to an Aggregator

To test for the scanning e ect, we investigate whether the removal of The AP content from Google News leads to a shift away from Google News. 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. Consequently consumers will shift away from Google News towards other aggregators. This test assumes that content from non-AP sources on Google News is not a close substitute for content from The AP. This assumption may be reasonable in light of the earlier discussion of the role of The AP. Founded in 1846 with an intent to create scale economies associated with telegraph transmission of the news, The AP is one of the largest news agencies in the world, and it is the only national news service in the US. Given that other major competitors are based abroad, it seems plausible that The AP is distinct from other non-AP sources.[2]

We collect 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 a liate 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).

[2] Mark Twain said in 1906, \There are only two forces that can carry light to all corners of the globe, only two, the sun in the heavens and the Associated Press down here" (Halberstam, 2007).

Table 2 reports the number of monthly visits, page views, and visit time to each aggregator during our period of study. Since the content removal occurs at the very end of the month on December 23, 2009, we consider November and December 2009 to be the period before content removal and January 2010 to be the period after content removal. If the scanning e ect persists, we would expect to observe a decline in visits and in reader attention through page views and visit time to Google News relative to Yahoo! News, since Google News would have a lower quality of scanning with the removal of The AP content while Yahoo! News would not. When we compare the metrics, we do not nd 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.

We also collect data from Experian Hitwise to check for any further evidence of changes in behavior. Hitwise \develops proprietary software that Internet Service Providers (ISPs)

metrics suggest that scanning behavior did not change for Google News relative to Yahoo! News after the removal of content from The AP.

In our tests for the scanning e ect, a potential concern may be whether Google News or Yahoo! News experienced any pre-existing trends prior to the dispute. For instance, while we do not nd a relative shift in metrics for Google News compared to Yahoo! News during the period of study, we consider whether we would have expected Google News to fare better than it did in the absence of the content removal. For instance, did Google News trend positively in the months leading to the removal of The AP content and then drop precipitously to the levels of Yahoo! News after the policy change?

To explore this alternative explanation, we collect additional data from Google Trends on weekly search activity for Google News and Yahoo! News for the year leading up to the dispute. In order to document whether any dramatic changes occur over the course of the year prior to December 2009, Figure 3 graphs the search indices for Google News and Yahoo! News from January to December 2009 where the search indices are normalized between 0 and 1. Search activity varies for both news aggregators over the course of the year. Reassuringly, we do not observe declining popularity for Google News in absolute levels or relative to Yahoo! News during the period prior to the dispute.

For a more formal test, we use the data from Google Trends to estimate a di erence-in-di erence analysis of search intensity for news aggregator j in week t:

$$\begin{aligned} searches_{jt} &= {}_0 + {}_1 Google_j \quad APContentRemoval_t \\ &+ {}_2 Google_j + week_t + {}_{jt} \end{aligned} \tag{1}$$

where *Google* is an indicator variable equal to one for searches on Google News, and *APContentRemoval* is an indicator variable equal to one for the weeks after the removal of The AP content from Google News on December 23, 2009. The vector *week* contains

weekly xed e ects to capture national variation in the volume and interest generated by

e ects. Potentially the two e ects are separate, as each relies on a di erent distribution of users. For instance, we may expect users who scan and skim content to di er from users who seek further information and drive tra c. In the aggregate, we nd that the overall set of customer characteristics across Google News and Yahoo! News are similar according to Table A-1 in the Appendix.[3] Therefore, as long as those customer characteristics adequately predict the way in which consumers use news aggregators, it is likely that the scanning and tra c e ects do not interact di erently across the two aggregators.

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 competition to other aggregators. We do not nd evidence that competing aggregators act as substitute platforms for one another. Perhaps if the reduction in quality a ected the customization of the website, we would observe an e ect. For instance, Athey and Mobius (2012) nd that the ability to customize stories to local content on Google News increased adoption of the news aggregator.

Our results should be interpreted with two caveats. First, it is possible that the length of the period in which The AP content was unavailable may have been too short to have had a noticeable impact, as users did not have su cient time to notice the change in quality and switch to alternative aggregators. This is one explanation for the lack of e ect we measure. Of course, unlike other products that are consumed less frequently or where quality is not readily ascertainable, the timeliness of news and the regularity of its consumption suggests that it may be reasonable to assume that enough consumers did visit the website and observe changes in quality. Nevertheless, our results should be interpreted with the caveat that consumers' awareness of the policy change may play a role. Second, even if readers are aware of the change, it is also possible that switching costs prevent readers from changing their behavior much. For instance, a reader may prefer to remain in the Google \environment" for other related services such as mail, search, etc. | all of which are easily

[3] Hitwise reports the fraction of users within each demographic category for a particular site.

accessible by embedded icons on the Google webpage. Typing in an alternative web address to navigate to a di erent news aggregator presents a higher cost than clicking on an icon, so such switching costs, while small, are probably present.

3.3 Testing for Tra c E ect: Downstream Visits after an Aggregator

tional and aggregator sites, and the average percentage of downstream visits they receive. As shown in Table 6, the top non-news websites re ect the top website brands on the Internet.

To verify that Yahoo! News could be considered an appropriate control group for Google News, we check 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 signi cantly 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.

Our preliminary analysis examines visits to news sites after navigating to an aggregator. Figure 4 illustrates the aggregate mean percentage of downstream tra c 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 tra c 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 5 illustrates that no such change in visits occurred between December 2008 and January 2009.

To formalize the insights provided by Figure 4, we run a di erence-in-di erences regression for the policy change and estimate the percentage of visits to website i after visiting news aggregator j in week t. We use a Generalized Linear Model (GLM) framework to

the policy change, then the measure of the share of outgoing tra c 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 tra c 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 tra c from each news aggregator separately, the share of outbound tra c to news sites would remain the same even with shifts in total tra c to each news site.

We estimate this speci cation 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 $g(\cdot)$ 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).

Our dependent data is proportional data and bounded between zero and one, so panel data methods propose the GLM framework to keep predicted values within the unit interval (Papke and Wooldridge, 2008). The GLM framework also maintains the advantage that no *ad hoc* transformations are required to handle the data at the extreme values near zero and one (Papke and Wooldridge, 1996). Given that the share of visits to a particular site may be small, this is attractive in our empirical setting. Furthermore, we interpret the results as an odds ratio. Because we use the binomial family with logit link, the coe cients may be used to compute an odds ratio, i.e., ratio of probabilities of success (visiting a news site) and of failure (visiting a non-news site) between Google News and Yahoo! News.

As Lechner (2010) points out, an arithmetic di erence between the expected value of the dependent variable before and after treatment will not di erence out the common trend for the group. Instead, because we work with fractional response data, we are able to take the ratio of the probabilities to compute the odds for the treatment and control group, and

the common trend for each group will proportionally di erence out.[6] We generalize to the di erence-in-di erences framework by computing the ratio of the odds for visiting news and non-news sites for the treatment and control groups of Google News and Yahoo! News:

$$\left[\frac{[Odds|Google=1;Post=1]}{[Odds|Google=1;Post=0]}\Big/\frac{[Odds|Google=0;Post=1]}{[Odds|Google=0;Post=0]}\right] = exp(\ _3) \tag{3}$$

where $G(E(Y)) = \ _0 + \ _1 News \ \ Google \ \ Post + \ _2 News \ \ Post + \ _3 News \ \ Google + \ _4 Google + \ $, and $Post$ is a dummy variable equal to one after the policy change.[7] To facilitate interpretation, we report the exponentiated coe cients or odds ratio for our results under the GLM logit link.

Table 7 reports the results in Column (1) for the full speci cation as described by equation (2).[8] The time period covers December 2009 to January 2010, which encompasses the weeks before and after the dispute on December 23, 2009 between The AP and Google News. In our setting, the odds are the probability of visiting a news site compared to the probability of not visiting a non-news site, or the share of visits to news site compared to the share of visits to all non-news sites. Note that since the coe cients are exponentiated, we interpret them and test for statistical signi cance relative to the value of one, which represents no e ect ($exp(0) = 1$). In other words, if the policy change has a negative e ect (coe cient is less than zero), we would expect an odds ratio or exponentiated coe cient to be less than

[6] For a simple example, suppose that $E(Y_i) = \ _0 + \ _1 X_i + \ _i$ where Y_i is the fraction of successes from n_i binomial trials for observation i, and X is a dummy variable equal to 0 or 1. Using the logit link for our link function $G(:)$, the odds are $E(Y)=(1 \ \ E(Y)) = exp(\ _0 + \ _1 X_1 + \ )$. Consequently, the odds for $X = 1$ are $(OddsjX = 1) = E(YjX = 1)=(1 \ \ E(YjX = 1)) = exp(\ _0 + \ _1 + \ )$, and the odds for $X = 0$ are $(OddsjX = 0) = E(YjX = 0)=(1 \ \ E(YjX = 0)) = exp(\ _0 + \ )$. We simplify the ratio of the two odds as $(OddsjX = 1)=(OddsjX = 0) = exp(\ _1)$.

[7] This is a simpli ed version of Equation r(0 [(j)]1717801)]1717801rum8I.68.-2(1))]TJ/F11 9.9626 Tf 19.372 0 Td [(=)]TJ/F

one; if the policy change has a positive e ect (coe cient is greater than zero), we would expect an odds ratio or exponentiated coe cient to be greater than one.

If news aggregators complement the news sources that they feature, we would expect the aggregators to direct further tra c to news sites. Accordingly, if the hosting of content from The AP by Google News prompts readers to seek further information, we would expect the removal of content from The AP to lead to a decline in referrals from Google News to other news sites. Since Table 7 reports the exponentiated coe cients, we would expect $_1$ to be negative, and therefore the odds ratio or the exponentiated coe cient of $_1$ to be less than one.

We nd that the exponentiated coe cient of 0.72 on *APContentRemoval Google News* indicates that the odds ratio is 0.72 or 72 percent of its level prior to the policy change. 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. 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 tra c to these news sites.

We also collect additional data 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 7, Column (2) compares January 2010 (when no content from The AP was available) to October (when The AP content was fully reinstated). If the complementary relationship between Google News and The AP exists, then we would expect an increase in visits to other news sites when Google News restores content from The AP.

As expected, we observe a positive e ect; as the odds ratio of visiting a news sites increased by 96% 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 \tra c e ect" does exist and that the relationship is complementary. Consumers do appear to use platforms to seek new and further content. One striking feature of how 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 displayed rather than merely a snippet. In light of this, our evidence of a tra c e ect rather than merely a scanning e ect is striking.

News sites on Google experience a 28 percent decrease in visits after the removal of The Associated Press articles. The magnitude of this estimate is comparable to concurrent and prior studies of technological adoption on consumers' consumption of content. Athey and Mobius (2012) study how the adoption of Google News toolbar a ects consumption of local news and nds that the additional content increases local news consumption by more than 26%, and over time (beyond an 8-week period) the increase persists at more than 14%. In our study, The AP represents both local and national news, so we would expect our results for tra c to be comparable. Furthermore, Xu et al. (2014) study how the adoption of a news app increases the probability of visiting a news site by nearly 30%.

If the claim in Cohen (2009) is true that Google sends a billion clicks each month to its partner news providers, then this percentage translates into a very large change in the number of clicks that news websites receive from Google News. While we do not know precisely the international breakdown, our data from Hitwise suggest that before the policy change, news media websites hosted in the US account for 40 percent of all clicks for the subset of users who use Google News. Therefore, this 28 percent decrease could imply an approximately 110 million decrease in visits each month from Google News users to news

media websites hosted in the US.[9]

As seen in Figure 1, The AP holds the topmost position under \Top Stories" on Google News. The large e ects of the policy may be expected if Google News places The AP in a prominent position on its website. While the exact algorithm is unknown for how Google News decides which stories to include and where to place them on its site, we do know that recency and popularity of the story play key factors. As The AP is the only remaining national newswire agency in the US, it is likely that The AP would provide breaking news on recent events and be very popular among readers. Consequently, The AP may often hold a prominent position on the page of Google News, and therefore we would expect the removal of The AP content to have a relatively large e ect on tra c from Google News.

When we consider the reinstatement, an even larger measured increase exists with the point estimate suggesting that clicks increase by 95 percent. One potential reason for the larger size of the reinstatement e ect is that when Google restored all the AP articles at once in fall 2010, the articles dominated the most valuable position on the webpage in Google News.

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 tra c e ect 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 and disseminate from The AP on their own site. We collect data on the number of stories featured from The AP on each news site.[10] We create a measure *APstories* of the fraction

[9] Forty percent of 1 billion clicks is 40 million clicks. A 28 percent reduction of 40 millions clicks is a decrease of 112 million clicks.

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

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 article and therefore removed from Google News when The AP hosted content was removed. The measure *APstories* re ects the strength of the relationship between the news site and The AP.

This data was collected from a cross-section of news sites in 2015. Our assumption is that the relationship between the news sites and The AP stories had a common trend over the years. This seems reasonable given that The Associated Press has a long history of relationships with newspapers, which seem unlikely to change quickly. The biggest concern is that newspapers who faced nancial turmoil could have severed links to The AP to save money, potentially in uencing our results. However, conversations with industry insiders suggest to us that newspapers facing nancial constraints are more likely to stop producing their own content rather than severing ties with The Associated Press. Therefore, it seems likely that The Associated Press relationships with newspapers have not changed dramatically in the time between 2009 and 2016.

Table 8 reports the results of equation (2) with our continuous measure *APstories* as our variable of interest instead of the indicator variable *News*. Once again, we report the exponentiated coe cients or odds-ratios. Consistent with our results on the tra c e ect, we nd that sites that are more embedded in The AP ecosystem experienced a larger drop in tra c relative to Yahoo! News. Columns (2)-(4) check robustness of the results to alternative de nitions 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

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 the reported results represent a percentage change in the odds of visiting compared to not visiting a news site. Using the conservative estimate of 0.52, for every one percentage point increase in the fraction of stories from The AP, the odds of visiting a site fall by 0.48 percent.[11]

4.2 The Availability of Content from The Associated Press

We also perform 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 on traffic during October 2010 when the dispute was resolved and compare this to traffic in December 2009; in both months, The AP content was featured on Google News. Table 7 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. The odds ratio is not statistically significant from one; in other words, the odds of visiting a news site did not change between these two periods.

4.3 Checking for a Pre-trend

As a final set of falsification checks, we 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 weekly data from November until the dispute in December 23, 2009, and we re-run our analysis with a dummy variable for the weeks in December instead of a post

variable. As shown in Table 7 Column (4), we do not find evidence of a pre-trend in the months preceding the contract dispute. The odds ratio is not significantly different from

odds-ratios of our results, so we interpret and test our estimated coe cients for statistical

6 Conclusion

despite lobbying for laws that require such consent. One possible reason is that ultimately aggregators may bene t platforms, and the purpose of the opt-in provision may be to increase bargaining power over payments to news providers.

Several limitations of this paper exist. First, our data is at the aggregate level, so we focus on uncovering heterogeneity in responses at the website level rather than at the consumer level. Thus we may not uncover other moderating factors that could explain the propensity to use new aggregators for scanning or tra c purposes. 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 because of questions over its generalizability. For example, the relative improvement in tra c from Google News to The AP websites may re ect improved terms in the deal they struck, which we do not observe. Third, we focus on the aggregation of news content which has attracted a lot of attention, but may have di erent search and consumption patterns from other content such as music and movies. Consequently our results may not generalize broadly to other content platforms. Notwithstanding these limitations, we believe this paper provides useful rst evidence about the e ects of digital aggregation on the consumption of copyrighted content.

Table 2: Monthly visits, page views, and visit time to Google News and Yahoo! News

Metric	Date	Google News	Yahoo! News
Visits	November 2009	75,667,000	224,160,000
Visits	December 2009	78,160,000	267,570,000
Visits	January 2010	79,373,000	262,700,000
Page views	December 2009	3%	7%
Page views	January 2010	3%	7%
Visit time	December 2009	22 seconds	5 seconds
Visit time	January 2010	22 seconds	5 seconds

Source: ComScore and Hitwise

Table 3: No evidence of scanning e ect during removal and reinstatement of content from The Associated Press

		(1)	(2)	(3)	(4)
APContentRemoval	Google	0.0485			
		(0.0763)			
APContentRestored	Google		-0.0720	-0.0235	
			(0.0707)	(0.0532)	
December	Google				0.153
					(0.0875)
APContentRemoval		-0.0143			
		(0.0664)			
Google		0.417	0.466	0.417	0.265
		(0.0419)	(0.0629)	(0.0419)	(0.0764)
APContentRestored			0.0260	0.167	
			(0.0841)	(0.0360)	
December					-0.0963
					(0.0663)
Observations		18	20	18	16

Note: Robust standard errors. *p < 0:1, **p < 0:05, ***p < 0:01. The outcome variable is the Google Trends index of search intensity for Google News or Yahoo! News, which is normalized between 0 and 1. Column (1) covers the period from December 2009 to January 2010, before and after the removal of The AP content from Google News on December 23, 2009. Column comoa29 fter the

Table 4: 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 weekly 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 de ned by Hitwise, excluding weather sites, international news sites, and top news aggregators.

Table 5: Top 40 news websites visited after Google News and Yahoo! News

	Avg Visit Pct
nytimes.com	0.029
abcnews.com	0.021
cnn.com	0.019
washingtonpost.com	0.017
wsj.com	0.017
nydailynews.com	0.016
reuters.com	0.014
examiner.com	0.013
time.com	0.012
foxnews.com	0.011
politico.com	0.011
msnbc.com	0.0083
people.com	0.0078
usatoday.com	0.0072
bloomberg.com	0.0051
nypost.com	0.0051
boston.com	0.0048
latimes.com	0.0048
usmagazine.com	0.0046
mercurynews.com	0.0044
edition.cnn.com	0.0040
bostonherald.com	0.0038
cbsnews.com	0.0037
pcworld.com	0.0037
sfgate.com	0.0033
npr.org	0.0032
businessweek.com	0.0031
csmonitor.com	0.0030
miamiherald.com	0.0030
philly.com	0.0030
theweek.com	0.0029
chron.com	0.0027
voanews.com	0.0026
freep.com	0.0025
seattletimes.nwsource.com	0.0022
dallasnews.com	0.0021
mcclatchydc.com	0.0019
startribune.com	0.0017
wired.com	0.0017

Table 6: Top 40 Non-news websites visited after Google News and Yahoo! News

	Avg Visit Pct
google.com	0.12
mail.yahoo.com	0.099
yahoo.com	0.072
facebook.com	0.062
search.yahoo.com	0.044
youtube.com	0.025
gmail.com	0.015
myspace.com	0.015
my.yahoo.com	0.013
mail.live.com	0.013
nance.yahoo.com	0.012
howlifeworks.com	0.010
msn.com	0.010
images.google.com	0.010
ebay.com	0.0100
cosmos.bcst.yahoo.com	0.0095
weather.yahoo.com	0.0079
blogsearch.google.com	0.0077
livescience.com	0.0075
nance.google.com	0.0072
weather.com	0.0067
omg.yahoo.com	0.0064
bing.com	0.0062
amazon.com	0.0059
members.yahoo.com	0.0058

Table 7: Downstream tra c from Google News and Yahoo! News during removal and reinstatement of content from The Associated Press

			(1)	(2)	(3)	(4)
			Removal	Reinstatement	Falsi cation	Falsi cation
APContentRemoval	Google	News	0.718			
			(0.127)			
APContentRestored	Google	News		1.955	1.347	
				(0.553)	(0.259)	
December	Google	News				0.964
						(0.115)
APContentRemoval	Google		1.120			
			(0.181)			
Google			1.092	1.187	1.115	1.291
			(0.125)	(0.142)	(0.0862)	(0.0768)
APContentRemoval	News		1.114			
			(0.0703)			
News	Google		0.811	0.575	0.718	0.667
			(0.109)	(0.0837)	(0.0849)	(0.0563)
APContentRestored	Google			0.902	1.016	
				(0.216)	(0.131)	
APContentRestored	News			0.601	0.680	
				(0.0899)	(0.0869)	
December	Google					0.998
						(0.105)
December	News					1.114
						(0.121)
Week Fixed E ects			Yes	Yes	Yes	Yes
Website Fixed E ects			Yes	Yes	Yes	Yes
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 tra c to websites after visiting Google News or Yahoo! News. The exponentiated coe cients are reported with the corresponding standard errors for interpretation as odds ratios. Note that since the coe cients are exponentiated, we interpret them and test for statistical signi cance relative to the value of one, which represents no e ect ($exp(0) = 1$). In other words, if the policy change has a negative e ect (coe cient is less than zero), we would expect an odds ratio or exponentiated coe cient to be less than one; if the policy change has a positive e ect (coe cient is greater than zero), we would expect an odds ratio or exponentiated coe cient to be greater than one. Column (1) covers the period from December 2009 to January 2010, before and after the removal of The AP content from Google News on December 23, 2009. Column (2) compares January 2010 (when no content from The AP was available) to October 2010 (the restoration of hosted articles by The Associated Press in Google

Table 8: The dispute had a larger e ect for sites that featured more stories from The Associated Press.

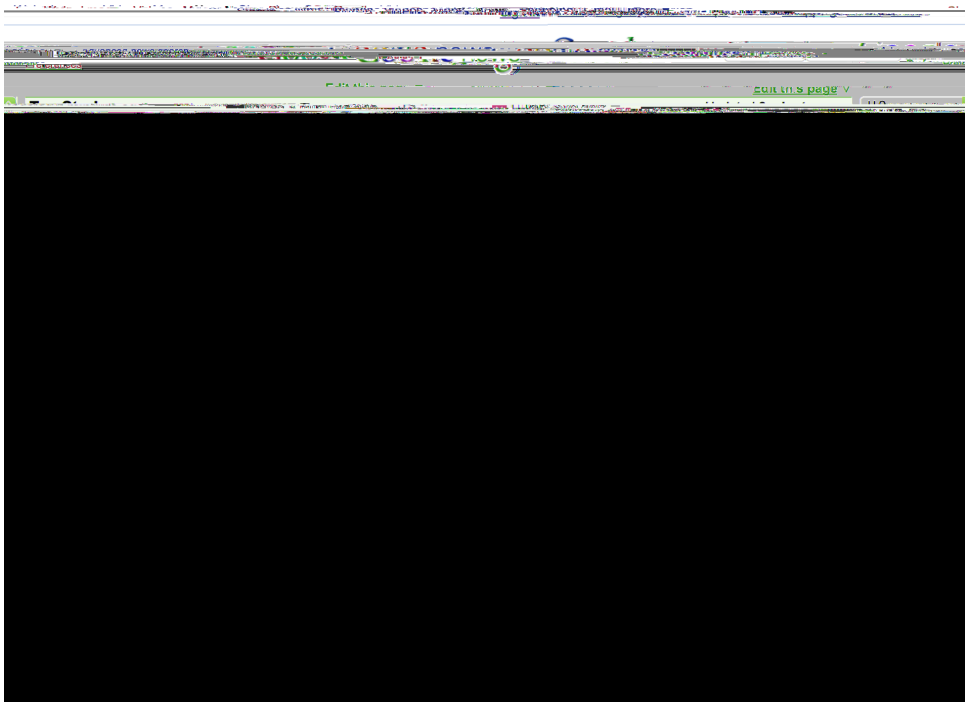
	(1) All	(2) No international	(3) No aggregators	(4) No international & no aggregators
PeriodDispute Google APstories	0.328**	0.319**	0.507***	0.511***
	(0.160)	(0.164)	(0.128)	(0.130)
PeriodDispute Google	1.052	1.061	0.929	0.928
	(0.133)	(0.144)	(0.0442)	(0.0457)
Google	1.068	1.046	1.158***	1.143***
	(0.0947)	(0.0994)	(0.0472)	(0.0481)
APstories	3.987***	4.080***	3.986***	4.079***
	(0.768)	(0.789)	(0.770)	(0.791)
PeriodDispute APstories	1.291	1.292	1.294	1.296
	(0.205)	(0.207)	(0.208)	(0.209)
APstories Google	0.776	0.827	0.590*	0.615*
	(0.306)	(0.337)	(0.173)	(0.181)
Website Fixed E ects	Yes	Yes	Yes	Yes
Week Fixed E ects	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 tra c to websites after visiting Google News or Yahoo! News. The table covers the period from December 2009 to January 2010. The variable *APstories* measures the percentage of stories that were from The Associated Press. The exponentiated coe cients are reported with the corresponding standard errors for interpretation as odds ratios.

Table 9: The dispute harmed sites that were either local or national

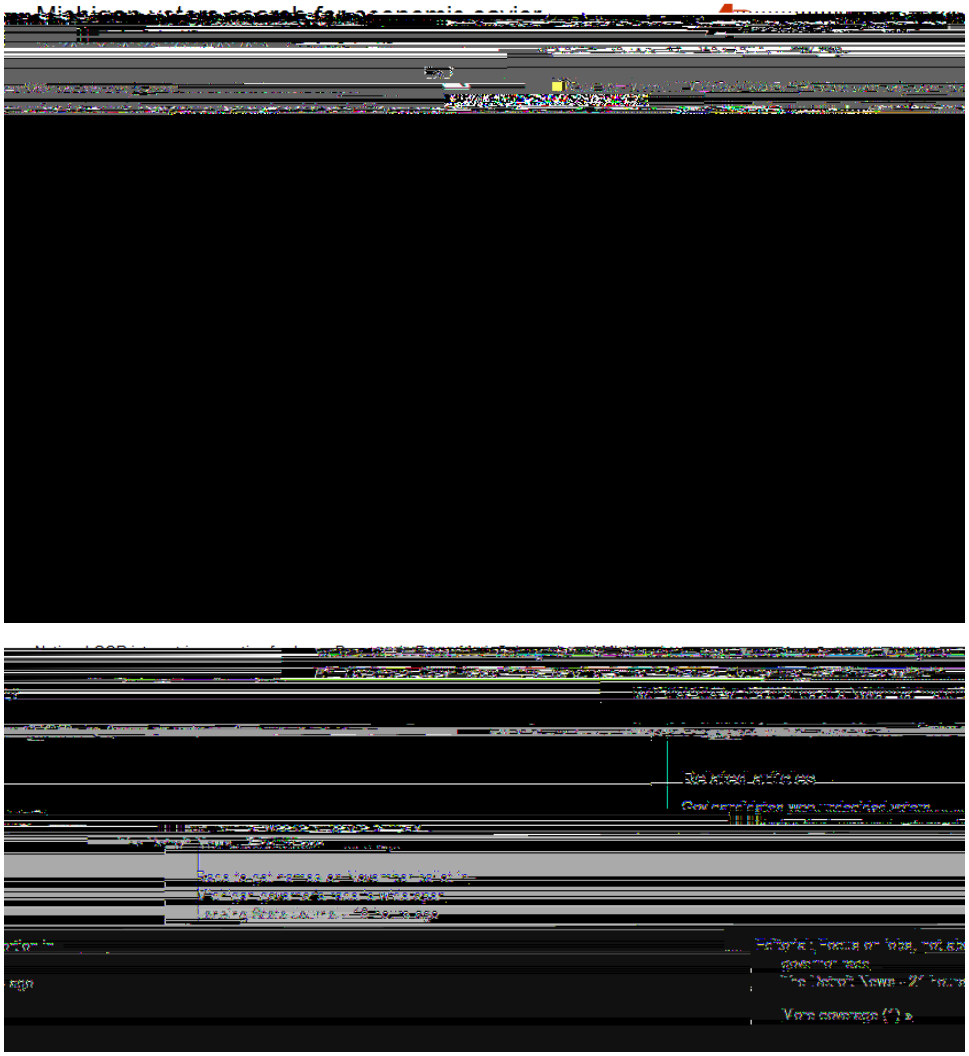
			(1) All	(2) No international	(3) No aggregators	(4) No international & no aggregators
PeriodDispute	Google	Local	0.817	0.817	0.817	0.817
			(0.0709)	(0.0709)	(0.0709)	(0.0709)
PeriodDispute	Google	National	0.742	0.742	0.742	0.742
			(0.0844)	(0.0844)	(0.0844)	(0.0844)
PeriodDispute	Google	News	0.917	0.900	1.081	1.081
			(0.156)	(0.167)	(0.0812)	(0.0829)
Google			1.092	1.064	1.219	1.203
			(0.125)	(0.134)	(0.0540)	(0.0554)
PeriodDispute	Google		1.120	1.141	0.950	0.950
			(0.181)	(0.203)	(0.0490)	(0.0510)
PeriodDispute	Local		1.122	1.122	1.122	1.122
			(0.0524)	(0.0524)	(0.0524)	(0.0524)
Google	Local		1.203	1.203	1.203	1.203

Figure 1: Screenshot of Google News



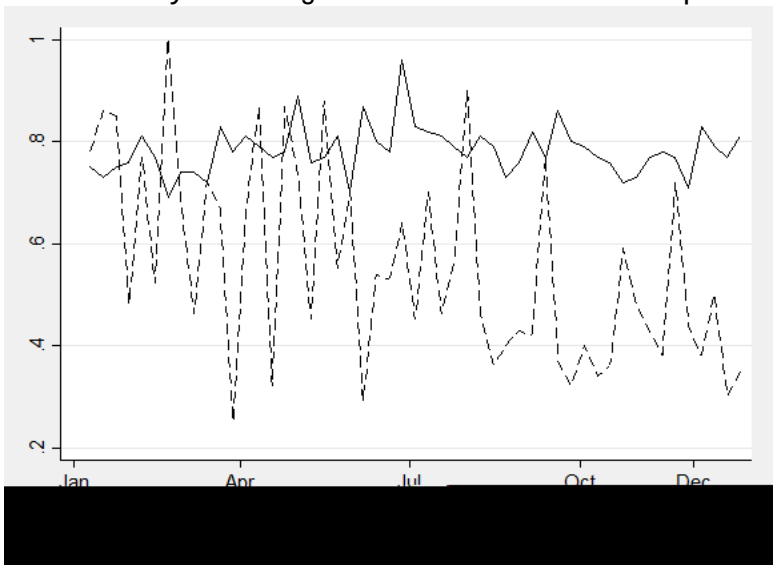
Note: Source: Wayback Machine (https://archive.org/web/) July 15, 2009.

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



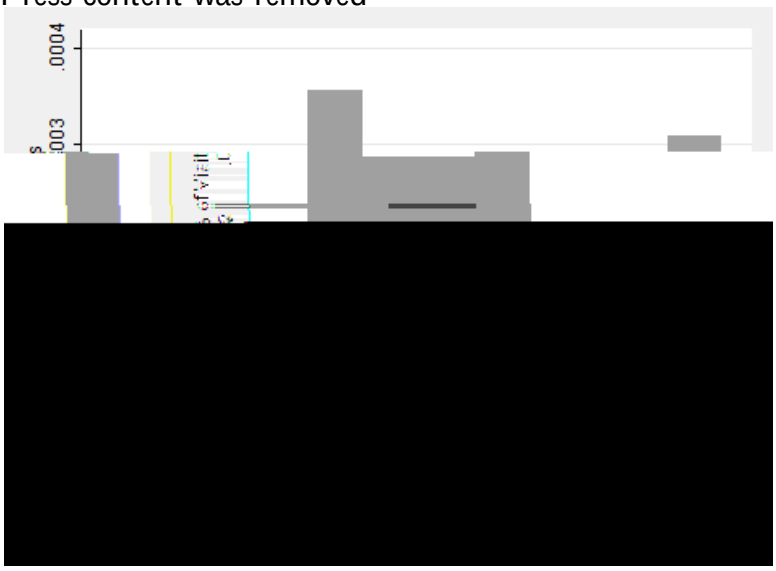
Note: Google News, August 2010. Text of article has been slightly edited to fit on page.

Figure 3: Search activity for Google News and Yahoo! News prior to the dispute



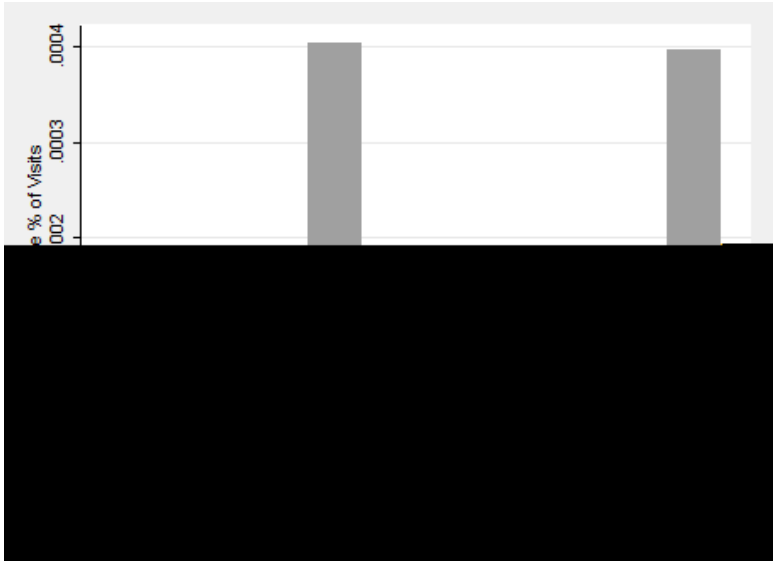
Note: This gure graphs the search index of Google News and Yahoo! News from Google Trends, which is normalized between 0 and 1. The vertical axis is the index of search intensity, and the horizontal axis is time from January to December 2009.

Figure 4: Downstream sites visited after Google News and Yahoo! News before and after The Associated Press content was removed



Note: This gure 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 5: Downstream sites visited after Google News and Yahoo! News in prior year when no content was removed (December 2008 and January 2009)



Note: This gure 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.

References

Ardia, D. (2008, June 16). Associated Press Sends DMCA Takedown to Drudge Report, Backpedals, and Now Seeks to De ne 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. Du o, and S. Mullainathan (2004). How much should we trust di erences-in-di erences estimates? *The Quarterly Journal of Economics 119* (1), 249{275.

Chiou, L. and C. Tucker (2010). How does pharmaceutical advertising a ect 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*.

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

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*.

Xu, J., C. Forman, J. Kim, and K. Van Ittersum (2014). News Media Channels: Complements or Substitutes? Evidence form Mobile Phone Usage. *Journal of Marketing 78*, 97{112.

Appendix

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

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 percentage of users within each demographic category. Statistics are reported for users of Yahoo! News, Google News, and the New York Times website.

Table A-2: Linear model of downstream tra c from Google News and Yahoo! News during removal and reinstatement of content from The Associated Press

			(1) Removal	(2) Reinstatement	(3) Falsi cation	(4) Falsi cation
APContentRemoval	Google	News	-0.00742			
			(0.00316)			
APContentRestored	Google	News		0.0102	0.00247	
				(0.00441)	(0.00434)	
December	Google	News				0.000126
						(0.00246)
APContentRemoval	Google		0.00147			
			(0.00218)			
Google			-0.0102	-0.00715	-0.0105	-0.0117
			(0.00569)	(0.00543)	(0.00475)	(0.00569)
APContentRemoval	News		0.00176			
			(0.00114)			
News	Google		0.0360	0.0245	0.0365	0.0379
			(0.00805)	(0.00728)	(0.00736)	(0.00818)
APContentRestored	Google			-0.00121	0.000197	
				(0.00279)	(0.00177)	
APContentRestored	News			-0.00697	-0.00505	
				(0.00263)	(0.00213)	
December	Google					0.0000501
						(0.00180)
December	News					0.00168
						(0.00186)
Week Fixed E ects			Yes	Yes	Yes	Yes
Website Fixed E ects			Yes	Yes	Yes	Yes
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 percentage of tra c (measured between 0 and 100) to a website after visiting Google News or Yahoo! News. The speci cations are estimated with OLS. 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 falsi cation check compares December 2009 and October 2010 when The AP content was available in Google News in both months. In Column (4), the falsi cation 010 whenths. Inal 2L C(051,)-3628(t)-302(w)29 n05

Table A-3: Downstream tra c for top sites from Google News and Yahoo! News

	(1) Removal	(2) Reinstatement	(3) Falsi cation	(4) Falsi cation
APContentRemoval Google News	0.569			
	(0.175)			
APContentRestored Google News		2.666	1.451	
		(1.369)	(0.477)	
December Google News				0.689
				(0.227)
APContentRemoval Google	1.304			
	(0.374)			
Google	0.700	0.914	0.817	0.809
	(0.261)	(0.177)	(0.104)	(0.0869)
APContentRemoval News	1.193			
	(0.120)			
APContentRestored Google		0.749	0.990	
		(0.333)	(0.233)	
APContentRestored News		0.524	0.635	
		(0.137)	(0.139)	
December Google				1.010
				(0.238)
December News				1.575
				(0.346)
Week Fixed E ects	Yes	Yes	Yes	Yes
Website Fixed E ects	Yes	Yes	Yes	Yes
Observations	1746	1960	1755	1755

Note: Robust standard errors clustered at website level. *$p < 0.1$, **$p < 0.05$, ***$p < 0.01$. The outcome variable is the fraction of tra c to a website after visiting Google News or Yahoo! News. The speci cations are estimated using the sample of websites with tra c levels among the top 100. The exponentiated coe cients are reported with the corresponding standard errors for interpretation as odds ratios. Note that since the coe cients are exponentiated, we interpret them relative to the value of one, which represents no e ect ($exp(0) = 1$). In other words, if the policy change has a negative e ect (coe cient is less than zero), we would expect an odds ratio or exponentiated coe cient to be less than one; if the policy change has a positive e ect (coe cient is greater than zero), we would expect an odds ratio or exponentiated coe cient to be greater than one. 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 falsi cation check compares December 2009 and October 2010 when The AP content was available in Google News in both months. In Column (4), the falsi cation check compares November 2009 and December 2009 to test for a pre-trend.

Table A-4: Robustness check of downstream tra c from Google News and Yahoo! News during removal and reinstatement of content from The Associated Press

	(1) Removal	(2) Reinstatement	(3) Falsi cation	(4) Falsi cation
APContentRemoval Google News	0.718			
	(0.127)			
APContentRestored Google News		1.955	1.347	
		(0.553)	(0.259)	
December Google News				0.964
				(0.115)
APContentRemoval Google	1.118			
	(0.172)			
Google	1.690	1.978	1.643	1.596
	(3.518)	(2.908)	(1.546)	(1.830)
APContentRemoval News	1.114			
	(0.0703)			
News Google	0.811	0.575	0.718	0.667
	(0.109)	(0.0837)	(0.0849)	(0.0563)
APContentRestored Google		1.375	1.386	
		(1.604)	(1.139)	
APContentRestored News		0.601	0.680	
		(0.0899)	(0.0869)	
December Google				1.045
				(0.161)
December News				1.114
				(0.121)
Week Fixed E ects	Yes	Yes	Yes	Yes
Website Fixed E ects	Yes	Yes	Yes	Yes
Trends	Yes	Yes	Yes	Yes
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 tra c to websites after visiting Google News or Yahoo! News. All regressions include direct controls for trends in visits to Google News and Yahoo! News. The exponentiated coe cients are reported with the corresponding standard errors for interpretation as odds ratios. Note that since the coe cients are exponentiated, we interpret them relative to the value of one, which represents no e ect ($exp(0) = 1$). In other words, if the policy change has a negative e ect (coe cient is less than zero), we would expect an odds ratio or exponentiated coe cient to be less than one; if the policy change has a positive e ect (coe cient is greater than zero), we would expect an odds ratio or exponentiated coe cient to be greater than one. 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 falsi cation check compares December 2009 and October 2010 when The AP content was available in Google News in both months. In Column (4), the falsi cation check