Social Media Sharing and Online News Consumption

Sinan Aral*† (sinan@mit.edu)
MIT Sloan School of Management

Michael Zhao* (mfzhao@mit.edu)
MIT Sloan School of Management

February 4, 2019

Abstract

Publishers' strategies in the digital age depend critically on whether social media substitutes for or drives online news consumption. Unfortunately, a number of empirical challenges, including endogeneity and the confounding effect of inherent newsworthiness, make identifying the causal effect of social media on news consumption difficult. We overcome these challenges by leveraging exogenous variation in regional news consumption to identify the cross region viewership effects of news sharing on social media. We apply this method to a novel data set that combines over 200 million page views of the New York Times (NYT) between April 8, 2013 to October 27, 2013 with data on daily weather patterns collected from 45,000 weather stations worldwide and a Twitter follower graph constructed from 10,000 Twitter users who shared NYT content with their 200,000 followers. Exploiting regional rainfall as a natural experiment, we first identify positive and significant cross-region peer effects in online news consumption. A 1% increase in aggregate viewership outside of a focal region causes viewership in that region to increase by approximately 0.34%. Further analysis suggests that social media is the primary driver of these cross-region peer effects. Specifically, cross-region peer effects in viewership are stronger when referred by social media than when referred by search engines. Furthermore, regions that are strongly-connected with each other on social media exhibit more positive and significant peer effects than regions that are weakly-connected. Taken together, these results suggest that social media sharing drives online news consumption.

1 Introduction

New media technologies have historically disrupted the contemporaneous media of the time, a phenomenon perhaps best captured by the 1979 pop song "Video Killed the Radio Star." Social media is no exception, directly competing with traditional media channels for people's time and attention. One key difference with social media, however, is that it may amplify the content of other media channels as well. These countervailing effects create a puzzle for media organizations and content producers: is social media helping or hurting their businesses?

*Sinan Aral and Michael Zhao contributed equally to this work.
†Corresponding Author.

The disruptive effects of social media show up in well-defined metrics like time spent (GlobalWebIndex 2017) and advertising market shares (McKinsey 2016). On the other hand, there are anecdotal examples of viral content that generates millions of dollars in online advertising. The capacity to monetize content viewership has spawned new professions such as full-time bloggers, YouTubers, and Instagram influencers. Even the course of popular culture has likely been shifted. In the absence of social media, would musicians like Justin Bieber or Ariana Grande be discovered? Would non-English songs like Despacito or Gangnam Style manage to reach the top of the US charts? At the core of this potential is a positive feedback loop (or "virtuous cycle"): the greater the viewership of a piece of content, the more likely it is to be widely shared; at the same time, as a piece of content continues to be shared, the larger and larger the potential audience. Unfortunately, the potential content amplification effects of social media, while anecdotally striking, are harder to measure systematically.

In this study, we investigate whether social media sharing drives online content consumption, focusing specifically on the online content of newspapers. Although newspapers have survived the technological disruptions of radio and television, they have experienced major difficulties since the turn of the new millennium. Pew Research reports that US newspaper circulation fell from 55.8 million in 2000 to 34.7 million in 2016[1] (Barthel 2017). Even more worrisome are the tremendous declines in advertising revenue, from a high of $49.4 billion in 2005 to just $18.3 billion in 2016. In response to such major losses of revenue, many newspapers have declared bankruptcy. Others have scaled back operations by laying off staff or reducing print publication. While some prominent newspapers seem to have found a degree of success by shifting to "subscription-first" business models, the importance of digital advertising for revenue generation is still readily apparent—growing from 17% of newspapers' advertising revenue in 2011 to 29% in 2016. Moreover, digital news consumption has increased dramatically, now with 93% of US adults consuming at least some of their news content from an online source (Stocking 2017). While social media is a major source of traffic for digital news (Newman et al. 2017), there are concerns that the distribution of news content on social media may actually be cannibalizing viewership.

With social media use increasing, it has never been more critical for news organizations to determine how social media sharing is impacting content viewership. However, there are two main empirical challenges to estimating this relationship. First, there are serious endogeneity concerns due to confounding factors like "inherent newsworthiness" and simultaneous equation bias from the feedback loop between viewership and sharing. Although instrumental variables (IV) estimation can address both of these issues (Hausman and Taylor 1983), valid instruments are difficult to find in this context. Second, the data collection requirements are daunting (Luca 2015). Ideally, one needs detailed data from both social media platforms and news organizations. However, it seems unlikely that a news organization could obtain fine-grained social media sharing data, especially given the recent negative press surrounding the data practices of social media platforms. It seems

[1] It is worthwhile to note that circulation has been steadily falling since 1990, however the rate of decline noticeable accelerated since 2005.

similarly doubtful that a social media platform could obtain the proprietary browsing data of a newspaper.

To address these challenges, we develop a novel empirical approach by combining two key ideas: regional localization and a peer effects framing. By regional localization we mean our analysis is conducted at the level of localized geographic regions. This allows us to take advantage of "local" instruments—variables that shock the viewership of particular regions but not others—to address the econometric concerns. We also redefine our objective in terms of identifying causal peer effects in news consumption rather than directly examining the cross-region effect of social media sharing on content viewership. This shift to a peer effects framework obviates the need to acquire high-quality social media sharing data, drastically simplifying the data requirements.[2] We use this approach to study online viewership at the New York Times (NYT). By exploiting variation in regional rainfall as a natural experiment, we identify positive and significant causal peer effects in cross-region news consumption. On average a 1% increase in outside-region (or external) viewership generates approximately 0.34% additional within-region (or regional) viewership. We also find this peer effect is significantly stronger for regions with more viewership.

We then examine whether social media sharing is the primary mechanism driving these peer effects. First, we analyze the NYT's web traffic referral data and find a stronger cross-region viewership peer effect on traffic referred from social media sources (Facebook and Twitter) than for traffic referred from search engines or news aggregators (Google, Yahoo, and Bing). Second, we build the region-to-region social media follower graph, parsed from Twitter user data, and find that the peer effect between regions that are "strongly-tied" on Twitter is significantly greater than the peer effect between regions that are "weakly-tied" on Twitter. Taken together, these results imply that, on average, social media sharing does positively impact online news consumption.

The contribution of our work is twofold. First, we develop a methodology that allows content producers to estimate whether social media is helping or hurting their businesses. Second, to the best of our knowledge, ours is the first paper to causally estimate the impact of social media sharing on online content consumption. Our work highlights the increasing importance of social media to the digital strategy of news content producers. Specifically, our results suggest that interventions targeted to increase social media sharing are likely to generate positive value for content producers.

2 Literature

Our paper draws from and contributes to two diverse streams of literature. Topically, our work relates to the growing literature on the impact of the Internet on the news industry and in particular, to work on the substitutability or complementarity of social media and online news consumption. We briefly overview the broader literature in section 2.1 before diving more deeply into work on the substitutability of social media and news consumption in section 2.2. Methodologically, our

[2]To be specific, high-quality social media sharing data is not necessary for the identification of causal peer effects. However, we make use of some social media data to build an approximate region-to-region connectivity graph to determine whether social media sharing is a primary mechanism behind these peer effects.

empirical strategy is inspired by approaches used to estimate peer effects in networks. Since this literature is deep and rich, we provide a brief overview of the most relevant work in section 2.3.

2.1 The Internet and News Industries

Early work in this space was primarily concerned with whether online news substituted for print news (Deleersnyder et al. 2002, Dimmick et al. 2004, Ahlers 2006, Gentzkow 2007). Using a variety of approaches, these papers all confirmed economically significant substitution effects. In addition to finding evidence of online substitution, Gentzkow (2007) calculated that consumer surplus increases by $45 million a year from free online news using a structural model that valued the entry of new goods. Despite the overall evidence of substitution, these papers, for the most part, note that online news media seemed to have little economic impact on the news industry at the time. However, it is important to note that all of these studies were conducted prior to the collapse of newspaper advertising revenue. A more recent study by Cho et al. (2016) paints a much starker picture: they find that the Internet is driving newspapers out of business, especially small local newspapers. This finding is quite concerning given the importance of local newspapers for political participation (Gentzkow et al. 2011).

Many researchers have since explored how news organizations (or content producers more generally) might succeed in the digital age. Several studies (Lee et al. 2014, Vu 2014, Sen and Yildirim 2015) suggest that editorial decision making, at least for some news organizations, has shifted towards increasing digital viewership and making money through advertising. Berger and Milkman (2012) examine what makes some pieces of content more viral than others, increasing views and ad impressions. Combining an observational study with lab experiments, they find that emotionally arousing news articles tend to be shared more online. News organizations have also increasingly adopted subscription-first business models. Generally, this is accomplished by instituting a digital paywall[3] that limits readers to a certain number of free articles per month or restricts access to "premium" or exclusive content (this is also referred to as "content gating"). As one might expect, digital paywalls seem to reduce the total amount of news consumed (Chiou and Tucker 2013), but generate more paid subscribers and circulation (Pattabhiramaiah et al. 2017). Aral and Dhillon (2018) investigate how changing the parameters of a paywall—the number of free articles and availability of sections—affect content consumption, subscription rates, and revenue. They find that, at the NYT, constraining paywall parameters reduced viewership but increased subscriptions, creating substantial net increases in monthly revenue. Closely related is a study by Lambrecht and Misra (2017) who find that content producers that offer both free and paid content should increase the share of free content available during periods of high demand.

There is also a significant body of work that examines the effect of specific Internet-era technologies on the economics of the news industry. For instance, Seamans and Zhu (2013) found that Craigslist devastated the market for classified advertising, estimating that newspapers lost of $5.4

[3]It is worth noting that digital paywalls are not exclusive to news content. For example, Youtube and Hulu both offer premium subscription services that expand the amount of available content.

billion in revenue from 2000 to 2007. Others have looked into the impact of news aggregators—websites or applications (e.g. Google News, Bing News, Yahoo News, etc.) that use human editorial judgement and computer algorithms to collect and curate news produced by other parties. Work on this topic is typically framed in terms of determining whether news aggregators are substitutes or complements for news consumption. Dellarocas et al. (2016) find that displaying more information about articles on a news aggregator decreases the probability that readers will click through to the full article. Though this result seems to suggest substitution, the other empirical research on this topic indicates that the complementarity effect dominates overall. Calzada and Gil (2016) and Athey et al. (2017), both studying a Google News shutdown in Spain, independently find that the shutdown caused news website visits to drop. A related paper by Chiou and Tucker (2017)—which exploits a dispute between Google News and the Associated Press—also finds evidence to support complementarity. As social media takes on the role of a content aggregator, a similar area of research has emerged around the substitutability or complementarity of social media.

2.2 Complementary and Substitution of Social Media on News Consumption

There has been recent interest in understanding how social media is changing news content consumption. Various theoretical mechanisms have been proposed to explain the potential patterns of substitution or complementarity with social media. For instance, social media may be substituting for news consumption through an "information redundancy effect."[4] Articles shared on social media usually include headlines and short snippets which decrease the informational value of clicking through to the actual article. This effect can be further exacerbated if the user opts to include a personal summary of the content when they share it. Another possible cause of substitution is a "time displacement effect." Here, the central idea is that social media use competes with news consumption for a consumer's time—as time spent on one medium increases, usage of other mediums decrease (Kayany and Yelsma 2000).

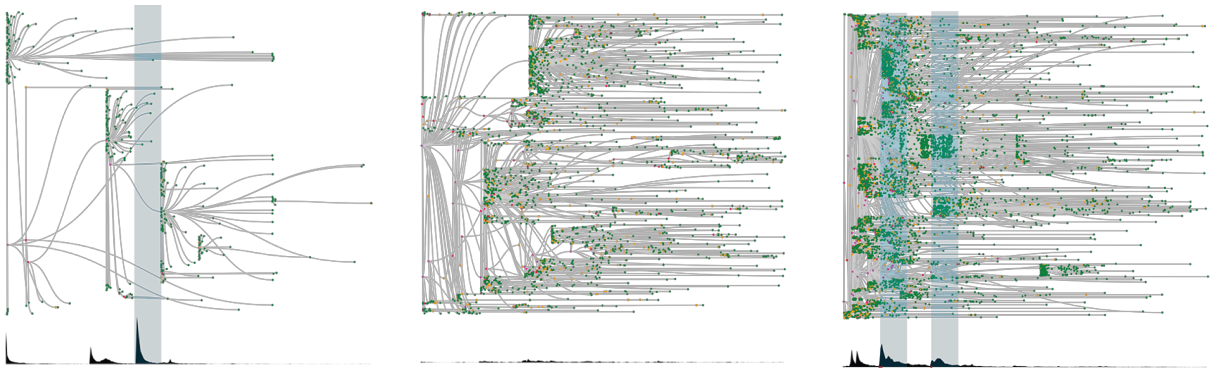
On the other hand, social media might complement news consumption through a "content discovery effect." Sharing a piece of content allows that content to reach people who wouldn't have seen it otherwise. In fact, the promotional power of social media may be particularly effective due to homophily (McPherson et al. 2001), the idea that "birds of a feather flock together." Since we tend to connect to others similar to us, content that is shared by our peers is likely to be more relevant to us. This effect may be even further enhanced due to "communication utility" (Atkin 1972) as we may want to discuss the article content with the person who shared it with us. Furthermore, reading shared content is also important for maintaining a shared information set with our peers. Alternatively, complementarity may result from a "habit formation effect." When using social media, users are routinely exposed to news content. This in turn may cause news consumption to become part of a larger ritual that includes social media use (McLeod and Chaffee 1973).

The difference between direct and indirect effects is also important. It is possible that different mechanisms may dominate at different time scales. Hence, even if substitution dominates

[4]This is also called the "scanning effect" in the news aggregator literature (Chiou and Tucker 2017)

indirectly (time displacement > habit formation), it still may be worthwhile to invest in a social media campaign if complementarity dominates directly (content discovery > information redundancy). Although broad macro trends seem to suggest that social media indirectly substitutes news consumption, empirically it is unclear which effects dominate directly. While particular pieces of content have almost certainly reaped massive benefits from social media, it is still unclear whether and to what degree these benefits generalize to news content (or online content more broadly). Anecdotal evidence in Aral (2013) seems to indicate the potential for both substitution and complementarity. He examines the Twitter cascades and web traffic of three New York Times articles, visualized in Figure 1.

Figure 1: Twitter Cascades and Click-through Web Traffic of 3 NYT Articles



(a) Article 1 (b) Article 2 (c) Article 3

Note: Visualizations by Niklaus Hanselmann. These figures illustrate the tweets and retweets (dots and lines) as well as the click-through web traffic (black density graphs below the cascades) for three NYT articles over time, visualizing the realtime association between Twitter activity and content viewership. The light blue bars in each article highlight particular periods of particularly heavy web traffic. In the first article, Twitter activity and content web traffic don't seem to correspond to each other. In the second article, there is a great deal of Twitter activity but very little web traffic. In the third article, periods of heightened Twitter activity seem to line up with periods of high web traffic.

For the first article, we can see a large spike in web traffic to the article, highlighted in the light blue region, that does not seem to coincide with any Twitter activity. In this case, it seems that social media activity and news consumption may be largely independent of each other. On the other hand, the second article enjoys a great deal of social media activity, but very little web traffic, suggesting that sharing might be substituting for content viewership. Lastly, the third article seems to display a positive feedback effect at work: periods of high Twitter activity readily correspond to periods of high web traffic (highlighted in the light blue regions), potentially the implying the dominance of complementarity effects.

While there might be significant variation in the substitution and complementarity patterns of news content at the article level, broader work suggests that the complementarity effect is stronger at the aggregate level. Early work by Hong (2012) looks at Twitter adoption and subsequent online traffic to 337 newspapers from January 2007 to December 2010. Using a monthly panel,

he found a significant positive association between Twitter adoption and the number of unique visitors to newspapers' websites even after including newspaper fixed effects and non-parametrically controlling for time trends. More recent work by Sismeiro and Mahmood (2018) examines the effect of a major Facebook outage on hourly web traffic to a major European news website. During the outage, they found a significant decrease in both traffic and unique visitors that lasted well beyond the outage itself. However, in direct contrast to this result, a recent news article (Schwartz 2018) showed that during a 45 minute Facebook outage, traffic to news websites increased overall by 2.3%, suggesting substitution was at work.[5]

Our work addresses two major gaps in the extant empirical literature. First, it provides, to the best of our knowledge, the first causal evidence of the complementarity between social media sharing and online news consumption. Second, we also focus particularly on the question of social media sharing rather than social media as a whole. This allows our results to more readily inform actionable interventions such as social sharing encouragements or viral marketing campaigns.

2.3 Peer Influence and Social Contagion

As we will describe, our approach to identifying the causal impact of social media on news consumption relies on identifying peer effects across regions that share links to news articles on social media networks. But, just as with the relationship between social media and news consumption, endogeneity issues present a challenge in identifying peer effects (Manski (1993)). It is now well known that behaviors and outcomes cluster among connected peers in networks. One explanation for this tendency is peer influence or the idea that my behavior is causing your behavior (or vice versa). However, these observed similarities may also be the result of other factors such as homophily (the tendency for like peers to form connections) ((McPherson et al. 2001), Smirnov and Thurner (2017)), confounding factors (correlated exposures to confounding events), and simultaneity (correlated outcomes resulting from peers simultaneously co-influencing each other).

Econometric identification of peer influence is a critically important question with significant real-world policy implications. Distinguishing whether observed behavioral clustering results from peer influence or other factors is critical in understanding what sorts of policy interventions are effective. For instance, peer-to-peer intervention strategies are not likely to be effective if clustered behavior is primary driven by homophily. Empirically, a number of approaches have been used to solve this identification issue including high-dimensional matching (Aral et al. 2009), structural modeling (Ghose and Han 2011), instrumental variables approaches (Bramoullé et al. 2009; Tucker 2008; Coviello et al. 2014; Aral and Nicolaides 2017), and randomized field experiments (Aral and Walker 2012; Bond et al. 2012; Banerjee et al. 2013; Kramer et al. 2014). Our approach is similar to the instrumental variables approaches used by Coviello et al. (2014) and Aral and Nicolaides (2017). These papers both exploit regional rainfall as a source of exogenous variation to study emotional and exercise contagion respectively. One difference with our empirical approach is that

[5] "What happens when Facebook goes down? People read the news" (http://www.niemanlab.org/2018/10/what-happens-when-facebook-goes-down-people-read-the-news/)

we focus on aggregate region-level outcomes rather than individual-level outcomes.

3 Data and Empirical Strategy

3.1 Data Sources

Our dataset is constructed primarily from the proprietary web activity logs of the NYT. This data consists of more than 2 billion individual events tracked on the NYT's internal servers from April 3, 2013 to October 31, 2013.[6] The raw data is very rich and granular consisting of millisecond-accurate timestamps, IP-address derived geolocation data, accessed URL, referrer URL, among many other fields. Since we are primarily concerned with content consumption, we limit ourselves to direct content pageviews. We further restrict ourselves to the date range between April 8, 2013 and October 27, 2013 so that we can fully capture each weekly news cycle (Monday to Sunday). Following this, we aggregate these content views up to the region-day level where regions are determined by clustering cities with high correlation in rainfall patterns and low geographic distance using a hierarchical clustering algorithm and restrict ourselves to the the top 500 US regions accounting for over 200 million content pageviews (70% of total content views).

We supplement this data with two additional data sources: the Global Historical Climatology Network (GHCN) database maintained by the National Oceanic and Atmospheric Association (NOAA) described in Menne et al. (2012) and Twitter user and follower data parsed using Tweepy.[7] The GHCN data contains daily observations of maximum temperature, minimum temperature and precipitation for some 45 thousand weather stations around the world (of which approximately 30 thousand are located in the continental United States). We use the geographic coordinates of each weather station in order to derive precipitation data for each region in our dataset. Our Twitter data is built by parsing a sample of 10000 "ordinary"[8] Twitter accounts with a tweet or retweet containing a link to a NYT article during our time frame.[9] As mentioned earlier, there are a number of concerns with the quality and precision of the self-reported locations in users' profiles. Hence, we are only able to obtain adequate coverage for the top 100 regions. We randomly sampled 100 accounts from each of these 100 regions and then obtained the self-reported locations (if available) in the profiles of all of the followers of those 10000 accounts to construct a region-to-region social media connectivity graph. A more comprehensive description of all of our data processing procedures is provided in Appendix A.

[6] Some prominent examples of tracked events include content pageviews, frontpage visits, searches, account settings changes, and even 404 "Page Not Found" Errors.

[7] "An easy-to-use Python library for accessing the Twitter API" (http://www.tweepy.org/).

[8] We exclude accounts with an extreme number of followers (10000+), so that we can parse a greater number of accounts more easily due to rate limits in the Twitter API.

[9] These tweets and retweets were also provided to us by the NYT.

3.2 Empirical Strategy

There are two major challenges to identifying the effect of social media sharing on news consumption: endogeneity and data collection. With endogeneity, one issue is the unobserved confounder of "inherent newsworthiness." Both news viewership and social media activity will be extremely high for major events like mass shootings, terrorist attacks, extreme weather disasters, or presidential elections. On the other hand, reporting on local scandals, book reviews, or esoteric stories about font are unlikely to be widely viewed or shared.

Additionally, there is also the econometric issue of simultaneous equation bias. Content viewership, at least in some cases, is likely driven by social media sharing.[10] But, at the same time, online news provides material for people to share and discuss. Written as a nonparametric system of equations we have:

$$V = f(S) + \epsilon$$
$$S = g(V) + \nu$$

where V denotes content viewership and S denotes social media sharing while ϵ and ν denote the error terms associated with the viewership and sharing equations respectively. By construction, V is structurally dependent on ν just as S is structurally dependent on ϵ which renders any "selection-on-observables" approach[11] insufficient, even when conditioning on the "right" covariates.

Though a classic econometric solution to these endogeneity problems is to use instrumental variables (IV), finding instruments in our context is difficult. Valid instruments need to satisfy two main restrictions: exclusion and relevancy. Unfortunately, it is generally quite difficult to satisfy the exclusion restriction since most covariates are likely to affect both V and S. One possible candidate would be something like a social media service interruption where a platform like Twitter or Facebook becomes inaccessible for some period of time. However, these platforms work extremely hard to make sure such events are exceedingly rare and are quickly resolved if they occur making such instruments less relevant. Moreover, to achieve identification with such an instrument, one would need a collection of such events. Any one-off service outage would be generically confounded by day-to-day variation in news consumption. To illustrate this point, suppose that there was a Facebook outage on the day of a major breaking news event like a terrorist attack. Without a credible means of controlling for inherent newsworthiness, a one-off instrument[12] would be strongly associated with increased news viewership since its effectively functioning as a fixed effect that captures the variation from the positive shock.

To overcome this challenge, we conduct our study at the level of localized geographic regions, something that we call "regional localization." This region-level analysis enables us to use "local"

[10] In some cases, social media activity itself can be newsworthy. For instance, President Trump's tweets routinely get a significant amount of news coverage. Viral phenomenon also commonly shows up in the news, e.g. ALS Ice Bucket Challenge, Laurel vs Yanny, the Dress, etc.

[11] Conditioning on some covariate set or using some kind of matching strategy

[12] Where "1" denotes an outage.

instruments, variables that shock particular regions but not others. Events such as local sports games, state-specific holidays (Patriot's Day, Nevada Day, Confederate Memorial Day, etc.), or music concerts could all potentially serve as a local instrument. In our case, we use rainfall. While it may not be obvious that the weather has an effect on news viewership and sharing, the basic idea is that when it is raining outside, people are more likely to stay indoors and spend time on the Internet reading the news and browsing social media. For example, Sen and Yildirim (2015)—who use weather as an instrument to study whether editorial decisions about news content take viewership into consideration—find that viewership of articles is 5%-8% higher on rainy days for a large Indian news provider.

Naturally, our approach relies on high-quality geolocation data to produce sensible results. However, this is where we run into the second major challenge: data collection. For our strategy to work optimally, we would need geolocated social media data on sharing events in addition to the network structure. In our case, the best we can obtain are Twitter users' self-reported profile locations which suffer from many problems associated with self-reported survey data. In particular, non-response is a significant issue, with around half of tweets or retweets lacking self-reported location information. Even for users with self-reported locations, there are still serious concerns about the accuracy and granularity of the locations provided.[13]

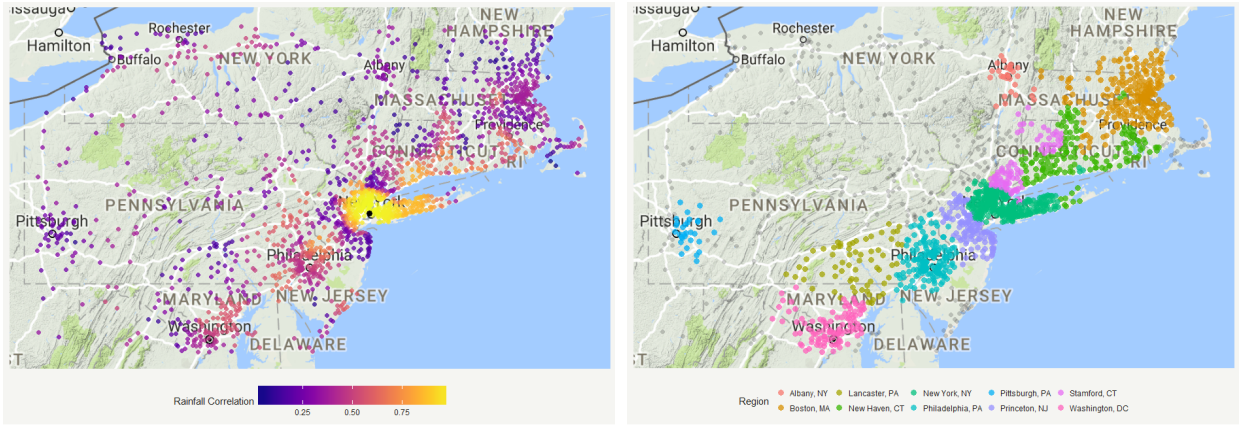
To address this, we reformulate our strategy to identify causal "peer effects" in regional viewership. What we mean by this is we shift our objective to identify the effect of viewership of other regions on a particular region, which obviates the need for high-quality geolocated social media data. To clarify our approach, consider the simple 2-region case:

$$V_{1t} = \alpha + \beta V_{2t} + \gamma R_{1t} + \epsilon_{1t}$$
$$V_{2t} = \alpha + \beta V_{1t} + \gamma R_{2t} + \epsilon_{2t}$$

In the first equation, viewership in region 1 on date t (V_{1t}) depends on viewership in region 2 on date t (V_{2t}) and rainfall in region 1 on date t (R_{1t}). Similarly, in the second equation, viewership in region 2 on date t depends on viewership in region 1 on date t and rainfall in region 2 on date t (R_{2t}). Not only is there a simultaneity problem, the error terms, ϵ_{1t} and ϵ_{2t} are also almost certainly correlated as a result of variation in day-to-day "inherent newsworthiness" that impacts both regions. But, as long as R_{1t} and R_{2t} are independent of ϵ_{2t} and ϵ_{1t} respectively, we can use R_{2t} as an instrument for V_{2t} in the first equation and R_{1t} as an instrument for V_{1t} in the second equation to identify β.

There are two potential concerns with our rainfall instrument. First, there may be correlation in

[13] A user's self-reported location may not match their actual location at the time of the tweet due to travel or a failure to update location after moving. There is also often a tendency to self-report the closest major city since it is much more recognizable to other people. Users can also intentionally mis-report their location if she strongly identifies with a particular place like her hometown rather than her current location. Other fairly popular self-reported locations are metaphysical in nature, e.g. "Every/Some/No-where", "My Happy Place", "Somewhere over the Rainbow", "On Cloud 9", etc. Aside from these issues, there are potential problems with granularity. While some self-reported locations are very specific—"Williamsburg", "San Francisco, CA", etc.—others are far too general for our purposes—"California", "United States", "Earth", etc.



(a) Rainfall Correlation with NYC (b) Top 10 Northeastern Regions

Northeast Cities These figures depict 1692 cities in the Northeast United States included in NYT data. (a) illustrates the Pearson correlation in rainfall of the these cities with New York City (displayed as the black point). Highly correlated cities are displayed as orange or yellow while more uncorrelated cities purple or blue. As we might expect, the surrounding cities have extremely similar rainfall to NYC. (b) shows the membership of top 10 regions on this map, with the remainder of cities as gray dots. These regions correspond quite well to correlation map in (a).

rainfall events from region to region which would violate the exclusion restriction. Our hierarchical clustering procedure partially addresses this issue since cities with similar weather will likely end up in the same region, especially if they are geographically close. Looking at Figure ??, and focusing on New York City, we see that cities with highly correlated rainfall are assigned to the same region as NYC. However, there are still potential problems that result from correlated rainfall across regions. For example, consider the rainfall correlation between NYC and Washington DC. Though it is not quite as high compared with NYC's surrounding cities, it is certainly still high enough for us to be concerned about a potential exclusion restriction violation. In order to address this concern, we restrict our analysis to regions with sufficiently uncorrelated rainfall. Second, our exclusion restriction may be violated due to national news coverage about weather events. To address this issue, we remove articles with weather-related content tags[14] from our analysis. Overall however, we are not particularly worried about this concern since weather-related content represents such a small fraction of NYT articles and pageviews[15] meaning that potential bias will be extremely small.

3.3 Regression Specifications

The standard model specification used in peer effects studies is the "linear-in-means" model where the outcomes of an ego unit are regressed on the average outcomes of that unit's peers (this is also known as an "outside-in" model). While Manski (1993) establishes that peer effects are generically not identified, later work by Brock and Durlauf (2001) and Durlauf and Tanaka (2008) show

[14] According to the NYT's internal content tagging system, accounts for 0.2% of pageviews

[15] Especially since 2013 was a rather mild year in terms of weather related disasters. According to NOAA (2018) only 8 major weather-related disasters occurred during our timeframe.

that instrumental variables are sufficient for identification. We make one minor adaptation to the standard linear-in-means model to better suit our context: rather than using the mean outcome of alters as our main independent variable, we use the summed outcome of alters.

3.3.1 Main Specification

Our main model specification is a log-log peer effects panel model:

$$\ln V_{it} = \beta \ln V_{-it} + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + alpha_i + \tau_t + \epsilon_{it} \tag{1}$$

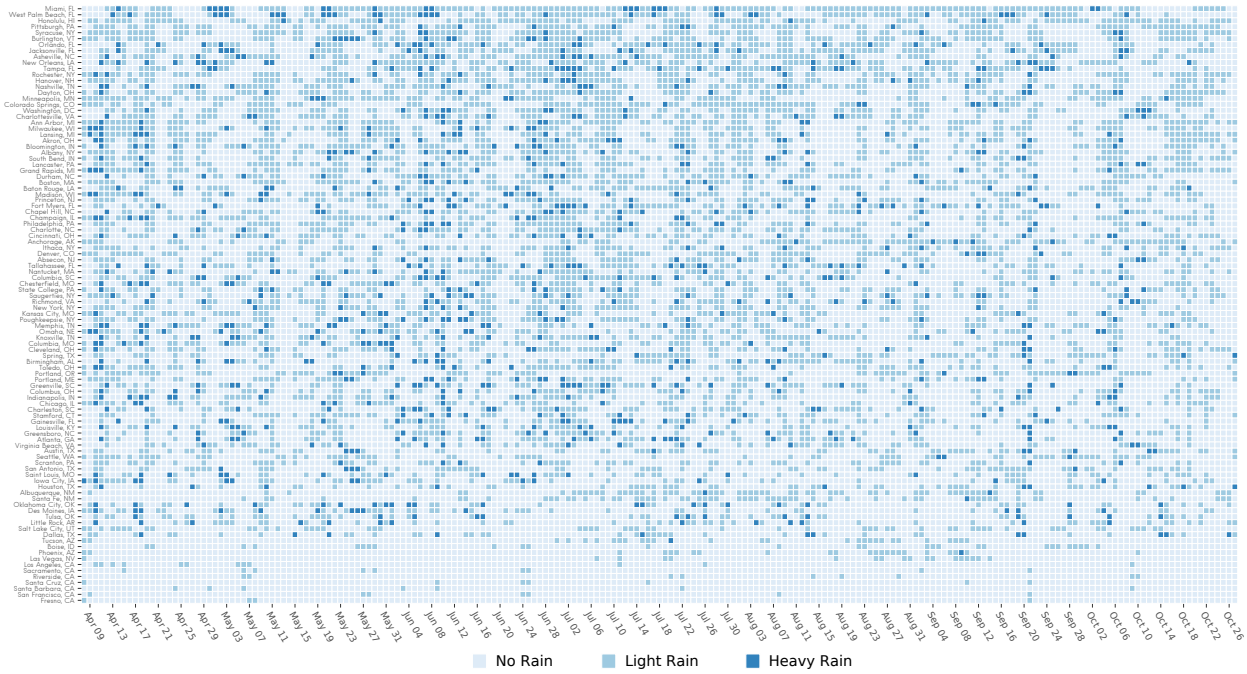
The dependent variable in this model is $\ln V_{it}$, the log views of NYT content in region i on date t, which we will refer to as "regional" viewership for expositional reasons. The main independent variable of interest is $\ln V_{-it}$, the log of the viewership of NYT content from all other regions with sufficiently uncorrelated precipitation (P) with region i on date t. Again, for ease of exposition, we will refer to this as "external viewership." More formally, we can write $V_{-it} = \sum_{j \in \boldsymbol{U}_i} V_{jt}$ where $\boldsymbol{U}_i$ is the set of all regions k where the absolute correlation in precipitation between k and i is less than 0.25 ($\boldsymbol{U}_i = \{k : |\rho(P_k, P_i)| < .25\}$). The associated parameter, β, denotes the peer effect of aggregated viewership in regions in $\boldsymbol{U}_i$ on the viewership in region i. Since we're using a log-log model specification, we interpret β as an elasticity—a 1% increase in V_{-it} generates β more viewership in V_{it}.

We operationalize regional weather by converting our continuous precipitation measure into two binary variables R^l_{it} and R^h_{it}, indicating when precipitation exceeds 0.22mm and 16.86mm respectively. We will refer to precipitation less than or equal to 0.22mm as "no rain," precipitation exceeding 0.22mm but less than or equal to 16.86mm as "light rain," and precipitation exceeding 16.86mm as "heavy rain." These values were determined using a 2-stage greedy grid search aimed to find the discretization thresholds that have the greatest explanatory power on regional viewership. The exact details of this procedure can be found in Appendix A.5. We opt for discrete indicators rather than the continuous measure since human behavior regarding the weather tends to be non-linear. For instance, the amount of outdoor activity that occurs sharply drops even if there is just a little bit or rain outside. Given the way R^l_{it} and R^h_{it} are coded, γ_1 captures the effect of light rain on regional viewership but γ_2 captures the additional marginal effect of heavy rain. An illustration of our discretized rain measure for the top 100 regions in our data can be found in Figure **??**.

As part of our panel specification, we include a set of region (α_i) and time fixed effects (τ_t[16]). In our context, region fixed effects control for regional heterogeneity arising from differences in factors such as population demographics, political leaning, etc.[17] The time-fixed effects are perhaps even more important, accounting for at least some of the variation in unobserved seasonal trends and

[16]For this study, we operationalize our time fixed here as a combination of week and day-of-week fixed effects. Using date-level fixed effects introduces a peculiar type of substitution endogeneity. People, generally speaking, can only view content from one place. If someone reads an NYT article in NYC, it simultaneously implies that she is not reading that article outside of NYC.

[17]While population characteristics are certainly changing over the course of our study, the amount of change that can occur over the 7 month period of our study are going to be quite minor



Rainfall for Top 100 Regions This plot illustrates the precipitation for the top 100 regions between April 3, 2013 to October 31, 2013. Each cell in this plot corresponds to the amount of precipitation a particular region (on the y-axis) received on on a particular day (on the x-axis). Precipitation is discretized into 3 categories: "No Rain" (less than or equal to 0.22mm of precipitation), "Light Rain" (greater than 0.22mm and less than or equal to 16.86mm of precipitation), and "Heavy Rain" (greater than 16.86mm of precipitation). The rainiest regions are found at the top of this plot (Miami, Fl; West Palm Beach, Fl; Honolulu, HI) while the driest regions are at the bottom (regions in California due to drought in 2013).

even inherent newsworthiness. Lastly, ϵ_{it} denotes the error term.

3.3.2 IV First-Stage

We use the following model specification for the first-stage of our instrumental variables regression:

$$\ln V_{-it} = \zeta_1 R^l_{-it} + \zeta_2 R^h_{-it} + \eta_1 R^l_{it} + \eta_2 R^h_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \tag{2}$$

The instruments for $\ln V_{-it}$ are R^l_{-it} and R^h_{-it}, the simple averages of light and heavy rainfall for regions with uncorrelated weather to i on date t respectively. More formally:

$$R^l_{-it} = \frac{1}{|\boldsymbol{U}_i|} \sum_{j \in \boldsymbol{U}_i} R_{1jt}$$
$$R^h_{-it} = \frac{1}{|\boldsymbol{U}_i|} \sum_{j \in \boldsymbol{U}_i} R_{2jt}$$

where $|\boldsymbol{U}_i|$ denotes the cardinality or number of regions in $\boldsymbol{U}_i$. Since R^l_{-it} is the fraction of regions experiencing light rainfall, we can interpret its associated parameter ζ_1 as the (approximate) percent in viewership if no regions in $\boldsymbol{U}_i$ are raining vs. if all regions in $\boldsymbol{U}_i$ are lightly raining. Accordingly, R^h_{-it} is the fraction of regions experiencing heavy rainfall, and given its coding, its associated parameter ζ_2 can be interpreted as the (approximate) percent change in viewership when comparing light rain in all regions in $\boldsymbol{U}_i$ relative to heavy rain in those same regions. R^l_{it} and R^h_{it} are included as control covariates since they show up in the main specification above. If our exclusion restriction assumption is correct, then we should expect the associated parameters η_1 and η_2 to be essentially 0.

Naturally, we also include both sets of fixed effects in this specification as well. While the time fixed effects τ_t behave similarly as they do in equation 1 above, the region fixed effects α_{-i} work a little bit differently. In this case, they absorb the aggregated population demographics, political leaning, etc. of all regions $j \in \boldsymbol{U}_i$. Programmatically however, this is equivalent to including a fixed effect for region i since $\boldsymbol{U}_i$ is determined by i. Finally, ν_{-it} is the associated error for our first stage specification.

3.3.3 Regional Heterogeneity

As we mentioned earlier, we expect there to be a great deal of heterogeneity in the strength of the viewership peer effect from region to region, especially considering our log-log model specification. We use the following model specification to investigate this regional heterogeneity:

$$\ln V_{it} = \beta^{top}(\ln V_{-it})^{top} + \beta^{mid}(\ln V_{-it})^{mid} + \beta^{bot}(\ln V_{-it})^{bot} + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{3}$$

Here, we modify our main model specification by breaking up our original main dependent variable into 3 distinct parts: $(\ln V_{-it})^{top}$, $(\ln V_{-it})^{mid}$, and $(\ln V_{-it})^{bot}$. These three terms simply represent

the interaction of $\ln V_{-it}$ with top_i, mid_i, and bot_i—binary variables indicating if a region i belongs to the top 167, middle 166, or bottom 167 regions in the regional viewership distribution in our timeframe. Here, we can interpret β^{top}, β^{mid}, and β^{bot} as the average cross-region viewership effect on the top, middle, and bottom regions respectively. Given the way these variables are coded, the parameter estimates are interpreted as the conditional average effects for the top, middle, and bottom regions respectively, rather than as an additive marginal effect assuming one group as the baseline group.

3.3.4 Viewership Referrals

To investigate the mechanism driving these peer effects, we examine the cross region viewership effect across two modes of content distribution: social media sharing and search. We specifically examine viewership that is referred from social media sources (urls from Facebook or Twitter) relative to viewership referred from search engines or news aggregators (urls from Google, Yahoo, or Bing). We use the following 2 regressions:

$$\ln V_{it}^{sm} = \beta^{sm} \ln V_{-it} + \gamma_1^{sm} R_{it}^{l} + \gamma_2^{sm} R_{it}^{h} + \alpha_i^{sm} + \tau_t^{sm} + \epsilon_{it}^{sm} \tag{4}$$

$$\ln V_{it}^{se} = \beta^{se} \ln V_{-it} + \gamma_1^{se} R_{it}^{l} + \gamma_2^{se} R_{it}^{h} + \alpha_i^{se} + \tau_t^{se} + \epsilon_{it}^{se} \tag{5}$$

These regressions simply substitute different dependent variables into our main model specification. $\ln V_{it}^{sm}$ denotes the log sum viewership with social media referrers (Facebook and Twitter) of region i on date t. On the other hand, $\ln V_{it}^{se}$ denotes he log sum viewership referred from search engines or news aggregators (Google, Yahoo, and Bing) of region i on date i. β^{sm} and β^{se} are the peer effect parameters that we are primarily interested in. α_i^{sm}, α_i^{se}, τ_t^{sm} and τ_t^{se} denote region and time fixed effects, while ϵ_{it}^{sm} and ϵ_{it}^{se} capture the error terms . If social media sharing is the dominant mechanism driving these peer effects, then we should expect the β^{sm} to exceed β^{se} .

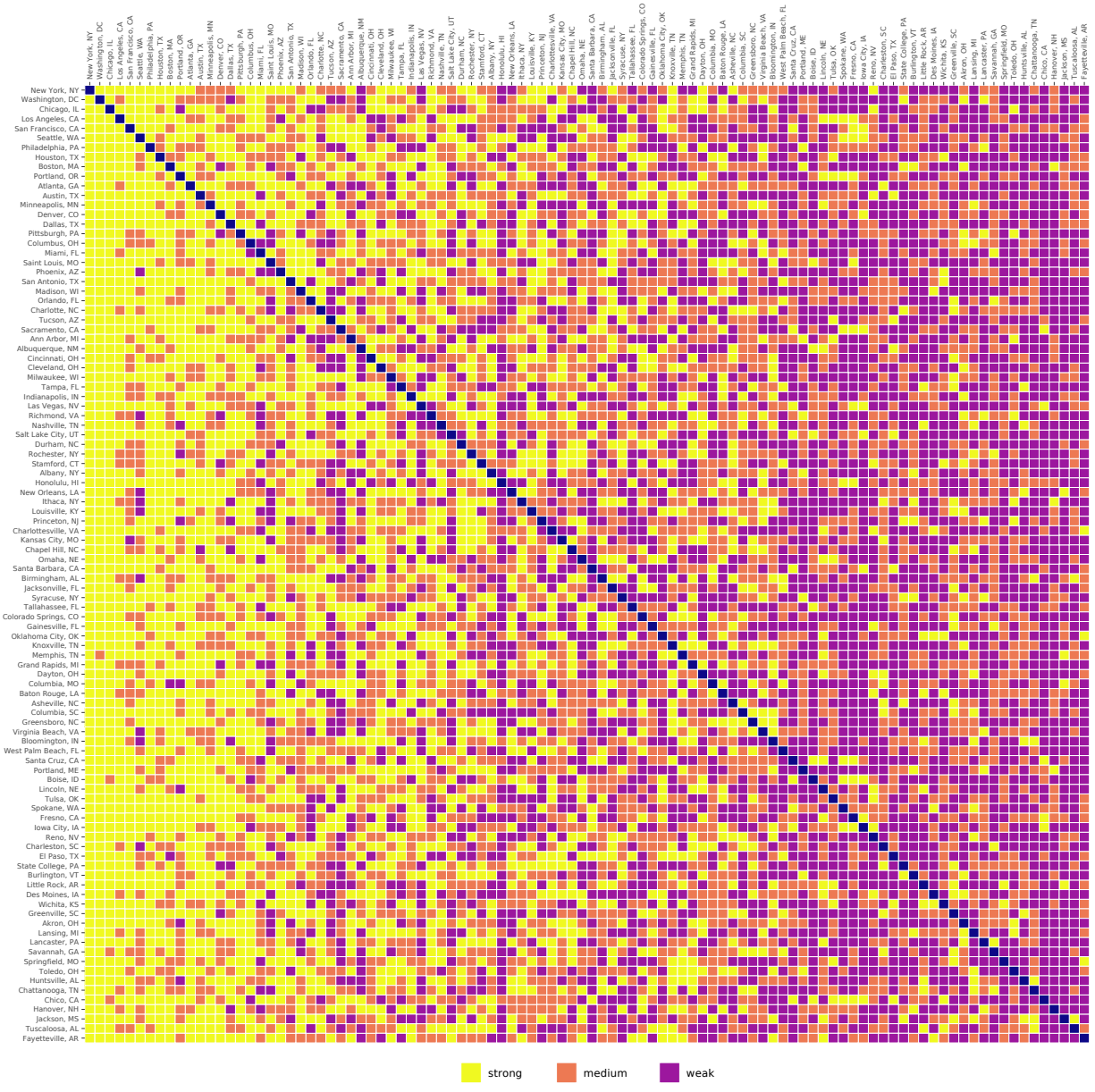
3.3.5 Twitter Connectivity

To further explore whether social media sharing is really driving these peer effects, we examine how social network connectivity mediates the strength of the peer effect. Using our Twitter follower data, we generate the region-to-region followship graph of the top 100 regions. For each region i, we classify each of other 99 j regions as "strongly-tied," "moderately-tied," and "weakly-tied" regions based on tie density. Strongly-tied regions are the top third of regions with the highest number of follower-followee[18] ties, moderately-tied regions are the middle third, and weakly-tied are the lowest third. We plot this followship graph as an adjacency matrix in Figure **??**.

For this part of our work, we use the following model specification:

$$\ln V_i t = \beta^{st} \ln V_{-it}^{st} + \beta^{mt} \ln V_{-it}^{mt} + \beta^{wt} \ln V_{-it}^{wt} + \gamma_1 R_{it}^{l} + \gamma_2 R_{it}^{h} + alpha_i + \tau_t + \epsilon_{it} \tag{6}$$

[18]The follower is the user in i who follows a user in j.



Region-to-Region Adjacency Matrix This a plot of directed adjacency matrix indicating region-to-region Twitter followee-follower tie density. Let i index the vertical axis and j index the horizontal axis. Excluding the cells where $i = j$ (along the dark blue diagonal), each cell represents whether users in region i follow many users in region j. Yellow ("strong"), orange ("moderate"), and purple ("weak") indicate whether the number of users followed in region j is in the top, middle, or bottom tertiles for region i. Regions are ordered from top-bottom and from left-right in terms of descending total viewership. Looking along the horizontal axis, we see that the leftmost, high population regions, are regions that tend to be more followed. As we move along the horizontal axis to smaller and smaller regions, we see a greater and greater amount of purple cells.

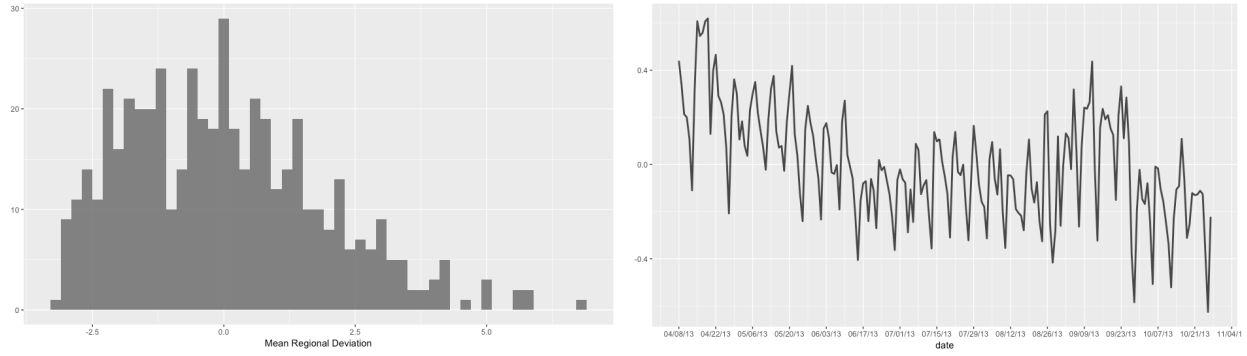
This specification modifies our main specification by replacing the primary independent variable $\ln V_{-it}$ with three new independent variables: $\ln V_{-it}^{st}$, $\ln V_{-it}^{mt}$, and $\ln V_{-it}^{wt}$. $\ln V_{-it}^{st}$ is denotes the log sum viewership of strongly-tied regions with sufficiently uncorrelated weather to i. $\ln V_{-it}^{mt}$ and $\ln V_{-it}^{wt}$ are the same, except for moderately- and weakly-tied regions. Naturally, α_i, τ_t, and ϵ_{it} are still as they were before. If social media sharing does play a major role in driving the cross-region viewership peer effect, then we should expect $\beta^{st} > \beta^{mt} > \beta^{wt}$.

4 Results

We begin this section with some basic descriptive statistics and visualizations of our data. We report the demeaned values for our viewership variables for data privacy reasons in Table 1 below. Our final data set consists of 101,500 observations (500 regions $\times$ 203 dates) and accounts for over 200 million pageviews. This full data set is used to estimate equations 1, 3.3.2, 3.3.3, and 3.3.4. However, as mentioned earlier, we only use 20,300 observations (100 regions $\times$ 203 dates) to estimate equation 3.3.5, due to the difficulty in joining self-reported Twitter profile locations to our constructed regions.

We also include a histogram of the mean regional deviation in log viewership and a plot of day-to-day shocks in regional viewership (see Figure 3) to illustrate the variation in $\log V_{it}$ from region to region and from day to day.

Figure 3: Variation in Log Viewership



(a) Regional Variation

(b) Date-to-Date Variation

Note: These are demeaned values.

4.1 Peer Effects in Online News Viewership

This section is organized as follows. We begin by validating our instruments and the first-stage regression of our IV model in section 4.1.1. We then move onto the estimation and analysis of our main regression 1 in section **??**. Lastly, we explore the heterogeneity of the peer effect relative to regional size in section 4.1.4.

Table 1: Description of Variables used in Regressions

Variable	Description	Mean	St. dev.	Min	Max
Main Specification Variables (3.3.1)					
$\ln V_{it}$	Demeaned log views of NYT content in region i on date t	0.000	1.905	−5.896	7.35
$\ln V_{-it}$	Demeaned log aggregate views of NYT content in regions with uncorrelated rainfall with i on date t	0.000	0.265	−1.222	0.698
R^l_{it}	Indicator for light rainfall (precipitation exceeding 0.22mm) in region i on date t	0.355	0.479	0	1
R^h_{it}	Indicator for heavy rainfall (precipitation exceeding 16.86mm) in region i on date t	0.052	0.222	0	1
First Stage Variables (3.3.2)					
R^l_{-it}	Weighted average of light rainfall indicators in regions $\boldsymbol{U}_i$ on date t	0.354	0.093	0.084	0.577
R^h_{-it}	Weighted average of heavy rainfall indicators in regions $\boldsymbol{U}_i$ on date t	0.052	0.033	0.000	0.206
Regional Heterogeneity Specification Variables (3.3.3)					
top_i	Indicator for whether region i is one the top 167 regions in total viewership	0.334	0.472	0	1
mid_i	Indicator for whether region i is one the middle 166 regions in total viewership	0.332	0.471	0	1
bot_i	Indicator for whether region i is one the bottom 167 regions in total viewership	0.052	0.472	0	1
Search vs Social Variables (3.3.4)					
$\ln V^{sm}_{it}$	Demeaned log aggregate views of NYT content referred from Facebook and Twitter in regions i on date t	0.000	1.882	−3.330	7.246
$\ln V^{se}_{it}$	Demeaned log aggregate views of NYT content referred from Google, Yahoo, and Bing region i on date t	0.000	1.829	−4.321	7.062
Twitter Connectivity Variables (3.3.5)					
$\ln V^{st}_{-it}$	Demeaned log aggregate views of NYT content in "strongly-connected" regions to i with uncorrelated rainfall on date t	0.000	0.484	−1.745	1.563
$\ln V^{mt}_{-it}$	Demeaned log aggregate views of NYT content in "moderately-connected" regions to i with uncorrelated rainfall on date t	0.000	0.466	−1.662	1.342
$\ln V^{wt}_{-it}$	Demeaned log aggregate views of NYT content in "weakly-connected" regions to i with uncorrelated rainfall on date t	0.000	0.663	−2.012	1.539

4.1.1 Instrument Validity and First-Stage Estimates

Even before examining the first stage itself, we first look at the coefficient estimates on R^l_{it} and R^h_{it} in Table 3. Looking specifically at columns 3 (TWFE) and 4 (IV), we can see that the models that incorporate both region and time fixed effects produce highly statistically significant light rainfall coefficients of around 0.02. Hence, if it is lightly raining in a particular region, then that region's viewership increases by approximately 2%. As for heavy rainfall, both TWFE and IV again produce similar statistically significant coefficients also coincidentally around 0.02. As mentioned earlier, due to the coding of R^l_{it} and R^h_{it}, the coefficient on R^h_{it} is interpreted as an additional marginal effect. Hence, if it is heavily raining in region i, then we should expect region i's viewership to increase by around 4%. Looking at column 2, not including time-fixed effects seems to slightly depress the coefficient estimates of both rainfall variables, but they still remain highly statistically significant. Most importantly however, these results confirm the basic intuition of instrumental variables approach since they indicate that rainfall has statistically and practically significant effects on a region's online news consumption. Next, we directly check the validity of the aggregated weather instruments themselves R^l_{-it} and R^h_{-it}. Table 2 presents the estimation of the first stage (eq. 2).

4.1.2 Instrument Validity and First-Stage Estimates

Even before examining the first stage itself, we fist look at the coefficient estimates on R^l_{it} and R^h_{it} in table 3. Looking specifically at columns 3 (TWFE) and 4 (IV), we can see that the models that incorporate both region and time fixed effects produce highly statistically significant light rainfall coefficients of around 0.02. Hence, if it lightly raining in a particular region, then we should expect that region's viewership to increase by approximately 2%. As for heavy rainfall, both TWFE and IV again produce similar statistically significant coefficients also coincidentally around 0.02. As mentioned earlier, due to the coding of R^l_{it} and R^h_{it}, the coefficient on R^h_{it} is interpreted as an additional marginal effect. Hence, if it is heavily raining in region i, then we should expect region i's viewership to increase by around 4%. Looking at column 2, not including time-fixed effects seems to slightly depress the coefficient estimates of both rainfall variables, but they still remain highly statistically significant. Most importantly however, these results confirm the basic intuition of instrumental variables approach since they indicate that rainfall has statistically and practically significant effects on a region's online news consumption. Next, we directly check the validity of the aggregated weather instruments themselves R^l_{-it} and R^h_{-it}. Table 2 presents the estimation of the first stage (eq. 2).

Looking at this table, we see that the estimated coefficients on R^l_{-it} and R^h_{-it} are both positive and highly statistically significant. Taking the parameter estimates as given however, we should observe that aggregated viewership is approximately 6.5% higher if it were lightly raining in all regions in $\boldsymbol{U}_i$ compared to absolutely no rain in the same regions. Again the coefficient on R^h_{-it} should be an additive marginal effect. Thus, these estimates suggest that when it is raining heavily in all regions in $\boldsymbol{U}_i$ then aggregated viewership should be approximately 34.8% higher compared

Table 2: First Stage

	Dependent variable:
	$\ln V_{-it}$
R^l_{-it}	0.0627***
	(0.0009)
R^h_{-it}	0.2353***
	(0.0021)
R^l_{it}	0.0006
	(0.0007)
R^h_{it}	0.0021
	(0.0013)
Wald F-stat	16033
Observations	101,500
R^2	0.9001

Note: *p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

with no rain.

While these numbers seem extreme, especially compared to the region-level rainfall effects, this in part is likely the result of the fact that there's never actually a case where it is either completely raining or completely dry in every single region. Related to this is the fact that R^l_{-it} and R^h_{-it} are aggregated measures based off the average across many regions. This means that the variance of these covariates will naturally be much smaller than the single-region counterparts R^l_{it} and R^h_{it}, which is likely to generate more extreme coefficient estimates[19]. Overall, we think the point values of these coefficients shouldn't be taken too literally.

More important is the question on whether R^l_{-it} and R^h_{-it} are meaningfully shocking our main covariate of interest $\ln V_{-it}$. To check the validity of our instruments, we conduct a Wald-Test and find an F-stat of 16033 (p-value = 0), substantially surpassing Stock and Yogo (2002)'s threshold for strong instruments. We therefore conclude that we do not have a weak instruments problem and that our instruments satisfy the relevancy condition. Since we have two instruments for a single endogenous variable, we also conduct a Sargan over-identification test and obtain a test-statistic of 22.047 (p-value = 1), suggesting that our instruments are indeed exogenous.

Given our peer effects setting, it is also important to check the coefficient estimates on the

[19]Recall that a regression coefficient in simple linear regression is given by $\frac{\text{Cov}(Y,X)}{\text{Var}(X)}$. In our case, the variance of R^l_{-it} is 26 times greater than R^l_{it} while the variance of R^h_{-it} is 46 times greater than R^h_{it}.

regional rainfall covariates R^l_{it} and R^h_{it} in the first stage. Significance here would suggest that there is some unobserved confounder that relates regional rainfall to external viewership. By extension, that same confounder would therefore likely relate aggregated external rainfall to regional viewership, indicating a violation of the exclusion restriction. Looking to the results, the coefficients on both R^l_{it} and R^h_{it} are neither statistically nor practically different from 0, further suggesting that our exclusion restriction holds.

4.1.3 Peer Effects Estimates

Table 3 presents the estimation of equation 1 using 4 different approaches: pooled-OLS (OLS) in column 1, region fixed effects (RFE) in column 2, region and time fixed effects or two-way fixed effects (TWFE) in column 3, and lastly, instrument variables (IV) in column 4.

Table 3: Main Results: Cross Region Peer Effect

	Dependent variable:			
	$\ln V_{it}$			
	(1) OLS	(2) RFE	(3) TWFE	(4) IV
$\ln V_{-it}$	−0.210	0.931***	1.050***	0.343***
	(0.200)	(0.011)	(0.010)	(0.077)
R^l_{it}	0.253***	0.017***	0.020***	0.021***
	(0.053)	(0.003)	(0.003)	(0.003)
R^h_{it}	0.006	0.014***	0.018***	0.019***
	(0.052)	(0.004)	(0.004)	(0.004)
Region FE	N	Y	Y	Y
Time FE	N	N	Y	Y
Observations	101,500	101,500	101,500	101,500
R^2	0.005	0.981	0.981	0.980

Note: *p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

If we think that variation in the "newsworthiness" of day-to-day events is the main driver of news consumption, then given our log-log specification, we might expect non-IV methods to produce a biased peer effect estimate of approximately 1[20]. Looking at the pooled OLS result in column 1 of table 3, we see that this is not case. Instead, we find a statistically insignificant coefficient estimate of -0.210. This negative point estimate is largely the result of the generic negative correlation

[20]Suppose that a terrorist attack occurs tomorrow and that doubles news consumption throughout the entire US. In this case, both external viewership and regional viewership would increase by 100% leading to a peer effect estimate of 1

between $\ln V_{it}$ and $\ln V_{-it}$[21]. Once we include region fixed-effects, the estimated peer effect jumps quite dramatically to a highly statistically significant 0.931 as seen in column 2, much closer to our a priori expectation of 1. Once adding on time fixed effects, which theoretically might be able to address some of the variation in day-to-day newsworthiness, the estimate slightly increases to a highly statistically significant 1.050 as seen in column 3.

These non-IV results suggest that panel methods are indeed insufficient to address the simultaneous equation bias of the peer effects framework. This leaves our IV estimates, which theoretically should be able to identity the cross-region peer effect. IV produces an estimate of 0.343 seen in column 4. Taking this point estimate as given, this suggests that a 1% increase in the total viewership of external regions (with uncorrelated weather) should increase regional viewership by approximately 0.34%. This result almost certainly confirms that regional viewership does indeed have positive spillover effects, generating additional viewership across regions. While the direction of this result agrees with the panel models, the magnitude of the coefficient is substantially different, confirming our concerns about potential endogeneity bias.

4.1.4 Regional Heterogeneity

We next look into how the strength of this peer effect might be heterogenous across regions. Table 4 presents the estimated results of equation 3 using TWFE in column (1) and IV in column (2).

Beginning with column (1), we should immediately be wary of endogeneity concerns since the coefficient estimates on $(\ln V_{-it})^{top}$, $(\ln V_{-it})^{mid}$, and $(\ln V_{-it})^{bot}$ are all rather close to 1. However, despite this problem, we see a very clear trend where the effect is strongest for the top regions and gets progressively weaker as we continue down through the middle and bottom regions. Comparing the point estimates against each other, the difference between the top and middle regions (0.040) is significant at the 10% level while the difference between the top and bottom regions is (0.052) is significant at the conventional 5% level. While the point estimates themselves are not meaningful, the observed directional trend is potentially credible. Even if the TWFE estimates are biased, it seems fairly implausible that this bias would run in different directions for the top, middle, and bottom regions.

The IV results in column (2) support this result. The point estimate on the top regions is the highest, followed by the estimate on middle regions, followed lastly by the estimate on the bottom regions. However, the overall trend here is much less clear since the standard errors of the IV model are naturally going to be much higher. Comparing the IV coefficients, we find that the difference between the top and middle regions (0.141) is significant at the conventional levels, but the difference between the top and bottom regions (0.150) is significant only at the 10% level due to the higher standard error on the bottom regions' coefficient.

Although not perfect, these results overall seem to suggest that cross-region viewership has a

[21] If region i is New York City, the highest viewership region in the data, then this automatically implies that NYC is not contributing to $\ln V_{-it}$. In the same vein, if i is a very minor region, then the high viewership regions will most likely be counted in $\ln V_{-it}$

Table 4: Regional Heterogeneity

	Dependent variable:	
	$\ln V_{it}$	
	(1) TWFE	(2) IV
$(\ln V_{-it})^{top}$	1.080^{***}	0.444^{***}
	(0.014)	(0.083)
$(\ln V_{-it})^{mid}$	1.041^{***}	0.303^{***}
	(0.019)	(0.085)
$(\ln V_{-it})^{bot}$	1.029^{***}	0.294^{**}
	(0.020)	(0.101)
R_{it}^{l}	0.020^{***}	0.021^{***}
	(0.003)	(0.003)
R_{it}^{h}	0.017^{***}	0.019^{***}
	(0.004)	(0.004)
Observations	101,500	101,500
R^2	0.981	0.980

Note: *p<0.05; **p<0.01; ***p<0.001.
Cluster-robust standard errors are reported.

stronger effect on higher viewership regions. Political homophily might be a factor that explains these results. Since the NYT is generally considered a "center-left" publication, the spillover effect from viewership to smaller regions that tend to be more politically conservative is likely to be smaller. Another possible factor is that the lower viewership regions tend to be demographically older and less likely to obtain news content from the internet compared with the high viewership regions, which would also work to suppress the magnitude of the estimated peer effect.

4.2 Is Social Media Sharing Driving these Peer Effects?

Our results point to positive and significant cross-region peer effects in news content consumption. However, it remains an open question whether social media is responsible for these peer effects. We investigate this first by comparing the estimated peer effect on viewership referred from social media sources to viewership referred from search engines and news aggregators. We also examine whether the region-to-region Twitter connectivity has a mediating effect on the strength of the peer effect.

4.2.1 Viewership Referrals: Social vs Search

Table 5 presents the estimation of the equations found in section 3.3.4. Specifically, Columns (1) and (2) report the the estimation of equation 4 using TWFE and IV. Similarly, columns (3) and (4) the estimation of equation 5, again using IV and TWFE respectively.

Table 5: WOM vs Search

	Dependent variable:			
	$\ln V_{it}^{sm}$		$\ln V_{it}^{se}$	
	(1) TWFE	(2) IV	(3) TWFE	(4) IV
$\ln V_{-it}$	1.337***	0.763***	1.107***	−0.055
	(0.023)	(0.151)	(0.009)	(0.071)
R_{it}^{l}	0.009**	0.012**	0.023***	0.023***
	(0.004)	(0.004)	(0.003)	(0.003)
R_{it}^{h}	0.028***	0.030***	0.012***	0.016***
	(0.007)	(0.007)	(0.003)	(0.004)
Observations	101,500	101,500	101,500	101,500
R^2	0.932	0.931	0.982	0.979

Note: *p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

We note that inferences made with the exact point estimates here should be treated with some caution. The referrer data we work with has some important limitations. One major problem is measurement error simply due to the limits on our ability to accurately map referrer URLs to the "channels" we're examining. Additionally, there are widely-known attribution problems with referrer data since referrer fields are often erroneously missing (copy-pasting or typing out the link manually) or unable to capture the true referrer (in-person WOM). Despite these issues, some clear trends emerge from these results. Both TWFE and IV models show substantially stronger estimated coefficients on content referred from social media than content referred from search engines. In particular, the IV results are quite striking. Looking at column 4, we see that the estimated effect of $\ln V_{-it}$ on $\ln V_{it}^{se}$ is -.055. Though this point estimate itself is negative, it is not statistically significantly different from 0. In retrospect, this result may not be too surprising. Although rainfall does have a positive and significant on viewership, it is still rather unlikely for a small boost in viewership in a number of regions, even if they are the large regions, to dramatically change search or news aggregator ranking results by a significant amount. On the other hand, the estimated effect of $\ln V_{-it}$ on $\ln V_{it}^{sm}$ is positive and highly statistically significant at 0.763, meaning that the cross-region effect on content referred from social media is quite strong. We feel these results overall provide powerful corroborating evidence that social media sharing is responsible for

the cross region peer effects.

4.2.2 Social Network Connectivity and Peer Effect Strength

Table 6 presents the results of estimating Equation 6. We continue to report results using both TWFE and IV in columns (1) and (2) respectively.

Table 6: How Social Network Connectivity Mediates Peer Effect Strength

	Dependent variable:	
	$\ln V_{it}$	
	(1) TWFE	(2) IV
$\ln V_{-it}^{st}$	0.996^{***}	0.669^{***}
	(0.025)	(0.066)
$\ln V_{-it}^{mt}$	0.981^{***}	0.652^{***}
	(0.032)	(0.096)
$\ln V_{-it}^{wt}$	0.973^{***}	0.403^{**}
	(0.035)	(0.098)
R_{it}^{l}	0.020^{***}	0.021^{***}
	(0.003)	(0.003)
R_{it}^{h}	0.021^{***}	0.024^{***}
	(0.006)	(0.005)
Observations	20,300	20,300
R^2	0.986	0.985

Note: *p<0.05; **p<0.01; ***p<0.001.
Cluster-robust standard errors are reported.

Looking at the TWFE estimates of $\ln V_{-it}^{st}$, $\ln V_{-it}^{mt}$, and $\ln V_{-it}^{wt}$ in column (1) we see that the estimated coefficient increases as Twitter connectivity increases. However, we unfortunately do not have enough power to determine whether these coefficients are statistically different from each other. Part of the issue here is that endogeneity bias is pushing all the coefficients towards 1, so variation between the coefficients is difficult to establish. Looking towards column (2) and the IV estimates. We see a qualitatively similar trend that the magnitude of the coefficient increases as tie density increases. In the case of IV, the difference of the effect of strongly- and weakly-tied regions (0.266) is statistically significant at the 1% level while the difference between moderately- and weakly-tied regions (0.249) is significant at the 5% level. However, we do not have enough power to determine whether the effect of strongly- and moderately-tied regions are significantly different from each other. Overall however, it seems fair to conclude that region-to-region tie density does

moderate the strength of the estimated peer effect, further supporting the idea that social media sharing is driving the peer effects.

4.3 Robustness Checks

We perform a series of checks to verify the robustness of our main results. We begin by testing 4 different operationalizations of our rainfall instrument: linear precipitation, quadratic precipitation, linear precipitation with a light rain indicator, and quadratic precipitation with a light rain indicator. We find that all these various operationalizations of rainfall produce extremely strong first-stage F-statistics. More importantly, IV estimates produced by these different operationalizes are qualitatively and quantitatively similar to original estimate seen in column (4) of table 3. The exact model specifications and regression results of these alternative operationalizations can all be found in Appendix B.1.

Next, we perform a falsification or placebo test on our weather instrument to check for instrument validity. We accomplish this by randomizing the order of our rainfall instruments since the weather from a random day should not be affecting viewership on a particular date t. We repeat this process 1000 times to generate a distribution of placebo results. Once randomized, not once out of the 1000 randomizations did the rainfall instruments exceed the strong instrument criteria determined by Stock and Yogo (2002). Accordingly, the IV estimates produced by these randomized rainfall instruments were only statistically significant at the 5% level in 10 out of 1000 trials. Moreover, none of the placebo estimates ever even approach the level of significance of real rainfall instruments. Given these results, it is quite unlikely that our estimated results in table 1 are driven by luck. Histograms of the first-stage Wald F-statistic and the p-values of the IV estimate can be found in Appendix B.2.

As mentioned in section 3.3, we adapted the standard linear-in-means model to better suit our setting. As a robustness check, we estimate the cross-region peer effect using the traditional formulation of the linear-in-means model using both TWFE and IV. These estimates are qualitatively similar to the results we presented earlier in the paper. Here, we continue to find evidence of endogeneity bias in the TWFE estimates since they substantially differ from the IV estimates. Similarly, the IV estimates are positive and highly statistically significant implying positive spillover in viewership. Quantitatively however, the point estimates themselves are significantly smaller than those estimated using equation 1. This is likely driven by the large degree of regional heterogeneity and the functional form assumptions of the linear-in-means model. The cross-region spillover effect of NYC's viewership increasing by 10% creates a much more substantial spillover effect than those generated by smaller regions. But, due to the assumptions of the specification, the contribution of NYC is averaged with the contributions of many more smaller regions, driving the estimated coefficient lower since the regional viewership distribution is very heavy tailed. The exact specification and regression results can be found in Appendix B.3.

While the linear-in-means "outside-in" model is certainly the most common framework used in the peer effect and social contagion literature, some papers, such as Coviello et al. (2014) opt for an

"inside-out" approach. Drawing on these types of models as an inspiration, we test a specification of our model that flips the dependent and main independent variable in equation 1. In this model, we are thus identifying the effect of region i's viewership on the total viewership of regions in $\boldsymbol{U}_i$. We estimate this specification using both TWFE and IV and find that the results are in line with our earlier findings. The exact model specification and results can be found in Appendix B.4.

We further investigate if time-series interactions are potentially biasing our results. For instance, ongoing stories may be causing autocorrelation in viewership. Hence, our estimated peer effect may be driven by this autocorrelation rather than being a "true" causal peer effect. To address these concerns, we estimated a set of model specifications that included auto-regressive terms of the main dependent variable. We found that including these autoregressive terms produced cross-region peer effect estimates qualitatively and quantitatively similar to our results above. All the AR model specifications and regression results can be found in Appendix B.5.

We also recheck all of our main results using different cutoff thresholds for our hierarchical clustering algorithm. Due to the nature of algorithm, a lower threshold will lead to a greater number of smaller clusters, where similarity within each cluster is higher. We find that the choice of threshold does not significantly impact our results qualitatively or quantitatively. These results can be found in in Appendix B.6.

Finally, due to the growing number of papers that employ weather as an instrument, one may wonder whether weather is truly a source of "exogenous variation." Since rainfall is causing other behavioral changes (like the reduced running found in Aral and Nicolaides (2017)), the underlying mechanism of the viewership peer effect might be these other behavioral changes rather than social media sharing. However, our results regarding the much stronger peer effect on social media traffic relative to search engine traffic implicitly addresses this concern. Browsing social media and using search engines are quite similar activities since they both involve using some kind of device to connect to the Internet to access content. If the peer effect is driven by an activity substitution mechanism, we wouldn't expect such a large difference between the estimated effects on social traffic and search traffic.

5 Discussion

Despite the seeming importance of social media to the visibility and consumption of online content, not much is actually known about how much or even whether social media can create value for online content producers. Causal identification has generally eluded the extant literature due to two main issues. First, the positive feedback nature of the relationship between social media and online content creates severe endogeneity issues. While instrumental variables can still provide identification under these circumstances, valid instruments are very difficult to find in this context. Second, the data requirements to answer this question are quite substantial. Ideally, we would want access to both browsing data from a content producer and detailed activity data from a social media platform.

To overcome these challenges, we developed a new empirical strategy that exploits regional variation in weather as an exogenous shock to regional viewership. With this model we identify causal cross-region peer effects in viewership. We find that a 1% increase in regional viewership generates about 0.34% additional viewership in outside regions. Additional analysis provide strong evidence that social media sharing is the primary mechanism driving these peer effects. Our results are quite robust to multiple operationalizations of our rainfall instruments, alternate model specifications, and different region-construction assumptions.

5.1 Managerial Implications

Our work has broad implications for news organizations and content producers more generally. First and foremost, our results are causal. Although prior work has established that social media is positively correlated with news consumption, the lack of causal identification has hindered businesses' ability to make calculated interventions. There are many plausible explanations for a generic positive association between social media use and online news consumption. Our results, however, strongly suggest that social media is having a significant positive causal effect on news viewership. As such, interventions aimed at increasing social media sharing may prove profitable. For instance, content producers can adopt sharing encouragement interventions by more actively asking their consumers to share their content or even potentially offering incentives to share. Another possible avenue could be to launch social media marketing campaigns or referral bonuses designed maximize the diffusion of content through social media.

Our work also shows that the effectiveness of using social media to drive viewership can vary significantly from region to region. The results from our regional heterogeneity specification suggest that campaigns which increase social media sharing are going to see much greater returns in the more populous regions.[22] By extension, it may be more prudent to rely on other methods like search or digital display advertising to drive content consumption in less populated regions.

Lastly, our empirical strategy can be employed by both researchers and firms alike in other contexts. As long as reliable geolocation data is available, our empirical strategy allows researchers to look for local sources of exogenous variation to use as instruments. Moreover, such instruments need not be limited to the weather. For example, local holidays or sporting events may potentially be viable candidates. For content producers, one major benefit of our approach is that it can provide causal estimates only requiring geolocated data from the content producer itself.

5.2 Limitations and Extensions

While our research provides significant insights into understanding the relationship between social media sharing and online news consumption, there are some limitations to our work. First, there are many dynamic substitution and complementarity effects that we do not consider. For instance, certain types of news content may be highly substitutable, if multiple articles all provide similar

[22] There already is a stronger peer effect on a percent-to-percent basis, even before taking into the account the extreme difference in population.

coverage about a single event. In this case, reading one article about the event might mean that it's less likely to read other articles about the same event. Alternatively, other types of news might be more complementary like reviews or opinion pieces which generally have a wider variety of viewpoints. Moreover, there may also be similarly complex interactions with radio and cable news as well. One way to try and understand some of these interactions may be to look at online news (or content more generally) at the content level, rather than region-day aggregates. Another limitation is that our work only focuses on "short-term" effects. Understanding the longer-term complementarity or competitive dynamics of social media and news consumption is beyond the scope of this paper.

Naturally there are also concerns about generalizability. The NYT is a rather exceptional news organization with more Pulitzer Prizes than any other newspaper. Its readership and influence outstrip most other news organizations and it is considered as the national "newspaper of record." Hence it is unclear to what degree our results will generalize to smaller or less influential news organizations. Similarly, it is also unclear whether we would find similar results for other types of content. News content has several distinctive properties. First, news content has notable alternative channels for consumption. While print circulation has been falling over the last decade, it still remains a major channel for the consumption of news content. Second, the relevance of news content tends to fade much more quickly than other types of content. For example, people still regularly listen to and enjoy music by the Beatles, whereas a news article from 1960's will have very little consumption value. While these reasons may suggest that social media's effect will be stronger for other types of online content (that lack such alternative channels), more research needs to be done.

There is still much to be understood about the role of different social media platforms or different types of word-of-mouth. Facebook, Twitter, Reddit, etc. all make different platform design choices. It would be interesting to understand what those platform-level differences imply about online content and business strategies designed to maximize viewership. Overall, there is still much that needs to be understood about the relationship between social media and online content.

5.3 Conclusion

Publishers strategies in the digital age depend critically on whether social media substitutes for or drives online news consumption. Unfortunately, a number of empirical challenges, including endogeneity and the confounding effect of inherent newsworthiness, have made identifying the causal effect of social media on news consumption difficult. We overcome these challenges by leveraging exogenous variation in regional news consumption to identify the cross region viewership effects of news sharing on social media. We identify positive and significant cross-region peer effects in online news consumption and demonstrate that social media sharing is the primary driver of these cross-region peer effects. Specifically, cross-region peer effects in viewership are stronger when referred by social media than when referred by search engines. Furthermore, social network connectivity mediates the strength of these peer effects. Regions that are strongly-connected on social media

exhibit more positive and significant peer effects than regions that are weakly-connected on social media. We also find these effects are more pronounced for more populous regions. Taken together, these results suggest that social media sharing drives online news consumption. This conclusion and the methods used to estimate it can help content producers make better decisions about how and when to incentivize social media use to drive viewership, revenues, and profit.

References

Ahlers, Douglas. 2006. News consumption and the new electronic media. *Harvard International Journal of Press/Politics* **11**(1) 29–52. doi:10.1177/1081180X05284317. URL https://doi.org/10.1177/1081180X05284317.

Aral, Sinan. 2013. To go from big data to big insight, start with a visual. *Harvard Business Review* URL https://hbr.org/2013/08/visualizing-how-online-word-of.

Aral, Sinan, Paramveer Dhillon. 2018. Digital paywall design: Implications for content demand and subscription rates .

Aral, Sinan, Lev Muchnik, Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51) 21544–21549.

Aral, Sinan, Christos Nicolaides. 2017. Exercise contagion in a global social network. *Nature Communications* **8**(14753). doi:10.1038/ncomms14753.

Aral, Sinan, Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* **337**(6092) 337–341. doi:10.1126/science.1215842. URL http://science.sciencemag.org/content/337/6092/337.

Athey, Susan, Markus M. Mobius, Jeno Pal. 2017. The impact of aggregators on internet news consumption URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2897960.

Atkin, Charles K. 1972. Anticipated communication and mass media information-seeking. *Public Opinion Quarterly* **36**(2) 188–199.

Banerjee, Abhijit, Arun G. Chandrasekhar, Esthet Duflo, Matthew O. Jackson. 2013. The diffusion of microfinance. *Science* **341**(6144). doi:10.1126/science.1236498. URL http://science.sciencemag.org/content/341/6144/1236498.

Barthel, Michale. 2017. Newspapers fact sheet. Tech. rep., Pew Research Center. URL http://www.journalism.org/fact-sheet/newspapers/.

Berger, Jonah, Katherine L. Milkman. 2012. What makes online content viral? *Journal of Marketing Research* **49**(2) 192–205. doi:10.1509/jmr.10.0353. URL https://doi.org/10.1509/jmr.10.0353.

Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, James H. Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**(7415). doi:10.1038/nature11421. URL http://doi.org/10.1038/nature11421.

Bramoullé, Yann, Habiba Djebbari, Bernard Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics* **150**(1) 41–55.

Brock, William A., Steven N. Durlauf. 2001. Interactions-based models. Elsevier, 3297 – 3380. doi:https://doi.org/10.1016/S1573-4412(01)05007-3. URL http://www.sciencedirect.com/science/article/pii/S1573441201050073.

Calzada, Joan, Ricard Gil. 2016. What do news aggregators do? evidence from google news in spain and germany URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2837553.

Chiou, Lesley, Catherine Tucker. 2013. Paywalls and the demand for news. *Information Economics and Policy* **25**(2) 61–69. doi:https://doi.org/10.1016/j.infoeconpol.2013.03.001. URL http://www.sciencedirect.com/science/article.pii/S016.

Chiou, Lesley, Catherine Tucker. 2017. Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy* **26**(4) 782–805. doi:10.1111/jems.12207. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jems.12207.

Cho, Daegon, Michael D. Smith, Alejandro Zentner. 2016. Internet adoption and the survival of print newspapers: A country-level examination. *Information Economics and Policy* **37** 3–19. doi:https://doi.org/10.1016/j.infoecopol.2016.10.001. URL http://www.sciencedirect.com/science/article/pii/S0167624516301196.

Coviello, Lorenzo, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, James H. Fowler. 2014. Detecting emotional contagion in massive social networks. *PloS one* **9**(3) e90315.

Deleersnyder, Barbara, Inge Geyskens, Katrijn Gielens, Marnik G. Dekimpe. 2002. How cannibalistic is the internet channel? a study of the newspaper industry in the united kingdom and the netherlands. *International Journal of Research in Marketing* **19**(4) 337–348. doi:https://doi.org/10.1016/S0167-8116(02)00099-X. URL http://www.sciencedirect.com/science/article/pii/S016781160200099X.

Dellarocas, Chrysanthos, Juliana Sutanto, Mihai Calin, Elia Palme. 2016. Attention allocation in information-rich environments: The case of news aggregators. *Management Science* **62**(9) 2543–2562. doi:10.1287/mnsc.2015.2237. URL https://doi.org/10.1287/mnsc.2015.2237.

Dimmick, John, Yan Chen, Zhan Li. 2004. Competition between the internet and traditional news media: The gratification-opportunities niche dimension. *The Journal of Media Economics* **17**(1) 19–33.

Durlauf, Steven N., Hisatoshi Tanaka. 2008. Understanding regression versus variance tests for social interactions. *Economic Inquiry* **46**(1) 25–28. doi:10.1111/j.1465-7295.2007.00076.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1465-7295.2007.00076.x.

Gentzkow, Matthew. 2007. Valuing new goods in a model with complementarity: Online newspapers. *American Economic Review* **97**(3) 713–744. doi:10.1257/aer.97.3.713. URL http://www.aeaweb.org/articles?id=10.1257/aer.97.3.713.

Gentzkow, Matthew, Jesse M. Shapiro, Michael Sinkinson. 2011. The effect of newspaper entry and exit on electoral politics. *American Economic Review* **101**(7) 2980–3018. doi:10.1257/aer.101.7.2980. URL http://www.aeaweb.org/articles?id=10.1257/aer.101.7.2980.

Ghose, Anindya, Sang Pil Han. 2011. An empirical analysis of user content generation and usage behavior on the mobile internet. *Management Science* **57**(9) 1671–1691.

GlobalWebIndex. 2017. Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes). Tech. rep., Statista. URL https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/.

Hausman, Jerry A., William E. Taylor. 1983. Identification in linear simultaneous equation models with covariance restrictions: An instrumental variables interpretation **51**(5) 152–1549.

Hong, Sounman. 2012. Online news on twitter: Newspapers social media adoption and their online readership. *Information Economics and Policy* **24**(1) 69–74.

Kayany, Joseph M., Paul Yelsma. 2000. Displacement effects of online media in the socio-technical contexts of households. *Journal of Broadcasting & Electronic Media* **44**(2) 215–229. doi:10.1207/s15506878jobem4402_4. URL https://doi.org/10.1207/s15506878jobem4402_4.

Kramer, Adam D. I., Jamie E. Guillory, Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* **111**(24) 8788–8790. doi:10.1073/pnas.1320040111. URL http://www.pnas.org/content/111/24/8788.abstract.

Lambrecht, Anja, Kanishka Misra. 2017. Fee or free: When should firms charge for online content? *Management Science* **63**(4) 1150–1165. doi:10.1287/mnsc.2015.2383. URL https://doi.org/10.1287/mnsc.2015.2383.

Lee, Angela M., Seth C. Lewis, Matthew Powers. 2014. Audience clicks and news placement: A study of time-lagged influence in online journalism. *Communication Research* **41**(4) 505–530. doi:10.1177/0093650212467031. URL https://doi.org/10.1177/0093650212467031.

Luca, Michael. 2015. User-generated content and social media. *Handbook of Media Economics*, vol. 1. Elsevier, 563–592.

Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* **60**(3) 531–542.

McKinsey. 2016. Global media report 2016. Tech. rep., McKinsey & Company. URL https://www.mckinsey.com/industries/media-and-entertainment/our-insights/global-media-report-2016.

McLeod, Jack M., Steven H. Chaffee. 1973. Interpersonal approaches to communication research. *American Behavioral Scientist* **16**(4) 469–499. doi:10.1177/000276427301600402. URL https://doi.org/10.1177/000276427301600402.

McPherson, Miller, Lynn Smith-Lovin, James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* **27**(1) 415–444.

Menne, Matthew J., Imke Durre, Russell S. Vose, Byron E. Gleason, Tamara G. Houston. 2012. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology* **29**(7) 897–910.

Newman, Nic, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, Rasmus K. Nielsen. 2017. Digital news report 2017. Tech. rep., Reuters Institute. URL http://www.digitalnewsreport.org/.

NOAA. 2018. Noaa national centers for environmental information (ncei) u.s. billion-dollar weather and climate disasters URL https://www.ncdc.noaa.gov/billions/.

Pattabhiramaiah, Adithya, S. Sriram, Puneet Manchanda. 2017. Paywalls: Monetizing online content .

Schwartz, Josh. 2018. What happens when facebook goes down? people read the news URL http://www.niemanlab.org/2018/10/what-happens-when-facebook-goes-down-people-read-the-news/.

Seamans, Robert, Feng Zhu. 2013. Responses to entry in multi-sided markets: The impact of craigslist on local newspapers. *Management Science* **60**(2) 476–493.

Sen, Anaya, Pinar Yildirim. 2015. Clicks bias in editorial decisions: How does popularity shape online news coverage? doi:10.2139/ssrn.2619440. URL http://dx.doi.org/10.2139/ssrn.2619440.

Sismeiro, Catarina, Ammara Mahmood. 2018. Competitive vs. complementary effects in online social networks and news consumption: A natural experiment. *Management Science* .

Smirnov, Ivan, Stefan Thurner. 2017. Formation of homophily in academic performance: Students change their friends rather than performance. *PloS One* **12**(8) e0183473.

Stock, James H, Motohiro Yogo. 2002. Testing for weak instruments in linear iv regression.

Stocking, Galen. 2017. Digital news fact sheet. Tech. rep., Pew Research Center. URL http://www.journalism.org/fact-sheet/digital-news/.

Tucker, Catherine. 2008. Identifying formal and informal influence in technology adoption with network externalities. *Management Science* **54**(12) 2024–2038. doi:10.1287/mnsc.1080.0897. URL https://doi.org/10.1287/mnsc.1080.0897.

Vu, Hong Tien. 2014. The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism* **15**(8) 1094–1110. doi:10.1177/1464884913504259. URL https://doi.org/10.1177/1464884913504259.

A Data Processing Procedures

A.1 NYT and GHCN Data Processing

We are very careful in handling the NYT data since it potentially contains fairly sensitive information. When parsing the dataset, we are careful to avoid any personal identifying information. The only fields that we look at are time of access, the url accessed, type of content, the derived geolocation, and the referrer URL. We are primary interested in consumption of actual content, hence we exclude any events that aren't associated with a piece of actual content. Since the NYT tracks the approximate duration readers stay on each webpage, a single pageview often results in multiple events in the web data. To account for this, we only look at the initial access of a piece of NYT content.

We aggregate these events to the city-date level. Since the NYT is based in the United States, we limit ourselves to viewership that occurs in the United States representing 72% of total pageviews. These pageviews are quite unevenly spread across 26166 cities (Boroughs are considered separate geolocations, for example, Brooklyn, NY and Queens, NY are separate from New York, NY) leading to a rather long-tailed distribution. We further exclude cities that do not have at least one pageview in each day across our dataset, leaving us with 5381 cities accounting for 97% of US pageviews. For each of the remaining cities, we use the Google Maps API to obtain geographic coordinate data.

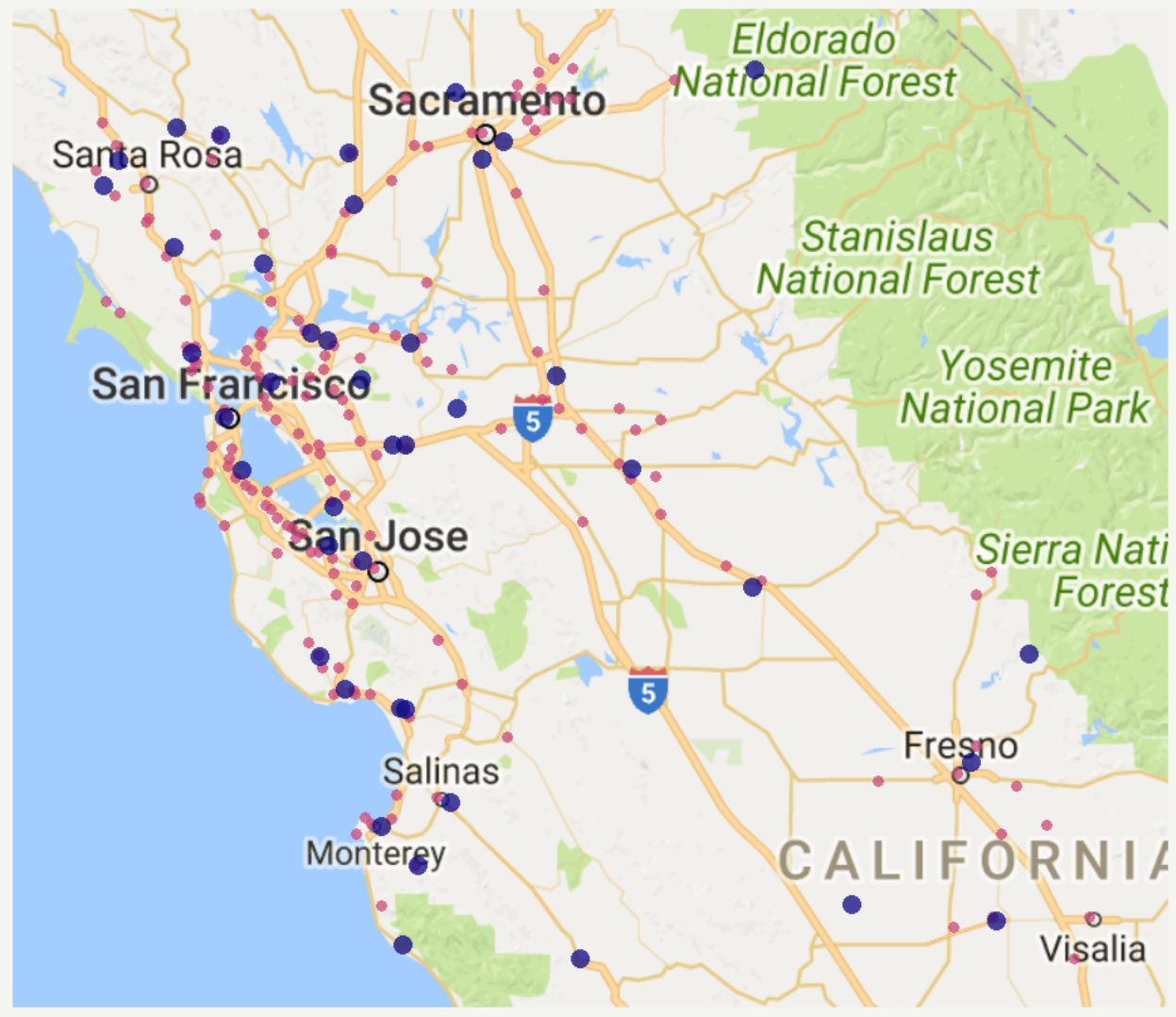
As we mentioned above, the GHCN data contains daily observations of maximum temperature, minimum temperature, precipitation, and geographic location for some 45 thousand weather stations around the world. since we are focusing on the US, we restrict ourselves to US-based weather stations. However, there is significant variation the number of reports from each weather station. We therefore limit ourselves to only the weather stations that aren't missing any precipitation observations for each one of the days in our time period, leaving us with 2,852 remaining weather stations.

A.2 Determining the Weather of Cities

In order for our identification strategy to work, we need to determine the amount of rainfall in each location in our dataset. This turns out to be relatively complicated. For some cases, there's only a single choice for a nearby weather station. However, for many others, there a number nearby weather stations that we can use as potential rainfall measurements. For example, looking at Northern California in figure 4, we can observe both of these cases. Near Fresno, weather stations are rather spread out with a fair amount of distance from station to station. However, in the SF Bay Area, weather station density (as well as city density) is considerable higher, with some weather stations practically next to one another.

Hence, to obtain weather data for each city in our dataset, we take the weighted average of the precipitation measurements from all weather stations located within 20 miles of each city. Our weighting scheme is based on proximity where closer stations have higher weights than further ones.

Figure 4: Northern California Cities and Weather Stations



This figure plots the locations of cities in the NYT dataset and weather stations in the GHCN dataset, subject to the filtering rules we describe above. Cities are denoted by the smaller purple points while weather stations are denoted by the larger dark blue ones.

Specifically, our weights are given by:

$$W_{ij} = \frac{20 - D_{ij}}{\sum_{k \in B_{20}(i)} 20 - D_{ik}}$$

W_{ij} denotes the weight of weather station j's precipitation measurement used to construct city i's rainfall measure as long as j is within 20 miles of i. D_{ij} is simply the geographic distance between city i and weather station j. We subtract this distance from 20 and divide by a normalizing constant to make sure the weights of all weather stations within in 20 miles sum up to 1. Although nearby weather stations have extremely correlated weather, this procedure allows to be less sensitive from potentially non-representative readings from a single weather station as there are some cases where a single weather station will have an outlier measurement relative to the nearby stations. A small number of cities did not have a weather station within 20 miles and were dropped from our data, leaving us with 4933 remaining locations.

A.3 Hierarchical Clustering

Since our identification strategy depends on rainfall and nearby locations generally have highly correlated rainfall, we wanted to create generate larger clusters of locations that all share highly similar weather. To create these clusters, we relied on an unsupervised machine learning algorithm known as hierarchical clustering. The hierarchical clustering procedure is very simple. First it initializes by recognizing each observation as its own individual cluster of one. The algorithm then proceeds to iteratively join the most similar clusters until a single cluster containing all the observations is reached. Hierarchical clustering is probably best known for generating the biological tree of life and the assignment of life into the various taxonomic classifications of domain, kingdom, phylum, class, order, family, genus, and species.

There are two main decisions to make when it comes to hierarchical clustering: the dissimilarity (also called the distance) metric, used to determine how "close" two individual observations are, and the linkage function, used to determine dissimilarity between multi-observation clusters. We construct our dissimilarity metric based on two factors, absolute correlation coefficient in rainfall and geographic distance. We combine these two factors together in the following manner:

$$D_{ij} = |\rho_{ij}| + G_{ij}/500$$

Here, the dissimilarity between two cities i and j is equal to the absolute Pearson correlation in rainfall between i and j ($|\rho_{ij}|$) plus the geographic distance between i and j (G_{ij}) divided by 500. Our choice of 500 here is fairly arbitrary, we simply wanted a distance so that cities with highly correlated rainfall that happened to be far away from one another were not placed into the same cluster. One way to think about our distance metric is that 500 miles of geographic distance contributes as much to "dissimilarity" as does going from completely uncorrelated weather to perfectly correlated weather.

Figure 5: Included Regions in the Continental United States

This figure plots the locations of the top 500 regions in our dataset (excluding Honolulu, HI and Anchorage, AL). Regions are centered on the highest viewership city in each region. The size of points reflect the relative viewership totals of these regions with higher viewership regions illustrated as larger points.

Using this dissimilarity metric, we ran the hierarchical clustering algorithm using several options for the linkage function. For each linkage function, we generated 1000 clusters and then compared how similar the clusters to each other. While there were minor differences in clusters from one linkage function to the next, for the most part the generated clusters were quite similar. Since our choice of linkage function didn't seem to matter all too much, we ended selecting simply using the average dissimilarity for the linkage function. While although we could've simply used the 1000 clusters we already generated when testing the different linkage functions, we opted instead to use clusters determined by a dissimilarity threshold. In particular, we chose a dissimilarity threshold of .5 leading to 542 clusters. For clusters with multiple observations, the average within-cluster rainfall correlation is 0.85. Our choice of .5 for the cutoff threshold here is again, somewhat arbitrary. We simply wanted a threshold help minimize some of the problems of self-reported Twitter locations yet maintained a high level of within-cluster rainfall correlation. We aggregate pageviews up to the region-day level where regions are defined the 542 clusters. The majority of the results presented in paper are estimated using the top 500 of the 542 regions. We present a plot of these regions in the continental United States, centered on the "main city" (city with the highest viewership in the cluster) in figure 5.

This clustering procedure helps mitigate the problem of Twitter users' reporting the nearest

major city rather than their actual locations since both are likely to be in same cluster. This also helps allows us to use accept slightly less granular self-reported locations than we would've before (for example, larger metropolitan areas like "NY Metropolitan Area" or "SF Bay Area").

A.4 Twitter Data Processing

Along with the weblog data, NYT provided us with a dataset containing every single tweet and retweet containing a bitly shortened URL to a piece of NYT content for approximately the same range as our time period. We used this to obtain a list of users with self-reported location data. Due to the problems working with self-reported locations, it was not possible for us to reliable attribute specific tweets to the locations we included in our dataset. However, the overall city-to-city tie density network should be a little less sensitive to issues mentioned above under a couple of reasonable assumptions: that the tie-density network is fairly stable over our time period and the identifiable self-reported locations tie-density network is a good representation of the "true" network.

Our main challenge here was accurately mapping self-reported locations to actual locations. Many different self-reported locations map to the same location in our dataset. For example, "NYC", "New York City", "Big Apple", "Midtown", "Wall Street", "NYU" etc. should all map to the same location for our purposes. We first lowercased all the self-reported city strings and filtered our converted all non-ASCII text. We then used some basic regular expressions to further normalize the location text. After this, we ran the cleaned text through the Google Maps API to recover geographic coordinates and location "type". Google Maps has several different location classifications, for example, a "sublocality" generally refers official sub-city areas like the borough of Brooklyn, a "locality" generally denotes a city or town, an "administrative_area_level_2" indicates a county, and an "administrative_area_level_1" designates a state (in the United States). There are also some unofficial types of importance, namely, "colloquial_area" which refers to the area of land that might make up a large metropolitan area like the SF Bay Area, the Tri-State Area, or Greater Los Angeles. We keep users with self-reported locations that have Google Maps types of "neighborhood","sublocality", "locality" and "colloquial_area". We specifically discard self-reported locations that has the Google Maps type of "route", even though "routes" are even more granular than localities. The problem with routes is the vast majority of "routes" are actually self-reported locations like "Cloud 9"—which the Google Maps API will return a result for "Cloud 9 Inn"—or "Nowhere"—which may be some local bar. Moreover, people generally don't self-report their own location with that degree of specificity (for the most part, the highest degree of specificity of commonly reported is at the level of "Williamsburg, Brooklyn" or "Midtown Manhattan" which both count as neighborhoods).

Using the geographic coordinates, we determine the closest city in the NYT dataset to each of the self-reported user locations. This way we can get region assignments for each of the included self-reported locations. In total, these self-reported locations are assigned to 174 of our regions. We then check the the ratio between between a region's total viewership and tweets and retweets

and exclude regions with ratios in top and bottom 5% of the ratio distribution. We further restrict our analysis to the top 100 remaining regions. Since it is not feasible (due to Twitter API limits) for us to examine the followers of every single account, we devise the following sampling procedure: we first exclude accounts with follower counts in the top 5% and accounts with fewer than 50 followers. Since the number of accounts associated with each region is very long tailed, we randomly sample 100 accounts from each of the 100 regions to make sure smaller regions are more accurately represented.

Using Tweepy, we access the Twitter API to obtain the followers (over 200000) of these 10000 users. Again making use of the Tweepy, we obtain the self-reported locations of the followers. Naturally, the follower self-reported locations have all the problems we described above. Hence, we use the same procedure as above to determine which region these follower accounts belong too. We then use this information to build the region-to-region directed network. Each directed edge e_{ij} in this network represents the the number of accounts in j that follow accounts in city i. Naturally, to account for the stratified sampling approach, we multiply the number of follower links appropriately (for example, if region i has 1000 accounts, we would then multiply e_{ij} by 10). Lastly, for each region i we classify the remaining 99 regions j as into 33 "strongly", 33 "mediumly", and 33 "weakly" tied regions based on the tertiles of region i's edge weight distribution. The adjacency matrix representing these classifications can be found in figure **??** in the main paper.

A.5 Operationalizing Rainfall with a 2-stage Greedy Grid Search

In our main paper, we operationalize rainfall into our model specifications by converting our continuous precipitation measure from the GHCN data into two binary indicators R_{it}^l and R_{it}^h denoting precipitation exceeding 0.22mm and 16.86mm respectively. We determine these values of 0.22 and 16.86 using a 2-stage greedy grid search to identify which values produce the greatest amount of explanatory power on log regional viewership $\ln V_i t$.

In the first stage stage of our approach, we performed an F-test comparing the "full model":

$$\ln V_{it} = \alpha_i + \tau_t + \gamma_1 R_{it}^l + \epsilon_{it}$$

where $R_{it}^l = \mathbf{1}(P_{it} > p_1)$ against the "restricted model" of:

$$\ln V_{it} = \alpha_i + \tau_t + \epsilon_{it}$$

for all possible values of p_1 starting from 0 going to 5 incremented by 0.01. We selected the value of p_1 that maximized the F-statistic of our test. In the second stage stage of our approach, we performed an F-test comparing the updated "full model":

$$\ln V_{it} = \alpha_i + \tau_t + \gamma_1 R_{it}^l + \gamma_1 R_{it}^h + \epsilon_{it}$$

where $R_{it}^h = \mathbf{1}(P_{it} > p_2)$ against the updated "restricted model" of:

$$\ln V_{it} = \alpha_i + \tau_t + \gamma_1 R_{it}^l + \epsilon_{it}$$

for all possible values of p_2 starting from 0 going to 50 incremented by 0.01. Again, we selected the value of p_2 that maximized the F-statistic of our test.

B Robustness Checks

B.1 Alternate Weather Instrument Specifications

We test 4 alternative operationalizations of our rainfall instrument given by the following four specifications:

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 P_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{7}$$

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 P_{it} + \phi_2 P_{it}^2 + \alpha_i + \tau_t + \epsilon_{it} \tag{8}$$

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 P_{it} + \gamma_1 R_{it}^l + \alpha_i + \tau_t + \epsilon_{it} \tag{9}$$

$$\ln V_{it} = +\beta \ln V_{-it} + \phi_1 P_{it} + \phi_2 P_{it}^2 + \gamma_1 R_{it}^l + \alpha_i + \tau_t + \epsilon_{it} \tag{10}$$

Equation 9 (linear precipitation) replaces the rainfall indicators with a linear term for precipitation in region i on date t (P_{it}). Equation 8 (quadratic precipitation) simply adds an additional quadratic precipitation term (P_{it}^2). Equation 9 (linear precipitation and rainfall indicator) combines both the linear precipitation term with binary indicator for rainfall. Lastly, equation 10 (quadratic precipitation and rainfall indicator) simply adds a quadratic precipitation term to the prior equation. The coefficient estimates of these specifications are presented in table 7 where columns (1) through (4) refer to equations 7 through 10 respectively.

Table 7: Alternative Weather IV

	Dependent variable:			
	$\ln V_{it}$			
	(1) TWFE	(2) IV	(3) TWFE	(4) IV
$\ln V_{-it}$	0.360***	0.545***	0.265***	0.462***
	(0.078)	(0.062)	(0.078)	(0.060)
P_{it}	0.001***	0.002***	0.0005***	0.001***
	(0.0001)	(0.0002)	(0.0001)	(0.0002)
P_{it}^2		−0.00002***		−0.00001***
		(0.00000)		(0.00000)
R_{it}			0.020***	0.017***
			(0.003)	(0.003)
Observations	101,500	101,500	101,500	101,500
R^2	0.980	0.981	0.980	0.981

*p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

Naturally, these different rainfall operationalizations imply different first stages as well. The

modified first stage specifications as presented below:

$$\ln V_{-it} = \alpha_{-i} + \tau_t + \psi_1 P_{-it} + \omega_1 P_{it} + \nu_{-it} \tag{11}$$

$$\ln V_{-it} = \alpha_{-i} + \tau_t + \psi_1 P_{-it} + \psi_2 P_{-it}^2 + \omega_1 P_{it} + \omega_2 P_{it}^2 + \nu_{-it} \tag{12}$$

$$\ln V_{-it} = \alpha_{-i} + \tau_t + \psi_1 P_{-it} + \psi_3 P_{it} + \zeta_1 R_{-it}^l + \eta_1 R_{it}^l + \nu_{-it} \tag{13}$$

$$\ln V_{-it} = \alpha_{-i} + \tau_t + \psi_1 P_{-it} + \psi_2 P_{-it}^2 + \psi_3 P_{it} + \psi_4 P_{it}^2 + \zeta_1 R_{-it}^l + \eta_1 R_{it}^l + \nu_{-it} \tag{14}$$

The first stage regression estimates can be found in table 8 where columns (1) through (4) refer to equations 11 through 14 respectively.

Table 8: Alternative Weather First Stage

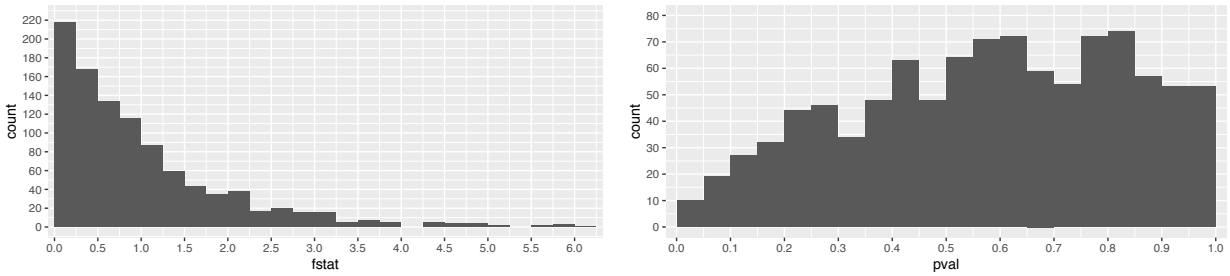
	Dependent variable:			
	$\ln V_{-it}$			
	(1) TWFE	(2) IV	(3) TWFE	(4) IV
P_{-it}	0.008^{***}	-0.010^{***}	0.006^{***}	-0.022^{***}
	(0.00004)	(0.0002)	(0.0001)	(0.0003)
P_{-it}^2		0.002^{***}		0.003^{***}
		(0.00003)		(0.00003)
R_{-it}^l			0.046^{***}	0.120^{***}
			(0.001)	(0.001)
P_{it}	0.0001^{**}	0.0002^{***}	0.0001^{*}	0.0002^{***}
	(0.00004)	(0.0001)	(0.00004)	(0.0001)
P_{it}^2		-0.00000^{***}		-0.00000^{***}
		(0.00000)		(0.00000)
R_{it}^l			0.0004	−0.0001
			(0.001)	(0.001)
Wald F-stat	38777	33881	17146	28361
Observations	101,500	101,500	101,500	101,500
R^2	0.900	0.901	0.900	0.901

*p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

B.2 Placebo Testing the Weather

We generate a placebo draw of our data by just randomizing the order of the rainfall indicators. We are careful to ensure that the pair of indicators are not broken up. We use this procedure to

Figure 6: Placebo Histograms



(a) Placebo First Stage Wald F-statistic Histogram

(b) Placebo IV Estimate P-value Histogram

generate 1000 placebo draws of our data. We estimate equation 1 using IV for each one of these draws. We report a histogram of the estimated first stage Wald F-statistic using these 1000 placebo draws in figure 6(a). We also report a histogram of the p-value IV coefficient estimate of $\ln V_{-it}$ of these draws in figure 6(b).

B.3 Linear-in-Means Specification

To ensure robustness of our results, we see if we can detect causal cross-region peer effects using an alternate model specification. Here, we look at the standard linear-in-means specification:

$$\ln V_{it} = \beta \left(\frac{1}{|\boldsymbol{U}_i|} \sum_{j \in \boldsymbol{U}_i} \ln V_{jt} \right) + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{15}$$

The only difference between this specification and our main specification given by equation 1 is the main independent variable of interest. For the linear-in-means specification, we use $\left(\frac{1}{|\boldsymbol{U}_i|} \sum_{j \in \boldsymbol{U}_i} \ln V_{jt} \right)$, the average log viewership of the regions in $\boldsymbol{U}_i$. We estimate this specification using both TWFE and IV. The results are presented in table 9.

B.4 "Inside-out" Specification

We also see if we can detect cross-region peer effects using the two following "inside-out" model specifications:

$$\ln V_{-it} = \xi \ln V_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \tag{16}$$

$$\ln V_{-it} = \xi \ln V_{it} + \zeta_1 R^l_{-it} + \zeta_2 R^l_{-it} + \alpha_{-i} + \tau_t + \nu_{-it} \tag{17}$$

The results of estimating these two specifications are presented in table 10

Table 9: Linear-in-Means Specification

	Dependent variable:	
	$\ln V_{it}$	
	(1) TWFE	(2) IV
$\frac{1}{\lvert \boldsymbol{U}_i \rvert} \sum_{j \in \boldsymbol{U}_i} \ln V_{jt}$	0.0003^{***}	0.0001^{***}
	(0.00000)	(0.00002)
R^l_{it}	0.020^{***}	0.021^{***}
	(0.003)	(0.003)
R^h_{it}	0.016^{***}	0.018^{***}
	(0.004)	(0.004)
Observations	101,500	101,500
R^2	0.981	0.981

*p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

B.5 Auto-regressive Models

To address concerns about potential time-series interactions, we test model specifications that include autoregressive terms:

$$\ln V_{it} = \beta \ln V_{it} + \lambda_1 \ln V_{i(t-1)} + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{18}$$

$$\ln V_{it} = \beta \ln V_{it} + \lambda_1 \ln V_{i(t-1)} + \lambda_2 \ln V_{i(t-2)} + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{19}$$

$$\ln V_{it} = \beta \ln V_{it} + \lambda_1 \ln V_{i(t-1)} + \lambda_2 \ln V_{i(t-2)} + \lambda_3 \ln V_{i(t-3)} + \gamma_1 R^l_{it} + \gamma_2 R^h_{it} + \alpha_i + \tau_t + \epsilon_{it} \tag{20}$$

All three of these equations are based off our main model specification, equation 1. The first equation adds in a single AR(1) term, $\ln V_{i(t-1)}$, the log viewership of region i on date $t-1$. The second equation adds an additional AR(2) term, $\ln V_{i(t-2)}$, and the third equations further includes an AR(3) term, $\ln V_{i(t-3)}$. The estimated results using of these model specifications using both TWFE and IV are included in table 11.

For both TWFE and IV estimates, we see that all 3 autoregressive model specifications produce nearly identical estimates to the results presented in columns (3) and (4) of table 3. Again, these results do seem to suggest the TWFE estimates may be still biased, given that they are consistently and substantially more positive than IV estimates across all our estimated results.

Table 10: Inside-Out Specifications

	Dependent variable:			
	$\ln V_{-it}$			
	(1) TWFE	(2) IV	(3) TWFE	IV (4)
$\ln V_{it}$	0.099***	0.078***	0.098*** 0.045*	
	(0.008)	(0.026)	(0.008)	(0.025)
$\ln R^{l}_{-it}$			0.067***	0.065***
			(0.001)	(0.001)
$\ln R^{h}_{-it}$			0.210***	0.224***
			(0.004)	(0.007)
Observations	101,500	101,500	101,500	101,500
R^2	0.909	0.909	0.910	0.907

*p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.

B.6 How does the Hierarchical Clustering Threshold Affect Estimates?

In this section, we examine how sensitive our results were to our choice of hierarchical clustering cutoff threshold. We re-estimated equation 1 using both TWFE and IV for many different threshold options. These are presented in table 12.

Table 11: Estimation of Autoregressive Models

	Dependent variable:					
	$\ln V_{it}$					
	(1) TWFE	(2) IV	(3) TWFE	(4) IV	(5) TWFE	(6) IV
$\ln V_{-it}$	0.850***	0.360***	0.889***	0.324***	0.899***	0.266***
	(0.013)	(0.053)	(0.012)	(0.055)	(0.012)	(0.052)
$\ln V_{i(t-1)}$	0.490***	0.509***	0.408***	0.439***	0.394***	0.430***
	(0.027)	(0.026)	(0.023)	(0.022)	(0.024)	(0.023)
$\ln V_{i(t-2)}$			0.152***	0.133***	0.104***	0.088***
			(0.006)	(0.006)	(0.006)	(0.006)
$\ln V_{i(t-3)}$					0.103***	0.095***
					(0.006)	(0.006)
R^l_{it}	0.011***	0.012***	0.012***	0.012***	0.012***	0.012***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
R^h_{it}	0.014***	0.015***	0.015***	0.016***	0.014***	0.015***
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
Observations	101,000	101,000	100,500	100,500	100,000	100,000
R^2	0.986	0.986	0.987	0.986	0.987	0.986

*p<0.05; **p<0.01; ***p<0.001. Cluster-robust standard errors are reported.