Social Media Sharing and Online News Consumption

by

Michael Zhao

B.A. University of California, San Diego (2011)
M.A. New York University (2014)

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Science in Management Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Signature redacted

Author . . . . . .
Sloan School of Management
August 10, 2018

Signature redacted

Certified by . . . . . .
Sinan Aral
David Austin Professor of Management
Thesis Supervisor

Signature redacted

Accepted by . . . . . .
Ezra Zuckerman Sivan
Alvin J. Siteman (1948) Professor of Entrepreneurship and Strategy
Deputy Dean
Professor, Technological Innovation, Entrepreneurship, and Strategic
Management and Work and Organization Studies

Social Media Sharing and Online News Consumption

by

Michael Zhao

Submitted to the Sloan School of Management
on August 10, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Management Research

Abstract

The ever increasing ubiquity of social media platforms has led to the emergence of an incredibly important positive feedback loop between social media sharing and online content consumption. The potential of this feedback loop is critical to marketers, publishers, politicians, and beyond. However, identifying causal effects in this context is very difficult. The data requirements are quite demanding, calling for data from both social media platforms and content producers. In addition, feedback loops inherently suffer simultaneous equation bias. Using regional rainfall as a natural experiment, we use a novel panel-IV strategy to identify positive and significant cross-region "peer effects" in online news viewership: a 1% increase in within-region viewership causes external viewership to increase by approximately 0.06%. Moreover, evidence suggests that social network sharing is a primary driver of these peer effects. We find that the peer effect is stronger on viewership referred from social network sources compared to viewership referred from search engines. Beyond this, we find that social network connectivity moderates the strength of this peer effect: "strongly-connected" regions exhibit more positive and significant peer effects relative to more "weakly-connected" ones. Our provides a first step in understanding how social media platforms generate value for online content producers.

Thesis Supervisor: Sinan Aral
Title: David Austin Professor of Management

Acknowledgments

We thank Christos Nicolaides for his time and guidance about working with the NOAA weather data. We also thank Dean Eckles, Paramveer Dhillon, Soroush Vosoughi, and the members of the Sloan Social Analytics Lab for all their helpful comments and suggestions.

Contents

1 Social Media Sharing and Online News Consumption **13**

1.1 Introduction 13

1.2 Literature 17

1.2.1 Social Media and News Consumption 17

1.2.2 Peer Influence and Social Contagion 22

1.3 Data and Empirical Strategy 23

1.3.1 Data Sources 23

1.3.2 Empirical Strategy 24

1.3.3 Regression Specifications 28

1.4 Results 33

1.4.1 Peer Effects in Online News Viewership 33

1.4.2 What's Driving these Peer Effects? 38

1.4.3 Robustness Checks 39

1.5 Conclusions 40

A Tables **43**

B Additional Data Processing **49**

B.1 NYT and NOAA Data Processing 49

B.2 Determining the Weather of Cities 50

B.3 Hierarchical Clustering 52

B.4 Twitter Data Processing 55

List of Figures

1-1 Twitter Cascades and Click-through Web Traffic of 3 NYT Articles . 20

1-2 Northeast Cities 27

1-3 Regional and Day to Day Viewership Variation 34

1-4 Rainfall for Top 100 Regions 34

1-5 Region-to-Region Adjacency Matrix 35

B-1 Northern California Cities and Weather Stations 51

B-2 Included Regions in the Continental United States 54

List of Tables

A.1 Description of Variables used in Regressions 43
A.2 Estimated Peer Effect 44
A.3 First Stage 44
A.4 Regional Heterogeneity 45
A.5 WOM vs Search 45
A.6 How Social Network Connectivity Mediates Peer Effect Strength . . . 45
A.7 Clustering Cutoff and Estimated Peer Effects 46
A.8 Estimation of Autoregressive Models 47

Chapter 1

Social Media Sharing and Online News Consumption

1.1 Introduction

The advent of social networking platforms, much like other innovations in communications technology, has dramatically changed the way we consume content. This is especially true for online digital content—nearly every single piece of content is affixed with some variation of "like, comment, and share". Multiple surveys [2, 38, 26] report that people use social media to both consume and disseminate content. Moreover, businesses recognize that social media and other forms of online word-of-mouth (WOM) play a major role in the visibility and consumption of their online content. Correspondingly, the piles of money thrown at viral marketing campaigns or social network advertising only seem to grow larger with each passing year.

Social media holds enormous potential. There are numerous examples of massively viral pieces of content that with billions of views subsequently generating millions of dollars from online advertising in the process. This unprecedented ability to monetize content viewership has spawned entirely new professions such as bloggers, YouTubers, Instagram models, etc. Even popular culture has been affected—it's hard to imagine how non-English songs like Despacito or Gangnam Style could've managed to reach the top of the US charts in the absence of social media. Perhaps most importantly,

these platforms can be harnessed by motivated (and potentially malicious) actors to achieve social or political action. In the wake of the 2016 US presidential election, understanding the role of social media platforms in disseminating fake or false news has been a major topic of interest both politically and academically [1, 31, 47]. Ultimately, 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.

It is due to the potential of this feedback loop social media has been hailed as a possible savior to the struggling news industry. Since the turn of the millennium, news organizations have generally struggled to adapt to the technological disruptions brought about by the internet. Pew Research reports that US newspaper circulation fell from 55.8 million in 2000 to 34.7 million in 2016 [11]. Advertising revenue has also dramatically declined, from $49.4 billion in 2005 to just $18.3 billion in 2016. In response to such trends, some newspapers have shifted to "subscription-first" models, focused on growing the subscription base rather than maximizing advertising revenue. Despite this shift in strategy, the importance of digital advertising is still readily apparent—growing from 17% of newspapers' advertising revenue in 2011 to 29% in 2016. Moreover, digital news consumption dramatically risen, now with 93% of US adults consuming at least some of their news content from an online source [44]. Social media plays an major role in this growth, with over half the US population now getting some news from social media [36]. As people spend more and more of their free time on social media [23], it has become increasingly important to consider the role of social media in the business strategy of news organizations (and content producers in general). In particular, the question of whether social media sharing drives news consumption has never been more critical.

Unfortunately, answering this question is extremely difficult. The are numerous econometric challenges that make identification of causal effects very difficult in this context. First off, social media sharing and news content consumption are confounded by "inherent newsworthiness". For example, both social media activity

and news viewership both soar during major events like the terrorist attacks, extreme weather disasters, and presidential elections. In addition, feedback loops are structurally endogenous, suffering from a particular kind of endogeneity called simultaneous equation bias or simultaneity. Beyond the econometric challenges, the mere data requirements to address this question are difficult to satisfy. Ideally, one needs detailed data from both social media platforms and content producers. From the perspective of a newspaper, it is implausible to collect the necessary data from a social media platform, especially considering all the recent negative news coverage about the data practices of these platforms. At the same time, these platforms don't seem to have any meaningful reason to acquire the proprietary data of a particular content producer. Moreover, it seems quite dubious for a newspaper to simply offer their proprietary browsing data to a social media company.

To address these challenges, we develop a novel empirical approach defined by two key features: a region-level of analysis and "peer effects" framework. First, the region-level of analysis enables us to leverage fixed effects to control for a significant amount of unobserved heterogeneity. More importantly, this allows us to look for "local" instruments—variables that shock particular regions but not others. In our case, we exploit regional rainfall patterns as a source of exogenous variation to address the econometric challenges. Second, reformulating our objective in terms of identifying causal peer effects in news consumption eliminates the need to acquire high-quality social media sharing data[1]. We use this approach to study the online viewership of the New York Times (NYT). Combining the NYT's web activity logs with historical weather records, we identify positive and significant causal "peer effects" in cross-region news consumption: on average a 1% increase in within-region (or regional) viewership generates approximately 0.06% additional outside-region (or external) viewership. Unsurprisingly, the strength of this peer effect varies significantly, with higher viewership regions exhibiting stronger peer effects. In addition, our results show that including both region and time fixed effects go a long way in

[1]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 network data in this study to show that social network sharing is a primary mechanism behind these peer effects as a secondary result.

reducing the potential bias on the estimated parameters.

While our approach allows us to causally estimate cross-region peer effects, one important limitation is that it does not provide any insight in the underlying mechanism driving these effects. One possible explanation—which we think to be the case—is that social media (and generally WOM) sharing is driving these effects: viewership in one region leads to sharing in that region which produces viewership in other regions. However, these peer effects could also be the consequence of search engine or news aggregator ranking effects. Under this line of reasoning, viewership in one region leads to higher search engine or news aggregator ranking which positively impacts viewership in other regions. To investigate which one of these mechanisms tends to dominate empirically, we exploit the NYT's webpage referrer data. Specifically, we estimate the cross-region viewership effect on traffic referred from social network and WOM sources (Facebook, Twitter, and Email) and on traffic referred from search engines and news aggregators (Google, Yahoo, and Bing). We find a significantly stronger effects on social sources, supporting the social media sharing mechanism.

Furthermore, if social media and WOM sharing is indeed the dominant mechanism behind these peer effects, then we would expect region-to-region link density to moderate the strength of the viewership peer effect. To illustrate this point, suppose we have 3 regions of equal population A, B, and C. Further assume that A is more strongly connected to B than to C. Given this, sharing from A should reach a wider audience and thereby generate more viewership in B compared to C. We test this hypothesis by building a directed[2] region-to-region followee-follower connectivity network using parsed Twitter[3] user data. Our results are essentially in line with our predictions: regional viewership more positively and significantly impacts the viewership in "strongly"-tied regions relative the the viewership in "weakly"-tied regions.

In summary, we find positive and economically significant peer effects in cross-

[2]Meaning that the tie from i to j is distinct from the tie from j to i.

[3]A popular micro-blogging platform that facilitates the dissemination short messages called "tweets" from users to their "followers"—other users who opt-in to receive a particular user's tweets. (`http://www.twitter.com/`)

region viewership of online news. Moreover, our findings suggest that social media and WOM sharing are a primary driver of these peer effects. The contribution of our work is twofold. First, we develop a methodology to allow content producers to determine whether social media sharing is helping or hurting their businesses. Second, this paper, to the best of our knowledge, is the first paper that causally identifies a positive effect of social media sharing on online news consumption. Naturally, our findings have important implications for news organizations (and possibly online content creators generally). Specifically, our results imply that it may be worthwhile to pursue specific interventions to increase the amount of social media sharing. Possible interventions could range from simple encouragements to share, redesigning the website user interface, or even offering monetary incentives to share.

1.2 Literature

Our paper is primarily draws from two diverse streams of research. First, our work relates topically to the literature on the impact of social media on news consumption or viewership. Second, our methodological approach is inspired by the some of work on social contagion and peer influence in social networks. We briefly cover the relevant literature in two separate subsections below.

1.2.1 Social Media and News Consumption

Over the past decade, we have simultaneously witnessed the rise of social media and the decline of the news industry. This has spawned a recurring debate on how exactly social media is impacting the news industry: is social media contributing to the decline of the news industry? Or is it a new business opportunity for news organizations to exploit[4]?.

The academic discussion surrounding this topic has been framed in terms on

[4]Two poignant headlines that highlight this debate: "Is social media destroying the news?" (`https://www.denverpost.com/2017/03/24/social-media-destroying-news/`) and "3 Ways Social Media Can Save the Newspaper Industry" (`https://socialmediaweek.org/blog/2013/08/ways-social-media-is-saving-the-newspaper-industry/`)

whether social media is a substitute or a complement for news consumption. Various theoretical mechanisms have been proposed to explain the potential patterns of substitution or complementarity. Starting with the substitution side, social media might be decreasing news consumption through a "information redundancy effect". Shared news articles can 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 due to a "time displacement effect". Here, the idea is that social media and news consumption compete with each other for a consumer's time: as time spent on one medium increases, usage of other mediums decrease [29].

On the complementarity side, social media may be increasing news consumption through a "content discovery effect". Sharing a piece of content on social media can increase its audience by allowing it to reach people who wouldn't have seen it otherwise. In fact, the promotional power of social media may be particularly effective due factors such as homophily [32] or "communication utility" [8]. In the case of homophily—the idea that "birds of a feather flock together"—we tend to connected to others similar to us, meaning that content that is shared by our peers is likely to be more relevant to us. Alternatively, we might derive communication utility from clicking-through to a shared news article and then discussing the article content with the person who shared it with us. Reading shared content is also an important aspect to maintaining a similar information set with our peers. Another possible cause of complementarity may result from a "habit formation effect". Since social media users are routinely exposed to news while using social media, news consumption may conceivably become part of a daily routine. Under this scenario, if social media use is then disrupted, a person may entirely abandon the other elements of this routine like reading the news.

As a further point of distinction, we would like to emphasize the difference between direct/short-term vs indirect/long-term effects. While this distinction is rather imprecise, one distinguishing feature is whether or not a specific effect can operate

at an individual article level. If so, we can consider it as a direct or short-term effect. For example, it is perfectly sensible to ask whether or not increased sharing of a particular news article will lead to decrease its consumption through information redundancy or increase its consumption content discovery[5]. It makes considerably less sense to ask the same question of time displacement or habit formation. We highlight this distinction because different mechanisms may be dominating at different time scales. Namely, even if substitution dominates in the long-term (time displacement > habit formation), it still might be worthwhile to invest in a social media campaign if complementarity dominates in the short-term (knowledge discovery > information redundancy).

Despite a fairly rich theoretical literature, researchers have only just begun to be explore this topic empirically. 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). An early case study by Aral and Hanselmann [3] seems to indicate potential for both, at least at the article level. They visualize the Twitter cascades and click-through web traffic of three NYT articles reproduced with permission in Figure 1-1. In the case of the first article, it seems that social media activity and news consumption are largely independent of each other. The second article may be a case where substitution effects dominate as social media use seems to be cannibalizing content viewership. Lastly, the third article might be an example of a positive feedback loop at work which supports the idea of complementarity.

Though there might be significant variation in the substitution and complementarity patterns of news content at the article level, empirical work at the aggregate level suggests complementarity effects dominate, at least in the short-term. Early work by Hong [27] looks at the the Twitter adoption and subsequent online traffic of 337 newspapers from the period of Jan 2007 to Dec 2010. Using a monthly panel, he found a positive association between Twitter adoption and the number of unique visitors to a newspapers website even after including newspaper fixed effects and non-

[5]Or in the case both, which effect dominates.

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

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

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 seems to line up with periods of high web traffic.

parametrically controlling for the time trend. Recent work by Sismeiro and Mahmood [41] more directly addresses this question of substitution or complementarity. Specifically, they examine the effect of a major Facebook outage on hourly web traffic to a major European news website. They find that during the outage, they find that a significant decrease in both traffic and unique visitors that lasts well beyond the outage hours. These two papers, to the best of knowledge, represent the empirical work on the impact of social media on news consumption.

One limitation in the current empirical literature is that this question of substitution vs complementarity of social media as a whole has somewhat obscured the identification of actionable causal insights. In particular, the specific question of whether social media sharing can generate viewership for news organizations has been overshadowed. In terms of counterfactuals, the former compares a social media vs no social media world. On the other hand, the counterfactuals compared in the latter question pertain to exogenously varying the degree to which content is shared on social media.

Causal identification is of particular importance to the social media sharing question. As mentioned earlier, social media sharing and news consumption a likely positively associated due to inherent newsworthiness. Naively interpreting this association as causal may be disastrous. For example, a news organization might adopt a strategy to promote sharing of content on social media that decreases viewership in the case where the information redundancy effect dominates. Even in the case where sharing does generate additional viewership, it's likely the positive association overestimates the true causal effect, which could lead to misallocation of resources.

As a final point in this subsection, it is worth mentioning the closely related work on the impact of news aggregators[6] on news consumption. As both social media and news aggregators act as intermediaries between consumers and news content, it should come as little surprise that the literatures share a number of commonalities. Much like the research on social media, work on this topic has focused on whether or not these aggregators are substitutes or complements for news content consumption. In fact, the mechanisms that potentially drive the substitutability or complementarity of social media also apply to news aggregators as well.

A number of studies have studied the impacts of news aggregator empirically. With a series of field experiments, Dellarocas et al.[21] find that displaying more information about articles on a news aggregator decreases the probability that readers will click through to the full article. This result suggests that an information redundancy effect occurs, at least at the level of individual articles. However, the other empirical research on this topic suggests that the complementarity effect generally dominates in the aggregate. Calzada and Gil [15] and Athey et al. [7], both studying a Google News shutdown in Spain, independently find that the shutdown caused news website visits to drop by 11% and 10% respectively. A related paper by Chiou and Tucker [18]—which exploits a dispute between Google News and the Associated Press—also finds evidence to support complementarity.

[6]News aggregators—websites or applications such as Google, Bing, Yahoo News, or Apple News—use human editorial judgement, computer algorithms, or some combination thereof to collect and curate news content produced by other parties.

1.2.2 Peer Influence and Social Contagion

Social contagion—the spread of thoughts, emotions, and behaviors—has been a major topic of interest for social scientists across multiple disciplines for decades. Theoretically, much work has been devoted to modeling the process of diffusion in networks [25, 48, 28, 16, 24]. Empirically, researchers have shown a wide variety of phenomenon such obesity [19], emotion [20, 30], voting participation [13], exercise [5], and even microfinance loans [10] can all propagate through social networks. Others have concentrated on understanding the factors that may influence the social contagion process itself [9, 6, 12, 46, 17].

One major focus in this literature is dedicated to better understanding the empirical fact that behavior or outcomes tend to cluster among connected peers. One explanation for this tendency is peer influence, namely my behavior is driving your behavior (or vice versa). However, these observed similarities may also be result other factors such as homophily [34], correlated exposure, or simultaneity. In the case of homophily, similarities in peer outcomes result from the fact that ties are formed between individuals that are more alike from the outset. For instance, Smirnov and Thurner [42] find that similarities in academic achievement of peer groups is driven largely by selection: students tend to form ties with those of similar academic achievement. As for the case of correlated exposure, clustered behavior results from exposure to the same or similar events. The classic example of this is with umbrella use: everyone has their umbrella because its raining outside. Regarding the case of simultaneity, correlated outcomes result from peers simultaneously co-influencing each other. Here there is a potential for fairly extreme snowball effects to occur: increasingly escalating dares or the formation of mobs.

Unfortunately, peer influence is often confounded by these other factors in observational data, something Manski [33] refers to as the "reflection problem". As one might expect, distinguishing peer influence from these other factors is critical important question since the efficacy of specific policy interventions depends on what exactly is driving the clustering of outcomes. 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 [4], structural modeling [22], instrumental variables approaches [14, 45, 20, 5], and randomized field experiments [6, 13, 10, 30].

In particular, our approach quite similar to the instrumental variables approaches used by Coviello et al. [20] and Aral and Nicolaides [5]. These papers both exploit regional rainfall as a source of exogenous variation to study emotional and exercise contagion respectively. However, one key difference with our empirical approach is that we focus on aggregate region-level outcomes rather than individual-level outcomes.

1.3 Data and Empirical Strategy

1.3.1 Data Sources

Our dataset is constructed primarily from the proprietary web activity logs of the NYT. This data covers each one of the billions of events tracked by the NYT's internal servers from April 3, 2013 to October 31, 2013[7]. The raw data is very rich and granular consisting of millisecond-accurate timestamps, IP-address derived geolocation data, accessed URL, referrer URLs, among many other fields. This data is relatively large, consisting of billions of individual (over 24 terabytes of storage space). We aggregate these events 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. For the purposes of our research, we restrict ourselves to the the top 500 US regions in terms of viewership which account for approximately 70% of total content views. 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).

We supplement this data with two additional data sources: the NOAA's historical

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

weather data complied by Menne et al. [35] and Twitter user and follower data parsed using Tweepy[8] The NOAA 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 sample of 10000 "ordinary"[9] Twitter accounts with a tweet or retweet containing a link to a NYT article during our timeframe[10]. We obtain the followers of these 10000 accounts and record the followers' self-reported location (if available) to construct a region-to-region connectivity graph. A more comprehensive description of all our data processing procedures is detailed in Appendix B.

1.3.2 Empirical Strategy

Estimating the causal effect of social media sharing on news consumption is very difficult. Content viewership, at least in some cases, is likely driven by social media sharing. At the same time, online content such as news provides material for people to share and discuss. Written as a nonparametric system of equations we have:

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

where V denotes content viewership and S denotes social media sharing while ϵ and η denote the error terms associated with the viewership and sharing equations respectively. Unfortunately, such a system inherently suffers from a type of endogeneity

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

[9]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.

[10]A couple of additional notes. First, due to the inherent problems with self-reported locations, we are only able to obtain adequate coverage for the top 100 regions[11] (100 accounts per region). Second, the Twitter account needed to have self-reported profile location that was matchable to one the regions in our dataset. For example, self-reported locations such as "New York City", "Williamsburg, Brooklyn", or "SF Bay Area" work; but locations such as "California", "Earth", or "United States" do not.

often referred to as simultaneity or simultaneous equation bias: by construction, V is structurally dependent on η just as S is structurally dependent on ϵ. Unfortunately, this issue renders any "selection-on-observables" approach[12] insufficient, as bias still persists even when conditioning on the "right" covariates.

Though the classic econometric solution for simultaneity is to use an instrumental variables (IV) approach, finding valid instruments in this context is difficult: most covariates are likely to affect both V and S. One ostensibly good candidate would be something like a Twitter or Facebook service interruption. However, social media platforms work extremely hard to make sure such events are exceedingly rare and are quickly resolved if they do occur. 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 with day-to-day variation in news consumption. For example 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[13] would be strongly associated with increased news viewership since its effectively functioning as a fixed effect for the duration of the outage.

To overcome this challenge, we develop a unique region-level IV strategy that takes advantage of regional sources of exogenous variation allowing us to identify cross-region causal effects. In our case, we use rainfall as our regional shock. While it may not be obvious that the weather has an effect on news viewership and sharing, the basic idea is that when the weather is poor, people are more likely to stay indoors and spend time on the Internet doing such things as reading the news or browsing social media. For example, Sen and Yildirim [40]—who use weather as an instrument to study whether editorial decisions about news content take into consideration clicks—find that viewership of articles is 5%-8% higher on rainy days for a large Indian news provider.

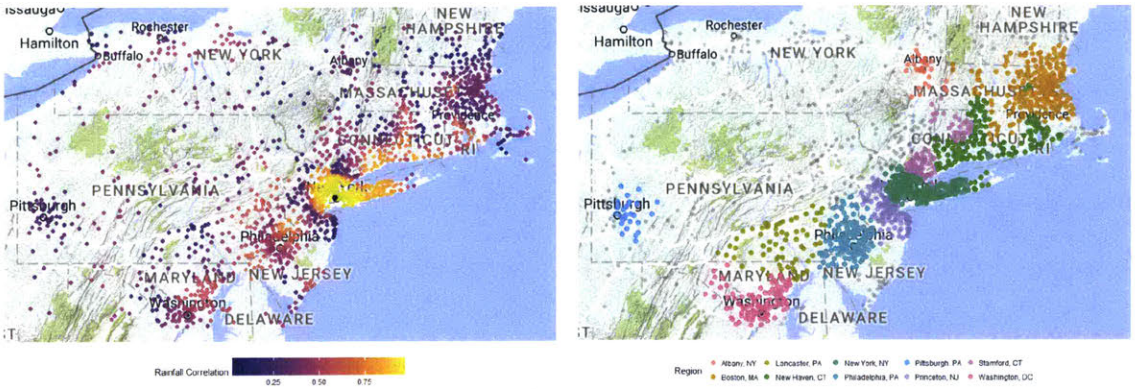
Ideally, we would apply our strategy to examine the cross-region causal effect of

[12]Conditioning on some covariate set or using some kind of matching strategy
[13]Where "1" denotes an outage.

social media sharing on news viewership, there are significant data requirements to overcome. Naturally, our approach relies on the precision of our geolocation data. Unfortunately, there are major concerns about the quality of geolocation data we can get for social media sharing. In our case, we are only able to obtain Twitter users' self-reported profile locations which suffer from all the classic problems associated with self-reported survey data. In particularly, non-response is a significant issue, with around half of tweets/retweets lacking Twitter profile location information. Beyond this, there are serious concerns about accuracy and granularity for the users that do include location information in their profile. 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 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. All this is to say that we are not particularly confident that using this data will produce credible estimates of causal effects.

While we aren't able to directly estimate the cross-region effect of social media on news viewership, we can still determine whether social media sharing is having a positive or negative effect on news consumption. As the quality of geolocation data in the NYT data is quite high, we can reformulate our strategy to identify causal "peer effects" in regional viewership. As mentioned above, instrumental variables is an empirical approach that can address the endogeniety concerns in peer effects

Figure 1-2: Northeast Cities



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

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).

framework. To hopefully clarify our approach, consider the simplified 2-region case:

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

Here, 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 almost certain correlated as a result in variation in day-to-day "inherent newsworthiness" that impacts both regions. Assuming that 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} for the second to identify β.

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

correlation in 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 1-2, 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 represent such a small fraction of NYT articles and pageviews[15] meaning that potential bias will be extremely small. Moreover, newsworthy weather events are likely to induce a negative rather than positive bias on our estimates (i.e., our estimates are more likely to be conservative rather than optimistic). In many instances, such disasters may damage infrastructure or cause evacuations in the affected regions thereby driving down regional viewership but increasing viewership elsewhere.

1.3.3 Regression Specifications

There are two common perspectives to studying peer effects: "outside-in", which focuses on the effect of peers on a focal unit, and "inside-out", which looks at the impact of a focal unit on its peers. Interestingly enough, the two papers that employ a similar weather-IV identification strategy to ours, Coviello et al. [20] and Aral and Nicolaides[5], each take one these perspectives: the former uses inside-out, while the

[14] According to the NYT's internal content tagging system

[15] Especially since 2013 was a rather mild year in terms of weather related disasters. According to [37] only 8 major weather-related disasters occurred during our timeframe.

latter uses outside-in. While both perspectives offer unique insights, for our context, an inside-out perspective has several key advantages.

First, we are only able to crudely model a partial region-to-region connectivity network. This alone already dramatically hurts the reliability of our estimates. This is compounded by the likely extreme heterogeneity in regional effect strength in our context[16] potentially creating even more bias.

Second, in our context, an outside-in approach would generate an extremely odd econometric problem of a shifting instrument set. For example, consider a simple 3-region case with a standard outside-in model: $V_{it} = \alpha + \beta V_{-it} + \gamma R_{it} + \epsilon_{it}$. Viewership in region i on date t (V_{it}) is regressed on the total sum of viewership of all other regions on date t ($V_{-it} = \sum_{j \neq i} V_{jt}$) and rainfall in i on date t (R_{it}). The set of instruments to identify β for region 1 would be $\{R_{2t}, R_{3t}\}$. However, the set of instruments for region 2 is $\{R_{1t}, R_{3t}\}$ while for region 3 it is $\{R_{1t}, R_{2t}\}$. In this 3-region case, each region has 2 instruments, the rainfall of the 2 other regions. This creates an implementation issue since we have 2 instrument vectors, but they don't refer to the same variable throughout the vector. This is not a problem if the region to region peer effect is relatively homogenous (and assuming the rainfall effect impacts each region similarly), often a fairly reasonable simplifying assumption used in other peer effects studies. However, as mentioned in the prior footnote, we should expect a great deal of regional heterogeneity in our case. To our knowledge, we have not discovered any literature that addresses this problem.

Third, even if the shifting instrument issue could be addressed, there's still the remaining problem of potentially large instrument set. Each additional region provides a potential new instrument in that region's rainfall, meaning in our case, we would be working with 500 instruments (and that's only assuming a single instrument from each region's rainfall). Such a large set of instruments may lead to overfitting of the endogenous variables [39].

Lastly and perhaps most importantly, the inside-out perspective generates more

[16]For example, the effect of NYC region is likely to be much stronger than the effect of Tulsa region.

interpretable and actionable managerial insights. Suppose that an online content producer decided to run a targeted marketing campaign to increase viewership in one particular region. The results of a inside-out model would allow us to understand the spillover effect of that campaign to other regions.

Main Specification

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

$$\ln V_{\boldsymbol{i}'t} = \alpha_{\boldsymbol{i}'} + \tau_t + \beta \ln V_{it} + \epsilon_{\boldsymbol{i}'t} \tag{1.1}$$

The dependent variable here is $\ln V_{\boldsymbol{i}'t}$, the log of the total number of views or viewership of NYT content from regions with sufficiently uncorrelated rainfall with region i on date t. More formally, we can write this as $\ln V_{\boldsymbol{i}'t} = \ln(\sum_{j \in \boldsymbol{i}'} V_{jt})$ where $\boldsymbol{i}'$ is the set of all regions k with an absolute rainfall correlation coefficient (with region i) below .25 ($\boldsymbol{i}' = \{k : |\rho(R_k, R_i)| < .25\}$). The main variable of interest is $\ln V_{it}$, the log viewership of region i on date t. It's associated parameter, β, denotes the peer effect of viewership in region i on the viewership of other regions. Since we're using a log-log model specification, we interpret β as an elasticity—an increase of 1% in V_{it} will generate β% more viewership in $V_{\boldsymbol{i}'t}$. As part of our panel specification, we include a set of region and time fixed effects. $\alpha_{\boldsymbol{i}'}$[17] denotes the aggregated fixed effect of regions in $\boldsymbol{i}'$ while τ_t denotes the time-fixed effect of date t[18]. Lastly, $\epsilon_{\boldsymbol{i}'t}$ denotes the error term.

[17]Note this is programmatically equivalent to including a fixed effect term for region i in our model specification as the set $\boldsymbol{i}'$ is fully determined by i.

[18]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 endogeniety. 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 elsewhere.

IV First-Stage

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

$$\ln V_{it} = \alpha_i + \tau_t + \gamma R_{it} + \eta_{it} \tag{1.2}$$

Our instrument for $\ln V_{it}$ is R_{it}, the rainfall in region i on date t. There are a number of ways we can operationalize our rainfall instrument. We simply define R_{it} as a simple binary variable indicating if precipitation in region i exceeded 0.22mm[19] on date t. For the sake of exposition, we will say it was "raining" in region i on date t if $R_{it} = 1$, and "not raining" if $R_{it} = 0$. γ here denotes the effect of rainfall on regional viewership: if it rains in region i on date t, then we can expect the viewership of content in region i on date t to increase by $100 * (e^{\gamma} - 1)\%$. α_i and τ_t denote region and time fixed effects respectively and η_{it} denotes the first stage error term. Since we have a single endogenous variable and single instrument, our system is exactly identified.

Regional Heterogeneity

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

$$\ln V_{i't} = \alpha_{i'} + \tau_t + \beta_1(\ln V_{it} * top_i) + \beta_2(\ln V_{it} * mid_i) + \beta_3(\ln V_{it} * bot_i) + \epsilon_{i't} \tag{1.3}$$

Here, we modify our main model specification by interacting $\ln V_{it}$ with top_i, mid_i, and bot_i—binary variables indicating if a region is in either the top, middle, or bottom thirds in terms of total viewership in our timeframe. Here, we can interpret β_1 as the

[19]This value was determined using a grid search over $[0, 5]$ in 0.01 steps. We selected the cut-off threshold that produces the strongest first-stage. Other cutoff values, and alternative ways of operationalizing our instrument (i.e. using precipitation as a continuous variable, multiple cutoffs, etc.) generally meet the criteria for "strong" instruments. These alternative instrument construction choices do not significantly change our second-stage results.

average peer effect of the top 167 regions, β_2 the average peer effect of the middle 166 regions, and β_3 the average peer effect of the bottom 167 regions.

Channels of Viewership

In order to investigate the mechanism driving these peer effects, we examine the effect of a region's viewership on the viewership of other regions from different channels, namely, Social network and WOM sources and search engines. We use the following 2 regressions:

$$\ln V_{i't}^{WOM} = \alpha_{i'} + \tau_t + \delta_1 \ln V_{it} + \epsilon_{i't}^{WOM} \tag{1.4}$$

$$\ln V_{i't}^{search} = \alpha_{i'} + \tau_t + \delta_2 \ln V_{it} + \epsilon_{i't}^{search} \tag{1.5}$$

These regressions are just substitute different dependent variables into our main model specification. $\ln V_{i't}^{WOM}$ denotes the log sum viewership with Social network or WOM referrers (Facebook, Twitter, Email) of regions with sufficiently uncorrelated weather with i. On the other hand, $\ln V_{i't}^{search}$ denotes he log sum viewership referred from search engines (Google, Yahoo, and Bing) of the same regions δ_1 and δ_2 are the associated peer effect parameters. Again, $\alpha_{i'}$ and τ_t are region and time fixed effects, while $epsilon_{i't}^{WOM}$ and $epsilon_{i't}^{search}$ denote the error terms for $\ln V_{i't}^{WOM}$ and $\ln V_{i't}^{search}$ respectively. If Social network and WOM are a primary driver of peer effects, then we should expect $\delta_1 > \delta_2$.

Twitter Connectivity

To further examine the role of Social network and WOM we also investigate the how Social network connectivity mediates the strength of the peer effect. Using our Twitter follower data, we generate region-to-region followship graph of the top 100 regions. For each region i, we classify each of other 99 j regions as "strong-tied", "medium-tied", and "weak-tied" regions based on tie density. "Strong-tied" regions are top third of regions with the highest number of followee-follower[20] ties, "medium-

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

tied" are the middle third, and "weak-tied" are the lowest third. For this part of our research, we use the following three model specifications:

$$\ln V_{i't}^{strong} = \alpha_{i'} + \tau_t + \theta_1 \ln V_{it} + \epsilon_{i't}^{strong} \tag{1.6}$$

$$\ln V_{i't}^{medium} = \alpha_{i'} + \tau_t + \theta_2 \ln V_{it} + \epsilon_{i't}^{medium} \tag{1.7}$$

$$\ln V_{i't}^{weak} = \alpha_{i'} + \tau_t + \theta_3 \ln V_{it} + \epsilon_{i't}^{weak} \tag{1.8}$$

Again, these regressions just substitute different dependent variables into our main model specification. $\ln V_{i't}^{strong}$ is simply the log sum viewership of strong-tied regions—with sufficiently uncorrelated weather–to i. $\ln V_{i't}^{medium}$ and $\ln V_{i't}^{weak}$ are the same, except for medium- and weak-tied regions. $\alpha_{i'}$ and τ_t are as still as they were before. Similarly, $\epsilon_{i't}^{strong}$, $\epsilon_{i't}^{medium}$, and $\epsilon_{i't}^{weak}$ denote the error terms for $\ln V_{i't}^{strong}$, $\ln V_{i't}^{medium}$, and $\ln V_{i't}^{weak}$ respectively. If Social network and WOM do play a strong role in driving the viewership peer effect between regions, then we should expect $\theta_1 > \theta_2 > \theta_3$.

1.4 Results

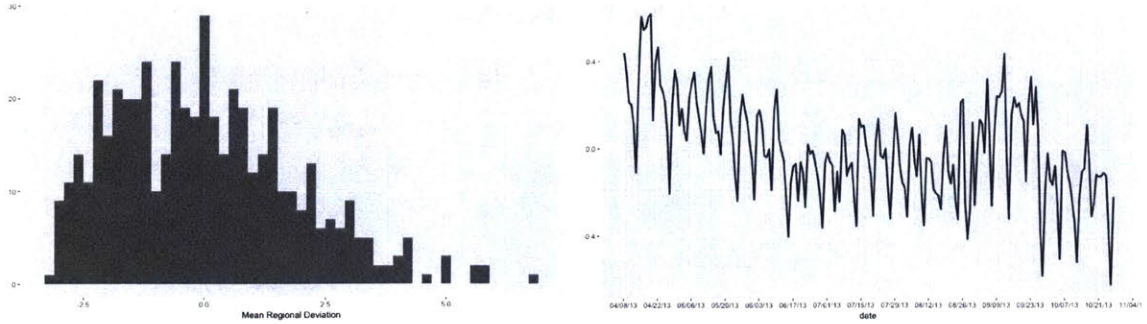
We begin this section with some basic descriptive statistics and visualization of our data. We report the demeaned values for our viewership variables for data privacy reasons table A.1 below. The top 8 variables make use of every observation in our dataset and are used to estimate equations 1 through 5 while the bottom 5 variables are used to estimate equations 6 through 8.

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 1-3) to illustrate the variation in $\log V_{it}$ from region to region and from day to day. Lastly, we present plot of the region-to-region Twitter followee-follower matrix in figure ??.

1.4.1 Peer Effects in Online News Viewership

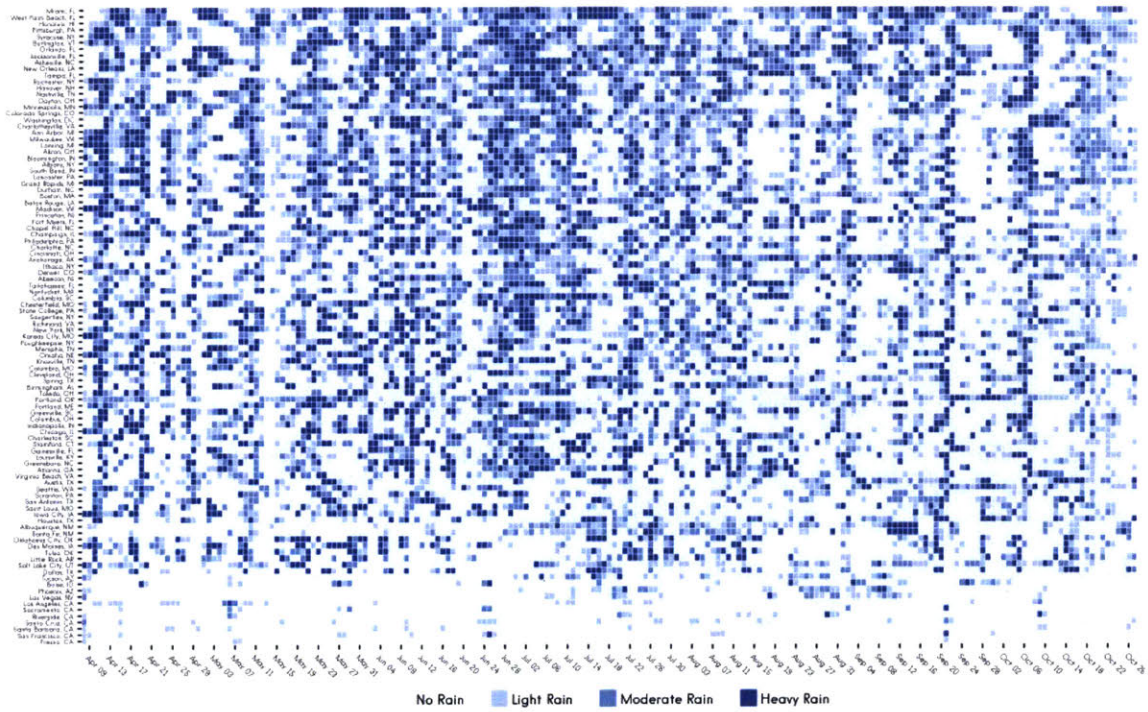
Table A.2 presents the estimation of equation 1.1 using 4 different approaches: pooled-OLS (OLS) in column 1, region fixed effects (RFE) in column 2, region and time fixed

Figure 1-3: Regional and Day to Day Viewership Variation



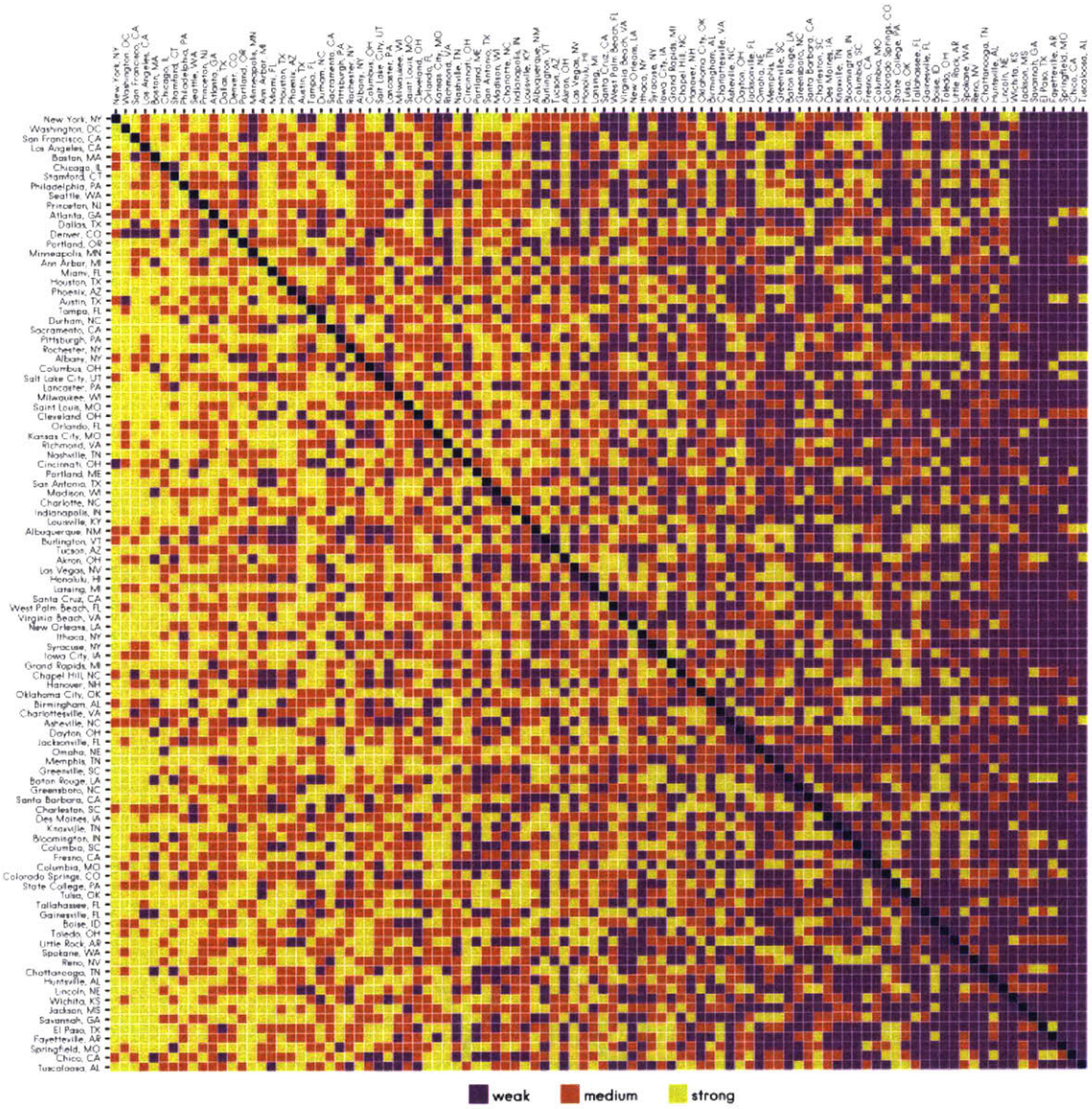
(a) Histogram of Mean Region Deviation (b) Date-to-Date Mean Deviation

Figure 1-4: 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 4 categories: "No Rain", "Light Rain", "Moderate Rain", and "Heavy Rain". No rain corresponds to 0 precipitation while light, moderate, and heavy rain correspond to the first, second, and third tertiles of the non-zero precipitation distribution (precipitation conditional on there actually being some rain).

Figure 1-5: 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 horizontal axis and j index the vertical axis. Excluding the cells where $i = j$ (along the dark blue diagonal), each cell represents whether region j follows many users in region i. Yellow indicates a high degree of followers ("strong"), orange indicates a moderate amount of followers ("medium"), while purple indicates a small amount of followers ("weak"). Looking along the horizontal axis, we see that the leftmost, high population regions, are regions that tend to be more followed. As we mover along the horizontal axis to smaller and smaller regions, we see a greater and greater amount of purple cells.

effects or two-way fixed effects (TWFE) in column 3, and lastly, instrument variables (IV) in column 4. somewhat surprisingly, OLS estimates the peer effect at a highly significant -0.006. A negative value here is rather odd since we often expect major confounding factors such as homophily and correlated exposure to drive estimates toward 1 in a log-log specification.

The OLS result seems mostly driven by the large degree of regional heterogeneity. Once we include region fixed effects, the estimated peer effect jumps quite dramatically to 0.441, perhaps indicating a major confounding factor at work. If we were to interpret this result causally, it would imply that on average, a 1% increase in viewership in region i would yield a .554% increase in the combined viewership of the regions in $\boldsymbol{i'}$, which doesn't seem extremely reasonable.

Indeed, once adding in time-fixed effects—which mitigate the correlated exposure problem by controlling for some of the day-to-day variation in "inherent newsworthiness"—we see the estimated peer effect drop to a potentially credible value of 0.099. However, are there still sources of endogeniety that might be biasing our estimation? Theoretically, we should still suspect some bias since including time-fixed effects doesn't address the issue of simultaneity. In order to explore this, we look to our panel IV estimates.

First-Stage and IV Estimate

Table A.3 presents estimation results of equation 1.2. Before moving onto our IV results, we first need to make sure that our proposed instrument, regional rainfall (R_{it}), satisfies the relevancy requirement. Looking at table A.3, the estimated rainfall effect is highly statistically significant 0.024. This means that when it rains, we expect the viewership in region i to increase by approximately 2.48%. Overall, our first-stage produces an F-stat of 93.14 substantially surpassing [43]'s threshold for strong instruments. We therefore conclude that we do not have a weak instruments problem.

Our second stage produces a peer effect estimate of 0.062 with a cluster-robust standard error of 0.027, indicating statistical significance at the 5% level. Taking the point estimate as given, this implies that on average, a 1% increase in region i's

viewership should increase the combined viewership in regions i' by 0.062%. However, we are not able to conclusively determine whether the IV estimate is significantly different from two way fixed effects panel model. However, the fact that the two-way panel model produces an estimate reasonably close to the IV estimates is seems to be a sign that the region and time fixed effects go a long way in terms of controlling for potential bias.

Regional Heterogeneity

As we've mentioned, we should expect a great deal of heterogeneity in the strength of the estimated peer effect especially when using a log-log model specification. Since the estimated peer effect is interpreted as the cross-region viewership elasticity, we ought to expect the elasticity of the New York region or Washington DC regions to be much more potent than the elasticity of the Albuquerque-NM or Indianapolis regions. Table A.4 presents the estimated results of equation 1.3 using both two-way fixed effects and IV.

Looking at column 1, we see all three estimated coefficients are highly statistical significant for TWFE. However, when moving over to column 2, only the top regions' peer effect is statistically significant at the 5% level for IV (though the middle regions' peer effect is significant at the 1% level). Despite the lack of precision with IV, the trend of the point estimates are exactly what we expected, regions with the more viewership also exhibit the stronger peer effects. Furthermore, even if TWFE is biased, it seems quite implausible that the bias would run in different directions for the top, middle, and bottom regions. What we mean by this, is that it seems quite unlikely this if the estimated peer effect of the top regions is positively biased, that the estimated peer effect of the middle or bottom regions would be negatively biased. Overall, we feel that this is strong evidence indicating heterogeneity in cross-region viewership elasticities of different regions, with higher viewership regions exhibiting stronger effects.

1.4.2 What's Driving these Peer Effects?

While our previous results credibly identify a causal peer effect, it is unclear what the underlying mechanism driving this peer effect. It's important to precisely pin down the primary mechanism as different mechanisms may imply different interventions to best take advantage of this peer effect. If for example, the peer effect is primarily driven by search engines or news aggregators so that increased viewership in one region may be driving up the search or aggregator ranking of a particular piece of content. In this case, news organization may want to invest more into search engine optimization. On the other hand, if social media sharing is the primary driver, then it may be a better idea to offer incentives to encourage people to share content.

Channels of Viewership: WOM vs Search

Table A.5 presents the results of estimating equations 1.4 and 1.5. Columns (1) and (2) report the estimated effect of region i's on the combined viewership in $\boldsymbol{i'}$ referred from Social network and WOM channels using TWFE and IV respectively. Similarly, columns (3) and (4) report the estimated effect region i's viewership on the combined viewership in $\boldsymbol{i'}$ referred from search engines using TWFE and IV respectively.

Interpreting the coefficient estimates themselves may not actually be too meaningful. Referrer data likely suffers from some degree of measurement error due to limits on ability to accurately map referrer URLs to the "channels" we're examining in this section. Moreover, there may be reliability since referrer fields can be erroneously missing (copy-pasting or typing out the link manually) or unable to capture the true referrer (In-person WOM recommendation). However, the important thing to look for in this table is the difference between columns (1) and (3) and columns (2) and (4). In both cases, the estimated coefficient for Social network and WOM is significantly stronger than the estimated coefficient for search engines. In spite of the problems we mentioned with referred data, we feel that this trend is highly suggestive that Social network and WOM are primary driving of the observed peer effects or at the very least, a more significant driver than search engines are.

Social network Connectivity

Table A.6 presents the results of estimating equations 1.6, 1.7, and 1.8. As with results above, we continue to report both TWFE and IV estimates. Columns (1) and (2) report the estimated coefficients for the combined for the regions in $\boldsymbol{i'}$ that are the most densely connected to i. Columns (3) and (4) report the coefficients for the regions $\boldsymbol{i'}$ that are somewhat connected to i. Lastly, columns (5) and (6) report the coefficients for the regions in $\boldsymbol{i'}$ that are only weakly connected to i.

As with table A.5 above, there's probably not too much value in interpreting the estimated coefficients themselves, especially given the all the problems we mentioned with using self-reported Twitter profile locations. Again, the key feature of this table is the comparison of coefficients between columns (1), (3), and columns (2), (4), and (6). Looking at the TWFE columns we see exactly the trend we would expect if Social network and WOM played a major role in driving cross-region peer effects in viewership. The IV results, while not quite as clean, are still consistent with this story. In column (2) we see that region i has a statistically significant effect on its strong-tied regions. The estimated coefficients in columns (4) and (6) are smaller than those in column (2), however the standard errors are fairly large. These IV columns by themselves are not the strongest evidence for our Social network and WOM mechanism. However, when with the results of section 4.2.2 and the clear trend from the TWFE columns, our findings suggest the Social network and WOM do have a major role in driving cross-region peer effects in the viewership of online news.

1.4.3 Robustness Checks

We perform a series of checks to verify robustness of our results. First, we 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 as

long as the threshold produces a reasonable number of clusters. Too few clusters will cause a great deal of measurement error since within-cluster similarity is low. Too many clusters causes the estimated peer effect to go towards 0 due to the greater of small regions with low viewership compared with number of bigger regions. Since the elasticity of small regions are going to be less than the big regions, increasing the ratio of small to big regions is only going to drag down the average cross-region viewership elasticity. These results can be found in Table A.7.

Another major concern is for potential time-series concerns to be in play. For example, 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. In order to this, we included a set of autoregressive model specifications. Including autoregressive terms does not significantly change our results, either qualitatively or quantitatively. The AR model regression results can be found in Table A.8.

1.5 Conclusions

Despite the seeming importance of Social network to the visibility and consumption of online content, not much is actually how much or even whether Social network and WOM can create value for online content producers. The econometric challenge is one of simultaneity—Social network sharing is driving content viewership and content viewership is driving Social network sharing—leading to biased estimates. The solution to this problem is to use instrumental variables, however, finding valid instruments in this is extremely challenging. To overcome this challenge, we develop a unique empirical strategy that takes exploits regional variation in weather as an regional shock. With this, we can identify cross-region causal peer effects in viewership. We find that a 1% increase in regional viewership will generate about 0.06% additional viewership in outside regions. Additionally, we provide strong evidence to suggesting that Social network and WOM is a primary mechanism that drives these peer effects. There are several additional key implications of our work.

On the managerial side, the estimated peer effects are directly analogous to viewership spillovers across regions. It allows businesses to understand how a region-specific marketing campaign might generate viewership outside that region. Naturally, this fact also has major implications for evaluating the effectiveness of such a campaign: a simple region-randomized experiment may underestimate the treatment effect. Combining the peer effect with Social network and WOM mechanism results suggests that Social network sharing, in expectation, should drive the increased consumption of online content. Hence, it might be a good idea for online content producers to offer incentives to induce sharing and promote their content—much like many other companies already do with their products. However, additional research still needs to be done in order to provide an estimate of the net value of such incentives.

There are also some important implications for academic research. Our empirical strategy can be employed in similar contexts where valid instruments may be similarly difficult to find. 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. Additionally, our results also suggest that using both region and time fixed effects can control for a significant amount of endogeneity. While the two-way panel model does still produces theoretically biased coefficients, it is also significantly more precise, and serves as a useful first step at reducing a bias.

Our research provides significant insights into understanding the relationship between social media sharing and online content consumption. However, its unclear to what degree our results will generalize to 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 to consume news content. While we might expect Social network to have a stronger effect on other types of online content (that lack such alternative channels), more research needs to be done.

Additionally, there are many dynamic substitution and complementarity effects

to consider. For example, for some types of news, one might only read a single article about an event meaning that a NYT article would be a substitute for a Washington Post article. On the other hand, for other types of news, reading a single article about an event may make one more likely to read other articles about that event. 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.

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

Appendix A

Tables

Table A.1: Description of Variables used in Regressions

Variable	Description	Mean	St. dev.	Min	Max
Main Specification Variables (1.3.3)					
$\ln V_{i't}$	Demeaned log aggregate views of NYT content in regions with uncorrelated rainfall with i on date t	0.000	0.275	−1.224	0.724
$\ln V_{it}$	Demeaned log views of NYT content in region i on date t	0.000	1.904	−5.892	7.358
R_{it}	Indicator for whether is it raining in region i on date t	0.354	0.478	0	1
Regional Heterogeneity Specification Variables (1.3.3)					
top_i	Indicator for whether i is one the top 167 regions in total viewership	0.334	0.472	0	1
mid_i	Indicator for whether i is one the middle 166 regions in total viewership	0.332	0.471	0	1
bot_i	Indicator for whether i is one the bottom 167 regions in total viewership	0.334	0.472	0	1
Search vs Social Variables (1.3.3)					
$\ln V_{i't}^{WOM}$	Demeaned log aggregate views of NYT content referred from Facebook, Twitter, and email in regions with uncorrelated rainfall with i on date t	0.000	0.381	−1.589	1.234
$\ln V_{i't}^{search}$	Demeaned log aggregate views of NYT content referred from Google, Yahoo, and Bing in regions with uncorrelated rainfall with i on date t	0.000	0.267	−1.085	0.690
Twitter Connectivity Variables (1.3.3)					
$\ln V_{i't}^{strong}$	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
$\ln V_{i't}^{medium}$	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
$\ln V_{i't}^{weak}$	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

Table A.2: Estimated Peer Effect

	Dependent variable:			
	$\ln V_{i't}$			
	(1) OLS	(2) RFE	(3) TWFE	(4) IV
$\ln V_{it}$	-0.006^{***}	0.441^{***}	0.099^{***}	0.062^{*}
	(0.0005)	(0.023)	(0.008)	(0.027)
Region FE	N	Y	Y	Y
Time FE	N	N	Y	Y
Observations	101,500	101,500	101,500	101,500
R^2	0.002	0.0561	0.915	0.914

Table A.3: First Stage

	Dependent variable:
	$\ln V_{it}$
R_{it}	0.024^{***}
	(0.003)
F-stat	93.14
Observations	101,500
R^2	0.979

Table A.4: Regional Heterogeneity

	Dependent variable:	
	$\ln V_{i't}$	
	(1) TWFE	(2) IV
$\ln V_{it} * top_i$	0.073***	0.093**
	(0.003)	(0.036)
$\ln V_{it} * mid_i$	0.044***	0.074
	(0.007)	(0.044)
$\ln V_{it} * bot_i$	0.034***	0.025
	(0.002)	(0.035)
Observations	101,500	101,500
R^2	0.917	0.914

Table A.5: WOM vs Search

	Dependent variable:			
	$\ln V_{i't}^{WOM}$		$\ln V_{i't}^{search}$	
	(1) TWFE	(2) IV	(3) TWFE	(4) IV
$\ln V_{it}$	0.164***	0.153**	0.084***	0.034
	(0.014)	(0.052)	(0.007)	(0.027)
Observations	101,500	101,500	101,500	101,500
R^2	0.778	0.909	0.778	0.906

Table A.6: How Social Network Connectivity Mediates Peer Effect Strength

	Dependent variable:					
	$\ln V_{i't}^{strong}$		$\ln V_{i't}^{medium}$		$\ln V_{i't}^{weak}$	
	(1) TWFE	(2) IV	(3) TWFE	(4) IV	(5) TWFE	(6) IV
$\ln V_{it}$	0.361***	0.116*	0.288***	0.059	0.193***	0.078
	(0.035)	(0.053)	(0.030)	(0.060)	(0.029)	(0.061)
Observations	20,300	20,300	20,300	20,300	20,300	20,300
R^2	0.977	0.971	0.973	0.967	0.982	0.982

Table A.7: Clustering Cutoff and Estimated Peer Effects

Thres.	N	TWFE	IV	Thres.	N	TWFE	IV
.250	970	0.068***	0.052***	.650	324	0.097***	0.065**
		(0.003)	(0.017)			(0.009)	(0.028)
.275	926	0.070***	0.051***	.675	299	0.100***	0.068*
		(0.003)	(0.018)			(0.010)	(0.032)
.300	871	0.072***	0.053***	.700	279	0.114***	0.071*
		(0.003)	(0.019)			(0.008)	(0.033)
.325	824	0.074***	0.056***	.725	258	0.115***	0.067*
		(0.003)	(0.019)			(0.008)	(0.036)
.350	774	0.076***	0.051**	.750	230	0.117***	0.046
		(0.004)	(0.019)			(0.009)	(0.041)
.375	732	0.077***	0.055**	.775	212	0.118***	0.037
		(0.004)	(0.020)			(0.010)	(0.044)
.400	700	0.078***	0.049**	.800	186	0.133***	0.054
		(0.004)	(0.021)			(0.012)	(0.047)
.425	658	0.080***	0.059**	.825	169	0.141***	0.047
		(0.005)	(0.021)			(0.014)	(0.050)
.450	625	0.080***	0.055**	.850	156	0.145***	0.061
		(0.005)	(0.023)			(0.015)	(0.051)
.475	590	0.082***	0.059**	.875	144	0.150***	0.051
		(0.005)	(0.023)			(0.016)	(0.055)
.500	542	0.083***	0.054**	.900	134	0.149***	0.036
		(0.006)	(0.024)			(0.017)	(0.056)
.525	504	0.085***	0.061**	.925	126	0.164***	0.041
		(0.006)	(0.025)			(0.018)	(0.060)
.550	465	0.087***	0.069**	.950	116	0.166***	−0.027
		(0.007)	(0.026)			(0.019)	(0.072)
.575	424	0.090***	0.077***	.975	107	0.173***	0.0005
		(0.006)	(0.025)			(0.020)	(0.067)
.600	387	0.090***	0.074**	1.000	101	0.185***	−0.003
		(0.007)	(0.026)			(0.023)	(0.065)
.625	356	0.094***	0.071**				
		(0.008)	(0.028)				

Table A.8: Estimation of Autoregressive Models

	Dependent variable:					
	$\ln V_{i't}$					
	(1) TWFE	(2) IV	(3) TWFE	(4) IV	(5) TWFE	(6) IV
$\ln V_{it}$	0.090***	0.052*	0.086***	0.056*	0.086***	0.055*
	(0.007)	(0.025)	(0.007)	(0.024)	(0.007)	(0.022)
$\ln V_{i't-1}$	0.288***	0.298***	0.349***	0.361***	0.355***	0.367***
	(0.003)	(0.007)	(0.003)	(0.010)	(0.003)	(0.009)
$\ln V_{i't-2}$			−0.140***	−0.148***	−0.161***	−0.168***
			(0.002)	(0.006)	(0.002)	(0.005)
$\ln V_{i't-3}$					0.044***	0.043***
					(0.001)	(0.001)
Observations	101,499	101,499	101,498	101,498	101,497	101,497
R^2	0.925	0.923	0.926	0.926	0.927	0.926

Appendix B

Additional Data Processing

B.1 NYT and NOAA 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 NOAA 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.

B.2 Determining the Weather of Cities

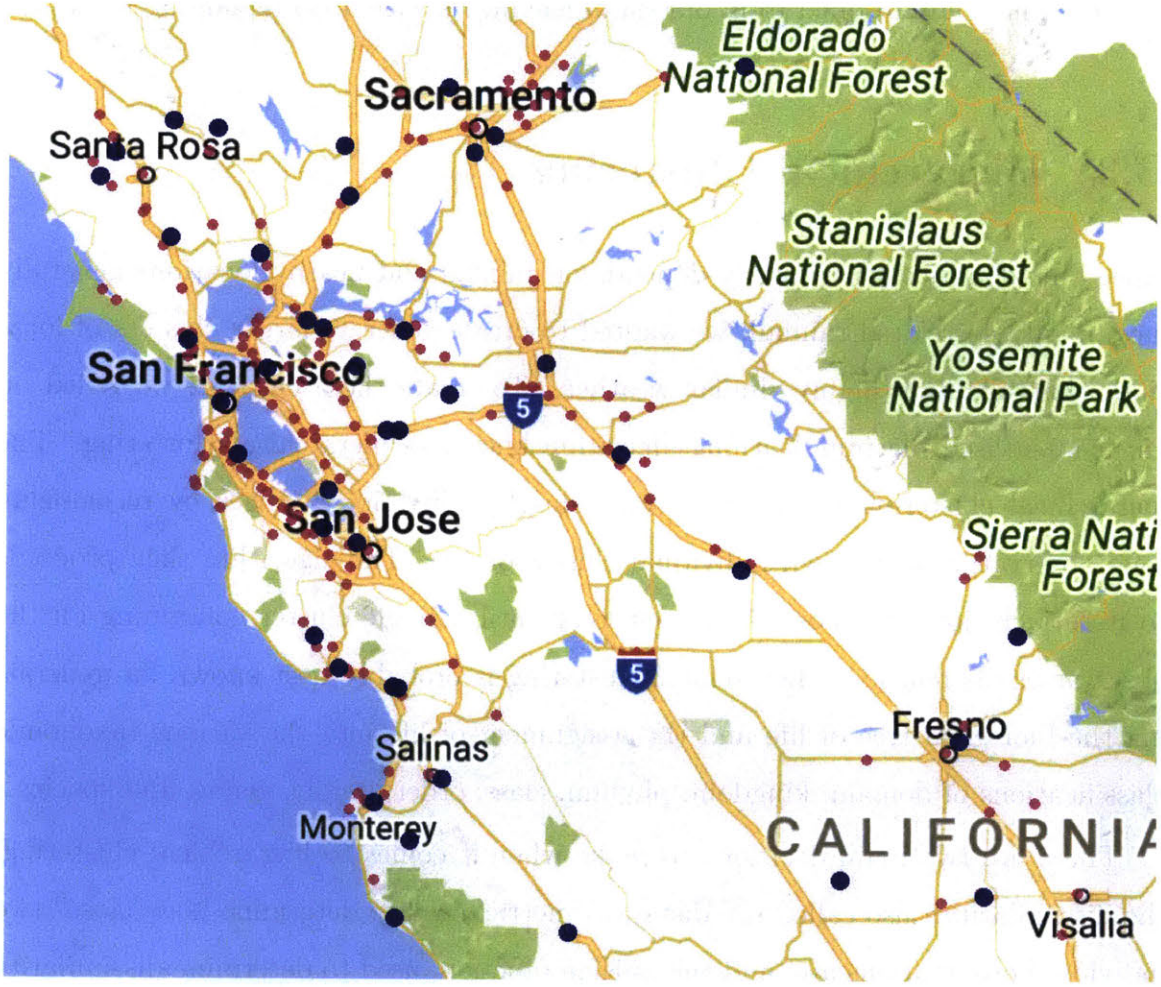
In order for our identification strategy to work, we need to determine the amount of rainfall in each location in our dataset. This turns out to be relatively complicated. For some cases, there's only a single choice for a nearby weather station. However, for many others, there a number nearby weather stations that we can use as potential rainfall measurements. For example, looking at Northern California in figure B-1, 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. 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

Figure B-1: Northern California Cities and Weather Stations



This figure plots the locations of cities in the NYT dataset and weather stations in the NOAA 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.

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.

B.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:

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

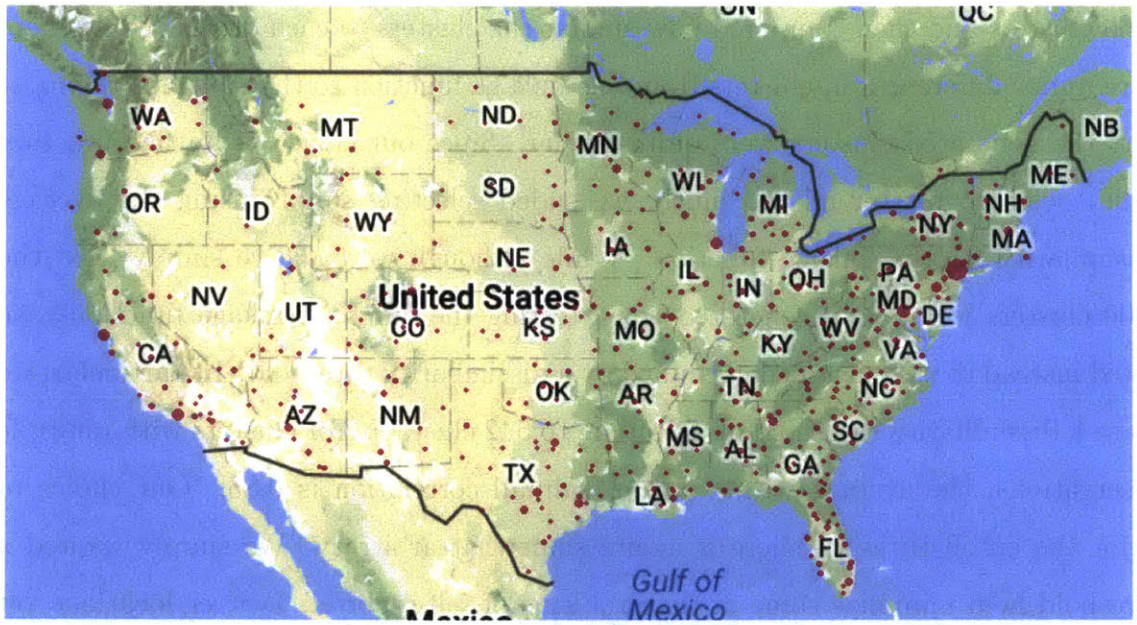
Here, the dissimilarity between two cities i and j is equal to the absolute Pearson correlation in rainfall between i and j ($|\rho_{ij}|$) plus the geographic distance between

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.

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 B-2.

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").

Figure B-2: 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.

B.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 major-

ity 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. Again, the follower self-reported locations have all the problems we described before. 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 multiple 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.

Bibliography

[1] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36, 2017.

[2] Deet T. Allsop, Bryce R. Bassett, and James A. Hoskins. Word-of-mouth research: Principles and applications. *Journal of Advertising Research*, 47(4):398–411, 2007.

[3] Sinan Aral and Nikolaos Hanselmann. To go from big data to big insight, start with a visual. *Harvard Business Review*, 2013.

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

[5] Sinan Aral and Christos Nicolaides. Exercise contagion in a global social network. *Nature Communications*, 8(14753), 2017.

[6] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[7] Susan Athey, Markus M. Mobius, and Jeno Pal. The impact of aggregators on internet news consumption. 2017.

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

[9] Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA, 2011. ACM.

[10] Abhijit Banerjee, Arun G. Chandrasekhar, Esthet Duflo, and Matthew O. Jackson. The diffusion of microfinance. *Science*, 341(6144), 2013.

[11] Michale Barthel. Newspapers fact sheet. Technical report, Pew Research Center, 2017.

[12] Jonah Berger and Katherine L. Milkman. What makes online content viral? *Journal of Marketing Research*, 49(2):192–205, 2012.

[13] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, and James H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415), 2012.

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

[15] Joan Calzada and Ricard Gil. What do news aggregators do? evidence from google news in spain and germany. 01 2016.

[16] Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.

[17] Justin Cheng, Lada Adamic, P. Alex Dow, Jon M. Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 925–936, New York, NY, USA, 2014. ACM.

[18] Lesley Chiou and Catherine Tucker. Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy*, 26(4):782–805, 2017.

[19] Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4):370–379, 2007. PMID: 17652652.

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

[21] Chrysanthos Dellarocas, Juliana Sutanto, Mihai Calin, and Elia Palme. Attention allocation in information-rich environments: The case of news aggregators. *Management Science*, 62(9):2543–2562, 2016.

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

[23] GlobalWebIndex. Daily time spent on social networking by internet users worldwide from 2012 to 2017 (in minutes). Technical report, Statista, 2017.

[24] Sharad Goel, Duncan J. Watts, and Daniel G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, EC '12, pages 623–638, New York, NY, USA, 2012. ACM.

[25] Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[26] Alfred Hermida, Fred Fletcher, Darryl Korell, and Donna Logan. Share, like, recommend. *Journalism Studies*, 13(5-6):815–824, 2012.

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

[28] Matthew O. Jackson and Leeat Yariv. Diffusion of behavior and equilibrium properties in network games. *American Economic Review*, 97(2):92–98, May 2007.

[29] Joseph M. Kayany and Paul Yelsma. Displacement effects of online media in the socio-technical contexts of households. *Journal of Broadcasting & Electronic Media*, 44(2):215–229, 2000.

[30] Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[31] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[32] Ammara Mahmood and Catarina Sismeiro. Will they come and will they stay? online social networks and news consumption on external websites. *Journal of Interactive Marketing*, 37:117 – 132, 2017.

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

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

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

[36] Nic Newman, Richard Fletcher, Antonis Kalogeropoulos, David A. L. Levy, and Rasmus K. Nielsen. Digital news report 2017. Technical report, Reuters Institute, 2017.

[37] NOAA. Noaa national centers for environmental information (ncei) u.s. billion-dollar weather and climate disasters. 2018.

[38] Kristen Purcell, Lee Rainie, Amy Mitchell, Tom Rosenstiel, and Kenneth Olmstead. Understanding the participatory news consumer. 2010.

[39] David Roodman. A note on the theme of too many instruments. *Oxford Bulletin of Economics and statistics*, 71(1):135–158, 2009.

[40] Anaya Sen and Pinar Yildirim. Clicks bias in editorial decisions: How does popularity shape online news coverage? 2015.

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

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

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

[44] Galen Stocking. Digital news fact sheet. Technical report, Pew Research Center, 2017.

[45] Catherine Tucker. Identifying formal and informal influence in technology adoption with network externalities. *Management Science*, 54(12):2024–2038, 2008.

[46] Johan Ugander, Lars Backstrom, Cameron Marlow, and Jon M. Kleinberg. Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109(16):5962–5966, 2012.

[47] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[48] Duncan J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.