

# Essays on Spillover Effects in Digital Media

by

Michael Zhao

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

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

© Massachusetts Institute of Technology 2020. All rights reserved.

Author .....  
Department of Management  
August 7, 2020

Certified by .....  
Sinan Aral  
David Austin Professor of Management  
Professor, Information Technology and Marketing  
Thesis Supervisor

Accepted by .....  
Catherine Tucker  
Sloan Distinguished Professor of Management  
Professor, Marketing  
Faculty Chair, MIT Sloan PhD Program



# Essays on Spillover Effects in Digital Media

by

Michael Zhao

Submitted to the Department of Management  
on August 7, 2020, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Management

## Abstract

In this thesis, I address spillover effects in digital media across 3 different essays.

In Chapter 1, I focus on estimating social spillovers in the consumption of online news. Exploiting regional rainfall as a natural experiment, I find strong statistical evidence of positive social spillovers in consumption of online New York Times content. Specifically, I find that a 1% increase in aggregate viewership outside of a focal region causes viewership in that region to increase by approximately 0.34%. Further analysis shows that these spillover effects are skewed toward the most popular content and are driven by social media sharing rather than search engines or news aggregators.

In Chapter 2, I estimate the spillover effect of paid advertising on organic mobile app installations. I analyze a spend shutoff “experiment” conducted by GameSpace, a major US-based mobile game developer. Using both difference-in-differences (DiD) and regression discontinuity in time (RDiT) approaches, I surprisingly find evidence that paid advertising is boosting organic installs. Moreover, a two-way fixed effects panel regression indicates that every \$100 in spend is associated with approximately 4 organic installs and 34.4 paid installs—estimates that are quantitatively consistent with our RDiT results.

In Chapter 3, I investigate spillover effects of statewide social distancing policies and reopenings during the COVID-19 pandemic on mobility behavior. Specifically, I find that if all alter states implement a shelter-in-place order, an ego county’s mobility decreases by 15-20%. Alter state reopenings have similarly large but opposite effect: once all alter states reopen, ego county mobility increases by 19-32%. These statewide policies also have a major impact on interstate travel: when a destination county is subject to a shelter-in-place order, its out-of-state traffic decreases by 13-18% but only for distant counties. However, once reopened, traffic from both nearby and distant counties increases by 12-13%.

Thesis Supervisor: Sinan Aral

Title: David Austin Professor of Management

Professor, Information Technology and Marketing





## **Acknowledgments**

To my parents, Ying Yu and Jun Zhao, who have given me the encouragement, support, and resources to succeed throughout my life. And to my wife, Zoey Zhou, who has graciously and patiently supported me throughout my long journey to earn my PhD.



# Contents

<b>1</b>	<b>Social Spillovers in Online News Consumption</b>	<b>15</b>
1.1	Introduction . . . . .	15
1.2	Related Literature . . . . .	17
1.3	Data and Empirical Strategy . . . . .	20
1.3.1	Data Sources . . . . .	20
1.3.2	Empirical Strategy . . . . .	21
1.3.3	Regression Specifications . . . . .	24
1.4	Results . . . . .	31
1.4.1	Peer Effects in Online News Viewership . . . . .	32
1.4.2	What is Driving these Peer Effects? . . . . .	40
1.4.3	Robustness Checks . . . . .	43
1.5	Discussion . . . . .	46
1.5.1	Managerial Implications . . . . .	46
1.5.2	Limitations and Extensions . . . . .	47
1.5.3	Conclusion . . . . .	48
1.6	Data Processing Procedures . . . . .	49
1.6.1	NYT and GHCN Data Processing . . . . .	49
1.6.2	Determining the Weather of Cities . . . . .	50
1.6.3	Hierarchical Clustering . . . . .	52
1.6.4	Twitter Data Processing . . . . .	53
1.6.5	Operationalizing Rainfall with a 2-stage Greedy Grid Search . . . . .	56
1.7	Robustness Checks . . . . .	57

1.7.1	Alternate Weather Instrument Specifications . . . . .	57
1.7.2	Placebo Testing the Weather . . . . .	59
1.7.3	Precipitation Correlation Threshold Sensitivity Analysis . . . . .	60
1.7.4	Including Weather Related Content . . . . .	61
1.7.5	Linear-in-Means Specification . . . . .	61
1.7.6	Auto-regressive Models . . . . .	62
1.7.7	Desktop Only Results . . . . .	63
1.7.8	How does the Hierarchical Clustering Threshold Affect Estimates? . . . . .	64
<b>2</b>	<b>Is Paid Advertising Cannibalizing Organic App Installs?</b>	<b>67</b>
2.1	Introduction . . . . .	67
2.2	Related Literature . . . . .	69
2.2.1	Advertising Shutoff Experiments . . . . .	69
2.2.2	Mobile Advertising . . . . .	70
2.3	Data . . . . .	71
2.4	Empirical Strategy and Model Specifications . . . . .	75
2.4.1	Difference-in-difference . . . . .	76
2.4.2	Regression Discontinuity in Time . . . . .	77
2.4.3	Fixed Effects Linear Panel Regression . . . . .	78
2.5	Results . . . . .	79
2.5.1	Difference-in-Difference . . . . .	79
2.5.2	Regression Discontinuity in Time . . . . .	82
2.5.3	Fixed Effect Panel Regressions . . . . .	85
2.6	Discussion . . . . .	86
<b>3</b>	<b>The Interdependence of Regional COVID-19 Reopenings in the United States</b>	<b>89</b>
3.1	Introduction . . . . .	89
3.2	Model Specifications . . . . .	93
3.3	Results . . . . .	94
3.4	Conclusion . . . . .	99
3.5	Materials and Methods . . . . .	100

3.5.1	Data . . . . .	100
3.5.2	Instrumental Variables Model . . . . .	110
3.5.3	Dyad-level Difference-in-differences . . . . .	115
3.5.4	Double Machine Learning Weather Controls . . . . .	117
3.5.5	Software . . . . .	120



# List of Figures

1-1	Northeast Cities . . . . .	23
1-2	Rainfall for Top 100 Regions . . . . .	26
1-3	Region-to-Region Adjacency Matrix . . . . .	30
1-4	Variation in Log Viewership . . . . .	32
1-5	Heterogeneity in Article-Level Peer Effects . . . . .	40
1-6	Northern California Cities and Weather Stations . . . . .	51
1-7	Included Regions in the Continental United States . . . . .	54
1-8	Placebo Histograms . . . . .	60
1-9	Threshold Sensitivity . . . . .	60
2-1	Examples of Mobile App Install Ads . . . . .	72
2-2	Scaled Time Series of Spend, Paid Installs, and Organic Installs . . . . .	74
2-3	Scaled Spend, Days 67 - 105 . . . . .	75
2-4	Parallel Trends . . . . .	80
2-5	Regression Discontinuity in Time . . . . .	83
3-1	Visualizations of Data . . . . .	92
3-2	Ego and Alter State Policy Impact on Mobility . . . . .	95
3-3	Origin and Destination State Policy Impact on Cross-State Mobility . . . . .	97
3-4	Origin and Destination State Policy Interactions . . . . .	98
3-5	County-Level Max Temperature across Four Consecutive Days . . . . .	104
3-6	County-Level Precipitation across Four Consecutive Days . . . . .	104
3-7	DiD Pre-Period Residuals . . . . .	107
3-8	Dyad Pre-Period Residuals . . . . .	116





# List of Tables

1.1	Description of Variables used in Regressions . . . . .	33
1.2	First Stage . . . . .	34
1.3	Main Results: Cross Region Peer Effect . . . . .	36
1.4	Regional Heterogeneity . . . . .	38
1.5	Article-Level Results . . . . .	39
1.6	WOM vs Search . . . . .	41
1.7	How Social Network Connectivity Mediates Peer Effect Strength . . . . .	42
1.8	Alternative Weather IV . . . . .	58
1.9	Alternative Weather First Stage . . . . .	59
1.10	With Weather Related Content . . . . .	62
1.11	Linear-in-Means Specification . . . . .	63
1.12	Estimation of Autoregressive Models . . . . .	63
1.13	Only Desktop Viewership . . . . .	64
1.14	Clustering Cutoff and Estimated Peer Effects . . . . .	65
2.1	Difference-in-Difference Estimates . . . . .	81
2.2	RDIT Estimates . . . . .	82
2.3	Panel Regressions . . . . .	85
3.1	Base Model Results . . . . .	108
3.2	Alters' Policies Model Results . . . . .	109
3.3	IV 2SLS Results . . . . .	113
3.4	IV LIML Results . . . . .	114
3.5	Dyadic Travel Results . . . . .	117

3.6 Dyadic Travel With Interactions . . . . . 118

# Chapter 1

## Social Spillovers in Online News Consumption

### 1.1 Introduction

With the advent of the internet and social media, an important question for content producers is how an individual's demand for content is influenced by her peers'. In this paper, we explore this question in the context of online news. Theoretically, these "peer effects"<sup>1</sup> could potentially either increase or decrease total news content consumption. On one hand, people can share and spread news articles to their peers, thereby broadening the viewership audience. On the other hand, peer consumption may obviate the need for my consumption: if my friend reads a movie review and communicates the information content to me, then I no longer need to read the review myself (or vice versa).

For news organizations, understanding how these social spillovers operate has never been more important. While newspapers face year-on-year declines in circulation and advertising revenue (Barthel 2017), online news consumption has increased dramatically, with 93% of US adults consuming at least some of their news content from an online source (Stocking 2017). If spillovers exist, whether they are positive or negative could have drastically different managerial implications. If positive, then strategies based on maximizing

---

<sup>1</sup>For those more accustomed to Manski (1993)'s terminology, when we say "peer effects" we usually are referring to Manski (1993)'s "endogenous effects."

digital viewership and ad revenue may be viable. If negative, then it may make more sense for content producers to focus on a narrower customer base and charge them for content.

Despite their importance, there is relatively little work that cleanly identifies social spillovers in content demand. Although it is well established that peer behaviors tend to be highly correlated, it is difficult to disentangle the effects of peer influence from confounding factors—an issue often referred to as “the reflection problem” (Manski 1993). We overcome this challenge by utilizing a hierarchical clustering algorithm to construct “local” geographic regions. These regions are constructed in such a manner that allow us to leverage local rainfall as a source of exogenous variation, thereby enabling an instrumental variables approach that can credibly identify cross-region spillover effects.

Applying this approach to the online viewership at the New York Times (NYT), we find statistically significant evidence of positive spillover effects. Specifically, our results indicate that a 1% increase in outside-region (or external) viewership generates approximately 0.34% additional within-region (or regional) viewership. We also find this spillover effect is significantly stronger for higher viewership regions. In a similar vein, an article-level analysis shows that the higher the viewership of the content, the more positive the spillover effect, suggesting that online content exhibits increasing returns to scale in demand (also known as network effects).

Additionally, we investigate the mechanism that drives these spillover effects. We analyze the NYT’s webpage referrer data and find stronger cross-region spillovers on traffic referred from social media sources (Facebook and Twitter) than for traffic referred from search engines or news aggregators (Google, Yahoo, and Bing). Expanding on this, we build the region-to-region social media follower graph, parsed from Twitter user data, and find that the peer effect between regions that are “strongly-connected” on Twitter is significantly greater than the peer effect between regions that are “weakly-connected” on Twitter. These results suggest that social media sharing drives online news consumption.

Overall, there are several contributions of our work. First, we are, to the best of our knowledge, the first empirical work to identify positive spillovers in demand for information goods like news content. Second, we uncover evidence that as viewership of a piece of content increases, the spillover effects become stronger. Third, our results indicate that

social media sharing, rather than search engine or news aggregator rankings, are driving these spillovers. Hence, our work provides the first clean causal evidence, albeit indirect, demonstrating social media’s positive impact on news consumption. Lastly, our empirical methodology can be replicated by content producers to determine if their content exhibits spillover effects, without the need for social network data.

## 1.2 Related Literature

Understanding peer effects has been a major topic of interest for social scientists at least since the seminal “Coleman Report” (Coleman 1968), which (of its many findings) found that an individual’s peers had an especially significant impact on her individual academic achievement. As a result, the report played a major role in shaping public policy in the years following its publication, with many citing its findings to support social integration.

Despite the academic interest, measurement of peer effects has proven to be quite challenging. The standard analytical approach is to use a “linear-in-means” model, where individual outcomes are regressed on the average of peer outcomes. However, the resulting estimate may be confounded by factors such as homophily (McPherson et al. 2001) or exposure to the same external stimuli. For instance, consider that both me and my friend are reading the news today. One explanation is that I caused my friend’s reading by sharing a news article with her (or vice versa). Another possibility is that we share many similar interests (which is why we are friends) that induce both of us to read the news. Lastly, there could be some kind of event that is driving everyone to read the news.

Disentangling peer effects from these other explanations is critically important and has significant real-world policy implications. In particular, peer effects can exhibit a “social multiplier” effect that can magnify the effect of an intervention. For example, Zhang and Zhu (2011) found that after Chinese Wikipedia blocked a significant number of contributors, non-blocked contributors decreased their contributions by 42.8%. Alternatively, consider a company trying to drive adoption of a product. If positive social spillovers exist, then seeding a few influential individuals within a social network could potentially generate a large cascade of product adoptions. In contrast, if homophily is primarily explaining

clustering in product adoptions, then running an ad campaign targeting individuals with similar characteristics to adopters is likely a better strategy.

Although Manski (1993) notes that identification of peer effects is generically confounded in observational data, recent studies have employed a number of approaches to address some of the challenges. Many employ large-scale randomized field experiments (Aral and Walker 2012, Bond et al. 2012, Banerjee et al. 2013, Kramer et al. 2014). Others have developed new observational methods ranging from high-dimensional matching (Aral et al. 2009), structural modeling (Ghose and Han 2011), and instrumental variables approaches (Bramoullé et al. 2009, Tucker 2008, Coviello et al. 2014a, Gilchrist and Sands 2016, Aral and Nicolaides 2017a). In particular, our approach is similar to those used by Coviello et al. (2014a) and Aral and Nicolaides (2017a), who both use exogenous variation in the weather to identify emotional and exercise contagion respectively. Like them, we take advantage of a key insight: while the weather may affect my behavior, it shouldn't directly affect the behavior of a friend who lives in a different area. However, the key difference with our approach is that it is designed to work without the need for social network data<sup>2</sup>. This is not trivial, as researchers or companies cannot necessarily get access to social network data. Even if such data is available, it may not be possible to join social network data with individual level outcomes.

Typically, the closest related work to ours are Moretti (2011) and Gilchrist and Sands (2016), who both study social spillovers in the context of movie consumption. In the case of Moretti (2011), the results suggest that the spillovers are the result of “social learning,” where consumers use their peers to gather information about a good with uncertain quality. On the other hand, Gilchrist and Sands (2016) instead contend that the spillovers are instead driven by a preference for a shared experience, though they cannot rule out some effect of learning. While both these papers uncover evidence of positive peer effects, it is unclear whether these findings can generalize to the consumption of online news. Unlike movies, most of the utility of news article lies with the information contained within them. Though people may have preferences for shared information, much like preferences for a

---

<sup>2</sup>Though we do make use of Twitter data to construct something of a proxy network, our main empirical analysis does not make use any network data.

shared experiences, if individuals are able to access the information in an article without reading it, then there is little further incentive to actually consume the content (“information redundancy”).

We also explore how digital technologies are mediating these social spillovers. Prior work by George and Waldfogel (2003) has found evidence of “preference externalities” in the consumption of traditional news: racial groups were more likely to buy a daily newspaper in regions where their racial group had greater population. However, in the context of digital news, such preference externalities are less relevant since digital news are not restricted by the limited space of a physical newspaper. Rather, the issue now is that there is too much content readily available. Hence, technologies like news aggregators and social media which allow consumers to more easily find content that suited to them are particularly relevant. In the case of news aggregators, prior work is typically framed in terms of determining whether news aggregators are substitutes or complements for news consumption. Dellarocas et al. (2016) find that displaying more information about articles decreases the probability that readers will click through to the full article. Though this result seems to suggest substitution due to an information redundancy effect<sup>3</sup>, other empirical research on this topic indicates that news aggregators on average increase news consumption by allowing content to reach a broader audience. Calzada and Gil (2016) and Athey et al. (2017), both studying a Google News shutdown in Spain, independently find that the shutdown caused news website visits to drop. A related paper by Chiou and Tucker (2017)—which exploits a dispute between Google News and the Associated Press—also finds evidence to support complementarity.

Research on how social media impacts news consumption also takes a complements vs substitutes framing. Anecdotal evidence from Aral (2013), which visualizes the Twitter cascades against the real time web traffic of 3 different NYT articles, seems to indicate the potential for both. More formal academic work has generally found a positive correlation between social media and news consumption. An early study by Hong (2012) looks at Twitter adoption and subsequent online traffic to 337 newspapers from January 2007 to December 2010. Using a monthly panel, he found a significant positive association between

---

<sup>3</sup>This has also been called the “scanning effect” by (Chiou and Tucker 2017)

Twitter adoption and the number of unique visitors to newspapers’ websites even after including newspaper fixed effects and non-parametrically controlling for time trends. More recent work by Sismeiro and Mahmood (2018) examines the effect of a major Facebook outage on hourly web traffic to a major European news website. They find that “online social networks are *positively associated* with news reading” (emphasis theirs’), as both viewership and number of unique visitors decreased both during and immediately after the shutdown. However, in direct contrast to this result, a recent news article (Schwartz 2018) showed that during a 45 minute Facebook outage, traffic to news websites increased overall by 2.3%, suggesting substitution was at work<sup>4</sup>. Our results inform this ongoing debate by providing, to the best of our knowledge, the first (albeit indirect) causal evidence that social media, on average, boosts news consumption.

## 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 consists of more than 2 billion individual events tracked on the NYT’s internal servers from April 3, 2013 to October 31, 2013.<sup>5</sup> The raw data is very rich and granular consisting of millisecond-accurate timestamps, IP-address derived geolocation data, accessed URL, referrer URL, among many other fields. Since we are primarily concerned with content consumption, we limit ourselves to direct content pageviews, totalling over 200 million in the US alone. 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 Global Historical Climatology Network (GHCN) database maintained by the National Oceanic and Atmospheric Association (NOAA) described in Menne et al. (2012) and Twitter user and follower pro-

---

<sup>4</sup>“What happens when Facebook goes down? People read the news” (<http://www.niemanlab.org/2018/10/what-happens-when-facebook-goes-down-people-read-the-news/>)

<sup>5</sup>Some prominent examples of tracked events include content pageviews, frontpage visits, searches, account settings changes, and even 404 “Page Not Found” Errors.



file data<sup>6</sup>. The GHCN data contains daily observations of maximum temperature, minimum temperature and precipitation for some 45 thousand weather stations around the world (of which approximately 30 thousand are located in the continental United States). We use the geographic coordinates of each weather station to derive precipitation data for each region in our dataset. Our Twitter data is built by parsing a sample of 10000 “ordinary”<sup>7</sup> Twitter accounts with a tweet or retweet containing a link to a NYT article during our time frame<sup>8</sup>. As mentioned earlier, there are a number of concerns with the quality and precision of the self-reported locations in users’ profiles. Hence, we are only able to obtain adequate coverage for the top 100 regions. We randomly sampled 100 accounts from each of these 100 regions and then obtained the self-reported locations (if available) in the profiles of all of the followers of those 10000 accounts to construct a region-to-region social media connectivity graph. A more comprehensive description of all of our data processing procedures is provided in Appendix 1.6.

### 1.3.2 Empirical Strategy

As mentioned earlier, it is difficult to identify causal peer effects in observational data. Of principal concern is the simultaneous equation bias inherent to estimating peer effects, which arises since an individual’s outcomes are dependent on peer outcomes and vice versa. For our context in particular, we also need to be worried about the unobserved confounding effect resulting from day-to-day variation “inherent newsworthiness”. During major events like presidential elections, terrorist attacks, or natural disasters, people will naturally be inclined to increase their news consumption. In fact, news websites saw increased traffic due to the coronavirus pandemic.

Work by Brock and Durlauf (2001) and Durlauf and Tanaka (2008) prove that instrumental variables can address these issues. Valid instruments need to satisfy two main restrictions: exclusion and relevancy. In practice, however, it is often difficult to find IVs that

---

<sup>6</sup>Parsed using Tweepy, “An easy-to-use Python library for accessing the Twitter API” (<http://www.tweepy.org/>).

<sup>7</sup>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.

<sup>8</sup>These tweets and retweets were also provided to us by the NYT.

satisfy both these requirements. Excluding correlated patterns, weather variation across regions is generally exogenous, thereby satisfying the exclusion restriction. Moreover, weather can have a large impact on behavior, potentially satisfying the relevancy restriction.

The central idea is that when it is raining outside, people are more likely to stay indoors and spend time on the Internet reading the news. For example, Sen and Yildirim (2015)—who use weather as an instrument to study whether editorial decisions about news content take viewership into consideration—find that viewership of articles is 5%-8% higher on rainy days for a large Indian news provider. In our own analysis (Table 1.3), we find that “light” rainfall increases regional viewership by about 2% while “heavy” rainfall increases viewership by about 4%<sup>9</sup>. To clarify our approach, consider the simplified 2-person case:

$$V_{1t} = \alpha_1 + \beta V_{2t} + \gamma R_{1t} + \epsilon_{1t}$$

$$V_{2t} = \alpha_2 + \beta V_{1t} + \gamma R_{2t} + \epsilon_{2t}$$

In the first equation, viewership of person 1 on date  $t$  ( $V_{1t}$ ) depends on viewership of person 2 on date  $t$  ( $V_{2t}$ ) and rainfall in region 1 on date  $t$  ( $R_{1t}$ ). Similarly, in the second equation, viewership of person 2 on date  $t$  depends on viewership of person 1 on date  $t$  and rainfall in region 2 on date  $t$  ( $R_{2t}$ ). As long as  $R_{1t}$  and  $R_{2t}$  are (conditionally) independent of  $\epsilon_{2t}$  and  $\epsilon_{1t}$  respectively, we can use  $R_{2t}$  as an instrument for  $V_{2t}$  in the first equation and  $R_{1t}$  as an instrument for  $V_{1t}$  in the second equation to identify  $\beta$ .

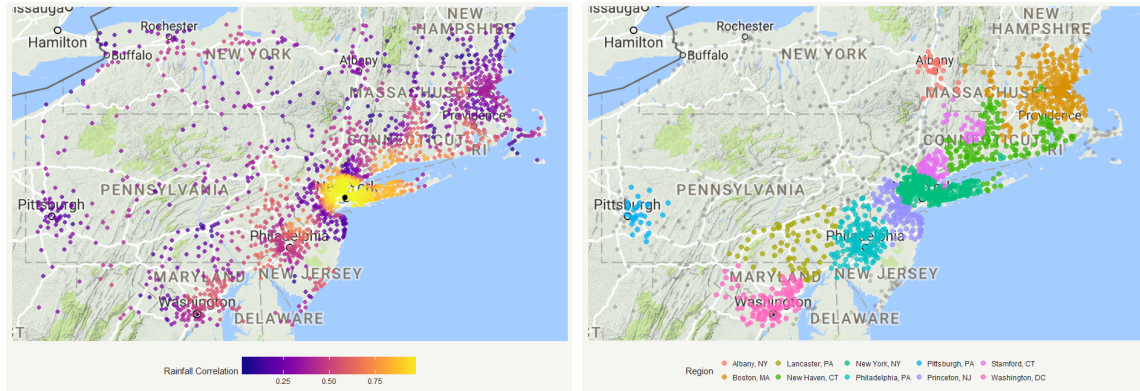
When expanding this to general linear-in-means model, the main endogenous variable of interest, average peer outcomes, is instrumented for by the average of peer weather as done by both Coviello et al. (2014a) and Aral and Nicolaides (2017a). In our case, although we have access to individual level data in the NYT data<sup>10</sup>, we do not know how the individuals are connected to each other. Though the social graphs of some social networks like Twitter are publicly available (to a degree), it is intractable to join the NYT user data to the social network data.

---

<sup>9</sup>It is worth noting that the estimated coefficients are highly statistically significant and remarkable consistent across most of our results.

<sup>10</sup>Either through cookies, unique device identifiers, or NYT user accounts.

Figure 1-1: 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 these cities with New York City (displayed as the black point). Highly correlated cities are displayed as orange or yellow while more uncorrelated cities are 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).

We instead aggregate our data to level of “localized” regions such that within-region weather is largely homogeneous, somewhat similar to the aggregation approach taken by Coviello et al. (2014a). In our case, this is accomplished using an unsupervised machine learning algorithm called hierarchical clustering (see Appendix 1.6.3 for the full details behind this procedure). However, since we do not have precise measures of the connectivity between regions, we instead use the summed viewership rather than mean viewership of other regions as our main endogenous variable (the exact model specification is described more formally in section 1.3.3 below).

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. To some degree, our hierarchical clustering procedure mitigates this concern since cities with similar weather will likely end up in the same region, especially if they are geographically close. Looking at Figure 1-1 and focusing on New York City, we see that cities with highly correlated rainfall are assigned to the same region as NYC. Moreover, as long as we include within-region rainfall in our regressions, outside-region rainfall should theoretically be conditionally independent. This however assumes that we have fully accounted

for rainfall’s effects in the model specification. Hence, as a further precaution, we further restrict our analysis to regions with sufficiently uncorrelated rainfall<sup>11</sup>.

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<sup>12</sup> from our analysis. Even if we miss some of the articles, weather-related content represents such a small fraction of NYT articles and pageviews<sup>13</sup> that potential bias will be extremely small. In fact, the five most viewed weather-related articles during our time period were about Hurricane Sandy, an event that occurred in 2012, which would certainly be uncorrelated with 2013 weather.

### 1.3.3 Regression Specifications

#### Main Specification

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

$$\ln V_{it} = \beta \ln V_{-it} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.1)$$

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

---

<sup>11</sup>We define “sufficiently uncorrelated” when the absolute correlation coefficient in rainfall between regions to be less than 0.25.

<sup>12</sup>According to the NYT’s internal content tagging system, accounts for 0.02% of pageviews

<sup>13</sup>Especially since 2013 was a rather mild year in terms of weather related disasters. According to NOAA (2018) only 8 major weather-related disasters occurred during 2013 and only 6 during our timeframe.

We operationalize regional weather by converting our continuous precipitation measure into two binary variables  $L_{it}$  and  $H_{it}$ , indicating when precipitation is less than 0.22mm and exceeds 16.86mm respectively. We will refer to precipitation less than or equal to 0.22mm as “no rain,” (which is the omitted category) precipitation exceeding 0.22mm but less than or equal to 16.86mm as “light rain,” and precipitation exceeding 16.86mm as “heavy rain.” These values were determined using a 2-stage greedy grid search aimed to find the thresholds that have the greatest explanatory power on regional viewership. The exact details of this procedure can be found in Appendix 1.6.5. We opt for discrete indicators rather than the continuous measure since human behavior regarding the weather tends to be non-linear. For instance, the difference in behavior between 0mm and 1mm of rain is going to be significantly larger than the difference between 20mm and 21mm of rain. Given the way  $L_{it}$  and  $H_{it}$  are coded,  $\gamma_1$  is interpreted as the average increase in regional viewership when it is lightly raining in that region. On the other hand,  $\gamma_2$  captures the additional marginal effect when it is heavily raining in a region<sup>14</sup> An illustration of our discretized rain measure for the top 100 regions in our data can be found in Figure 1-2.

Although we do not rely on fixed effects for identification, we include a set of region ( $\alpha_i$ ) and time fixed effects ( $\tau_t$ <sup>15</sup>) to account for some of the unobserved variance. In our context, region fixed effects control for regional heterogeneity arising from differences in factors such as population demographics, political leaning, etc<sup>16</sup>. On the other hand, time-fixed effects help account for seasonal trends and even some degree of inherent newsworthiness. Lastly,  $\epsilon_{it}$  denotes the error term.

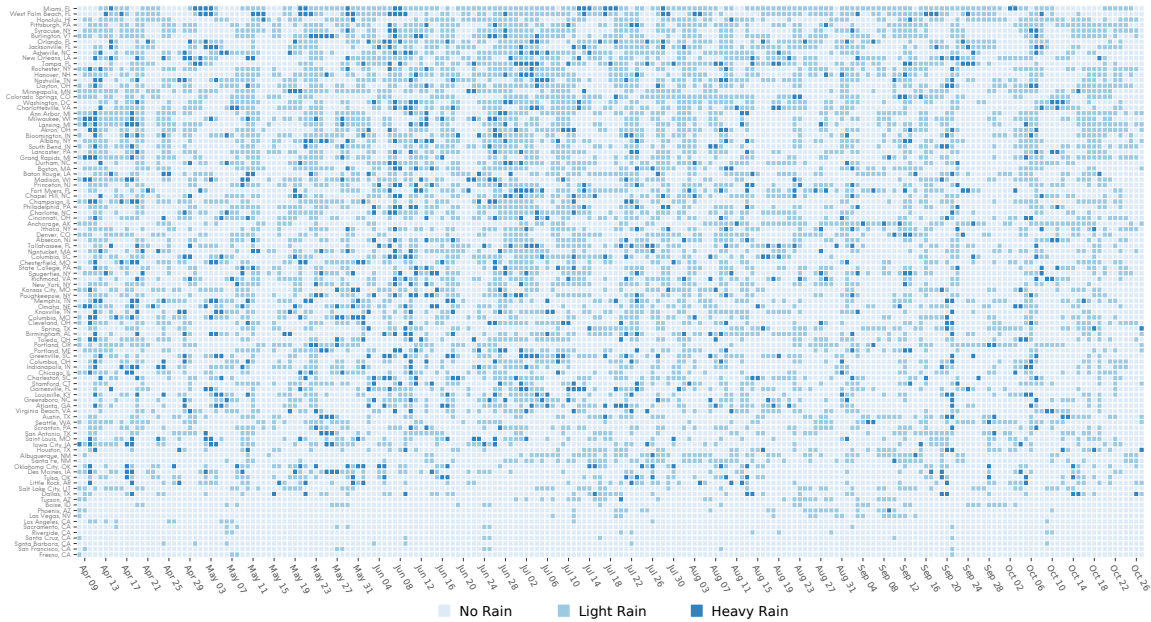
---

<sup>14</sup>It is worth noting that  $\gamma_1$  and  $\gamma_2$  are theoretically identified, if not by the exogeneity of rainfall, then under difference-in-difference like assumptions.

<sup>15</sup>For this study, we operationalize our time fixed effects here as a combination of week and day-of-week fixed effects. Using date-level fixed effects introduces a peculiar type of mechanical bias. To clarify, consider that  $V_{it} + V_{-it} = V_t$  (Not exactly true, since some regions are excluded from  $V_{-it}$ , but quite close). This then implies that  $V_{it} = V_t - V_{-it}$ . Including date-fixed effects, which absorb  $V_t$ , into regression of  $V_{it}$  on  $V_{-it}$  thereby mathematically induces a negative coefficient estimate (this is what we mean by mechanical bias). One way to practically think about this is a person only can read the news from one place at a time. Hence if someone reads an NYT article in NYC, it simultaneously implies that she is not reading that article outside of NYC.

<sup>16</sup>While population characteristics are certainly changing over the course of our study, the amount of change that can occur over the 7 month period of our study are going to be quite minor

Figure 1-2: Rainfall for Top 100 Regions



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

## IV First-Stage

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

$$\ln V_{-it} = \zeta_1 L_{-it} + \zeta_2 H_{-it} + \eta_1 L_{it} + \eta_2 H_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \quad (1.2)$$

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

$$L_{-it} = \frac{1}{|U_i|} \sum_{j \in U_i} L_{jt}$$
$$H_{-it} = \frac{1}{|U_i|} \sum_{j \in U_i} H_{jt}$$

where  $|U_i|$  denotes the cardinality or number of regions in  $U_i$ . Since  $L_{-it}$  is the fraction of regions experiencing light rainfall, we can interpret its associated parameter  $\zeta_1$  as the (approximate) percent in viewership if no regions in  $U_i$  are raining vs. if all regions in  $U_i$  are lightly raining. Accordingly,  $H_{-it}$  is the fraction of regions experiencing heavy rainfall, and given its coding, its associated parameter  $\zeta_2$  can be interpreted as the (approximate) percent change in viewership when comparing light rain in all regions in  $U_i$  relative to heavy rain in those same regions.  $L_{it}$  and  $H_{it}$  are included as control covariates since they show up in the main specification above. If our exclusion restriction assumption is correct, then we should expect the associated parameters  $\eta_1$  and  $\eta_2$  to be essentially 0.

Naturally, we also include both sets of fixed effects in this specification as well. While the time fixed effects  $\tau_t$  behave similarly as they do in equation 1.1 above, the region fixed effects  $\alpha_{-i}$  work a little bit differently. In this case, they absorb the aggregated population demographics, political leaning, etc. of all regions  $j \in U_i$ . Programmatically however, this is equivalent to including a fixed effect for region  $i$  since  $U_i$  is determined by  $i$ . Finally,  $\nu_{-it}$  is the associated error for our first stage specification.

## Regional Heterogeneity

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

$$\ln V_{it} = \beta^{top}(\ln V_{-it})^{top} + \beta^{mid}(\ln V_{-it})^{mid} + \beta^{bot}(\ln V_{-it})^{bot} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.3)$$

Here, we modify our main model specification by breaking up our original main dependent variable into 3 distinct parts:  $(\ln V_{-it})^{top}$ ,  $(\ln V_{-it})^{mid}$ , and  $(\ln V_{-it})^{bot}$ . These three terms simply represent the interaction of  $\ln V_{-it}$  with  $top_i$ ,  $mid_i$ , and  $bot_i$ —binary variables indicating if a region  $i$  belongs to the top 167, middle 166, or bottom 167 regions in the regional viewership distribution in our timeframe. Here, we can interpret  $\beta^{top}$ ,  $\beta^{mid}$ , and  $\beta^{bot}$  as the average cross-region viewership effect on the top, middle, and bottom regions respectively. Given the way these variables are coded, the parameter estimates are interpreted as the conditional average effects for the top, middle, and bottom regions respectively, rather than as an additive marginal effect assuming one group as the baseline group.

## Article Level Analysis

To dive deeper, we also examine spillover effects on an article level. However, since viewership is very heavy tailed across both articles and regions, this creates a significant amount of sparsity in the full article-region-date dataset, which makes econometric estimation difficult. To address this, we restrict our analysis to the first-day viewership of the top 1000 most viewed articles in our time period. We use the following model specification:

$$\ln V_{ijt} = \beta \ln V_{-ijt} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \theta_t + \epsilon_{ijt} \quad (1.4)$$

This essentially is our main specification adapted to an article level. Here the main dependent variable  $\ln V_{ijt}$  is simply the log viewership of article  $j$  in region  $i$  on date  $t$ . Similarly,  $\ln V_{-ijt}$  is the log combined viewership of article  $j$  in regions outside of  $i$  with sufficiently



uncorrelated weather on date  $t$ . Everything else is as it was before with the exception that our previous time-fixed effects  $\tau_t$  are replaced by proper date-fixed effects  $\theta_t$ .

### Viewership Referrals

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

$$\ln V_{it}^{sm} = \beta^{sm} \ln V_{-it} + \gamma_1^{sm} L_{it} + \gamma_2^{sm} H_{it} + \alpha_i^{sm} + \tau_t^{sm} + \epsilon_{it}^{sm} \quad (1.5)$$

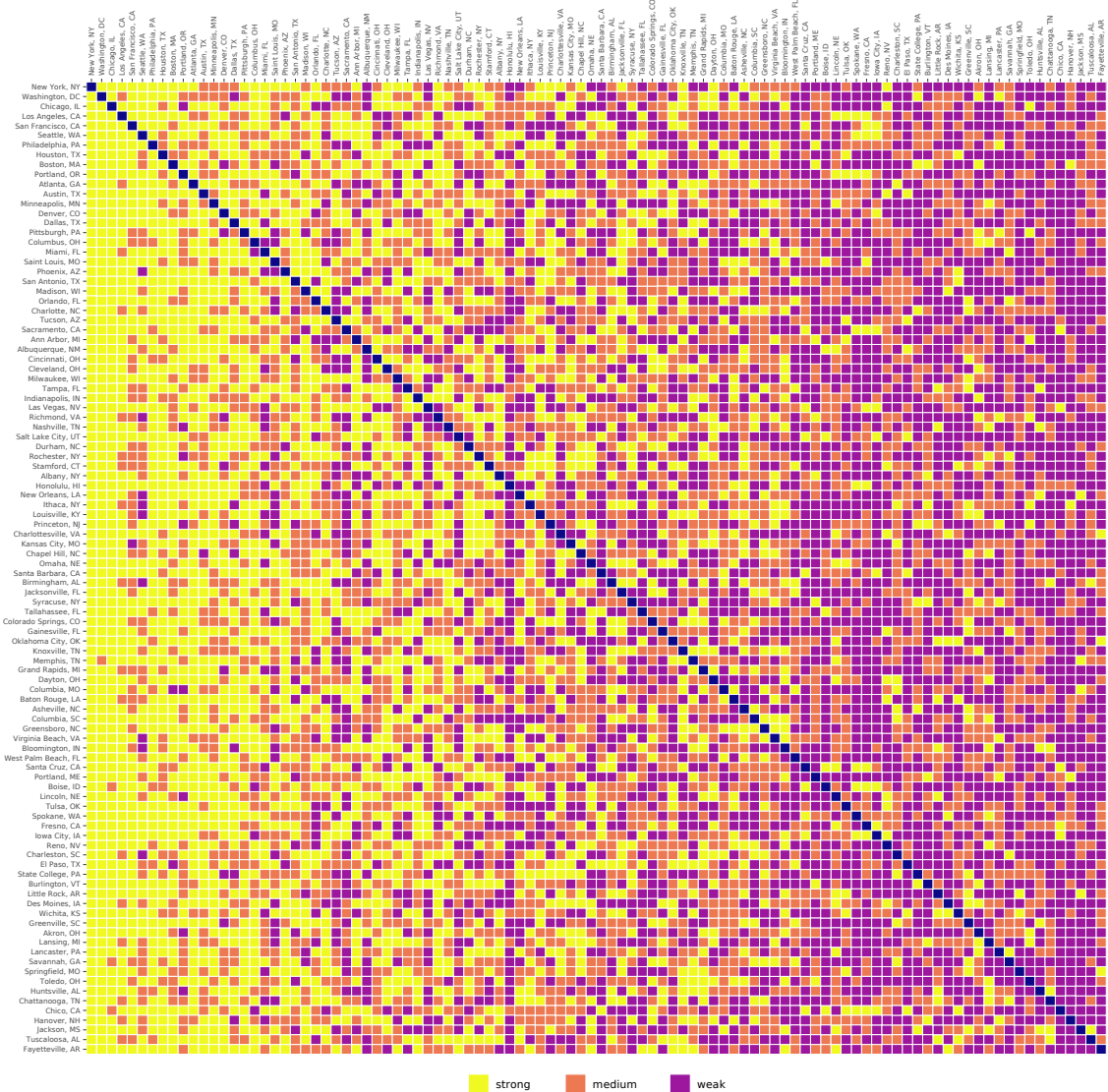
$$\ln V_{it}^{se} = \beta^{se} \ln V_{-it} + \gamma_1^{se} L_{it} + \gamma_2^{se} H_{it} + \alpha_i^{se} + \tau_t^{se} + \epsilon_{it}^{se} \quad (1.6)$$

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

### Twitter Connectivity

To further explore whether social media sharing is really driving these peer effects, we examine how social network connectivity mediates the strength of the peer effect. Using our Twitter follower data, we generate the region-to-region follower graph of the top 100 regions. For each region  $i$ , we classify each of other 99  $j$  regions as “strongly-connected,” “moderately-connected,” and “weakly-connected” regions based on tie density. Strongly-connected regions are the top third of regions with the highest number of

Figure 1-3: Region-to-Region Adjacency Matrix



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

follower-follower<sup>17</sup> ties, moderately-connected regions are the middle third, and weakly-connected are the lowest third. We plot this follower graph as an adjacency matrix in Figure 1-3.

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

$$\ln V_{it} = \beta^{sc} \ln V_{-it}^{sc} + \beta^{mc} \ln V_{-it}^{mc} + \beta^{wc} \ln V_{-it}^{wc} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.7)$$

This specification modifies our main specification by replacing the primary independent variable  $\ln V_{-it}$  with three new independent variables:  $\ln V_{-it}^{sc}$ ,  $\ln V_{-it}^{mc}$ , and  $\ln V_{-it}^{wc}$ .  $\ln V_{-it}^{sc}$  denotes the log sum viewership of strongly-connected regions with sufficiently uncorrelated weather to  $i$ .  $\ln V_{-it}^{mc}$  and  $\ln V_{-it}^{wc}$  are the same, except for moderately- and weakly-connected regions. Naturally,  $\alpha_i$ ,  $\tau_t$ , and  $\epsilon_{it}$  are still as they were before. If social media sharing does play a major role in driving the cross-region viewership peer effect, then we should expect  $\beta^{sc} > \beta^{mc} > \beta^{wc}$ . One potential concern with this specification is that our parameters may be capturing homophily in the region-to-region Twitter connectivity network rather than true heterogeneity in the strength of the spillover effect. However, from an econometric perspective, this type of homophily is not uniquely different from the latent homophily that confounds peer effects estimations generally. As such, as long as variation in rainfall is not correlated with Twitter connectivity, this issue should not impact our estimation.

## 1.4 Results

We begin this section with some basic descriptive statistics and visualizations of our data. We report the demeaned values for our viewership variables for data privacy reasons in Table 1.1 below. Our final primary dataset consists of 101,500 observations (500 regions  $\times$  203 dates) and accounts for over 200 million pageviews. This full dataset is used to estimate equations 1.1, 1.3, 1.5, and 1.6. Equation 1.4 is estimated on a region-article aggregated dataset that contains 500,000 (500 regions  $\times$  1000 articles) observations. Lastly, we use a

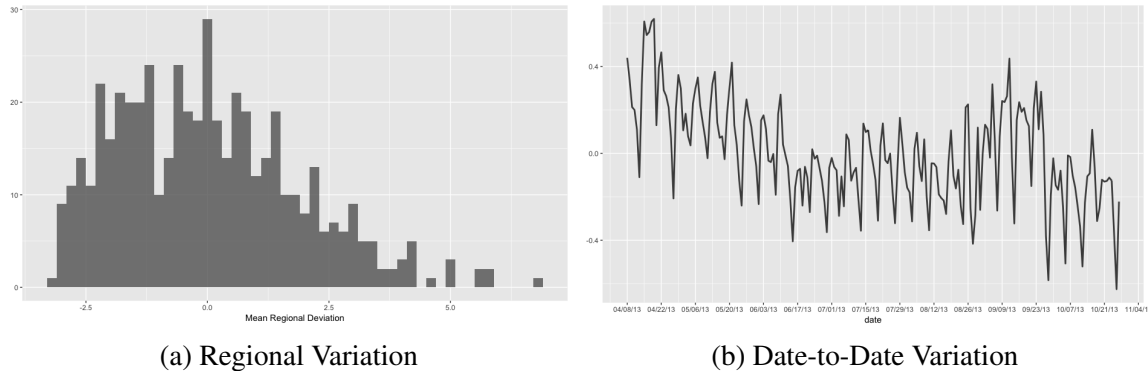
---

<sup>17</sup>The follower is the user in  $i$  who follows a user in  $j$ .

smaller 20,300 observation dataset (100 regions  $\times$  203 dates) to estimate equation 1.7, due to the difficulty in joining self-reported Twitter profile locations to our constructed regions.

We also include a histogram of the mean regional deviation in log viewership and a plot of day-to-day shocks in regional viewership (see Figure 1-4) to illustrate the variation in  $\log V_{it}$  from region to region and from day to day.

Figure 1-4: Variation in Log Viewership



Note: These are demeaned values.

### 1.4.1 Peer Effects in Online News Viewership

This section is organized as follows. We begin by validating our instruments and the first-stage regression of our IV model in section 1.4.1. We then describe the estimation and analysis of our main regression Equation 1.1 in section 1.4.1. Next, we explore the heterogeneity of the peer effects relative to regional size in section 1.4.1. Lastly, we provide some article level results in section 1.4.1. For all results, region and date clustered robust standard errors are reported as there may be both region-level and date-level error correlation.

#### Instrument Validity and First-Stage Estimates

Even before examining the first stage itself, we first look at the coefficient estimates on  $L_{it}$  and  $H_{it}$  in Table 1.3. Looking specifically at columns 3 (TWFE) and 4 (IV), we can see that the models that incorporate both region and time fixed effects produce highly statistically significant light rainfall coefficients of around 0.02. Hence, if it is lightly raining in a particular region, then that region’s viewership increases by approximately 2%. As for heavy

Table 1.1: Description of Variables used in Regressions

Variable	Description	Mean	St. dev.	Min	Max
<b>Main Specification Variables (1.3.3)</b>					
$\ln V_{it}$	Demeaned log views of NYT content in region $i$ on date $t$	0.000	1.905	-5.896	7.350
$\ln V_{-it}$	Demeaned log views of NYT content in regions with uncorrelated rainfall on date $t$	0.000	0.265	-1.222	0.698
$L_{it}$	Indicator for light rainfall (precipitation exceeding 0.22mm) in region $i$ on date $t$	0.355	0.479	0	1
$H_{it}$	Indicator for heavy rainfall (precipitation exceeding 16.86mm) in region $i$ on date $t$	0.052	0.222	0	1
<b>First Stage Variables (1.3.3)</b>					
$L_{-it}$	Weighted average of light rainfall indicators in regions $U_i$ on date $t$	0.354	0.093	0.084	0.577
$H_{-it}$	Weighted average of heavy rainfall indicators in regions $U_i$ on date $t$	0.052	0.033	0.000	0.206
<b>Regional Heterogeneity Specification Variables (1.3.3)</b>					
$toR_i$	Indicator for whether region $i$ is one the top 167 regions in total viewership	0.334	0.472	0	1
$mid_i$	Indicator for whether region $i$ is one the middle 166 regions in total viewership	0.332	0.471	0	1
$bot_i$	Indicator for whether region $i$ is one the bottom 167 regions in total viewership	0.334	0.472	0	1
<b>Article-Level Specification Variables (1.3.3)</b>					
$\ln V_{ijt}$	Demeaned log views of article $j$ in region $i$ on date $t$	0.000	1.877	-2.168	10.158
$\ln V_{-ijt}$	Demeaned log views of article $j$ in regions with uncorrelated rainfall on date $t$	0.000	1.202	-6.838	3.972
<b>Search vs Social Variables (1.3.3)</b>					
$\ln V_{it}^{sm}$	Demeaned log aggregate views of NYT content referred from Facebook and Twitter in regions $i$ on date $t$	0.000	1.882	-3.330	7.246
$\ln V_{it}^{se}$	Demeaned log aggregate views of NYT content referred from Google, Yahoo, and Bing region $i$ on date $t$	0.000	1.829	-4.321	7.062
<b>Twitter Connectivity Variables (1.3.3)</b>					
$\ln V_{-it}^{sc}$	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
$\ln V_{-it}^{mc}$	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
$\ln V_{-it}^{wc}$	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

rainfall, both TWFE and IV again produce similar statistically significant coefficients also coincidentally around 0.02. As mentioned earlier, due to the coding of  $L_{it}$  and  $H_{it}$ , the coefficient on  $H_{it}$  is interpreted as an additional marginal effect. Hence, if it is raining heavily in region  $i$ , then we should expect region  $i$ 's viewership to increase by around 4%. Looking at column 2, not including time-fixed effects seems to slightly depress the coefficient estimates of both rainfall variables, but they still remain highly statistically significant. Most importantly however, these results confirm the basic intuition of the instrumental variables approach since they indicate that rainfall has statistically and practically significant effects on a region's online news consumption. Next, we directly check the validity of the aggregated weather instruments themselves  $L_{-it}$  and  $H_{-it}$ . Table 1.2 presents the estimation of the first stage (eq. 1.2).

Table 1.2: First Stage

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

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

Looking at this table, we see that the estimated coefficients on  $L_{-it}$  and  $H_{-it}$  are both positive and highly statistically significant. Taking the parameter estimates as given how-

ever, we should observe that aggregated viewership is approximately 6.5% higher when it was lightly raining in all regions in  $U_i$  compared to when there was no rain in the same regions. Again the coefficient on  $H_{-it}$  is an additive marginal effect. Thus, these estimates suggest that when it is raining heavily in all regions in  $U_i$  aggregated viewership is approximately 34.8% higher than when there is no rain.

While these numbers seem extreme, especially compared to the region-level rainfall effects, this is likely because it is neither completely raining nor completely dry in every single region simultaneously. Also,  $L_{-it}$  and  $H_{-it}$  are aggregated measures averaged across many regions. Thus the variance of these covariates is naturally much smaller than in single-region counterparts  $L_{it}$  and  $H_{it}$ , which is likely to generate more extreme coefficient estimates<sup>18</sup>. Overall, we think the point values of these coefficients shouldn't be taken too literally.

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

## Peer Effects Estimates

Table 1.3 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 effects or two-way fixed effects (TWFE) in column 3, and lastly, instrument variables (IV) in column 4.

If we think that variation in the “newsworthiness” of day-to-day events is the main driver of news consumption, then given our log-log specification, we might expect non-IV

---

<sup>18</sup>Recall that a regression coefficient in simple linear regression is given by  $\frac{\text{Cov}(Y,X)}{\text{Var}(X)}$ . In our case, the variance of  $L_{-it}$  is 26 times greater than  $L_{it}$  while the variance of  $H_{-it}$  is 46 times greater than  $H_{it}$ .

Table 1.3: Main Results: Cross Region Peer Effect

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

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

methods to produce a correlational estimate of approximately 1<sup>19</sup>. Looking at the pooled OLS result in column 1 of table 1.3, we see that this is not case. Instead, we find a statistically insignificant coefficient estimate of -0.210. This negative point estimate is largely the result of the generic negative correlation between  $\ln V_{it}$  and  $\ln V_{-it}$ <sup>20</sup>. Once we include fixed-effects, the estimated peer effect jumps quite dramatically to a highly statistically significant 0.931 and 1.050 as seen in columns 2 and 3, much closer to our a priori expectation of 1.

These non-IV results suggest that panel methods are indeed insufficient to address the simultaneous equation bias present in our context. In contrast, IV produces an estimate of 0.343 as can be seen in column 4. Taking this point estimate as given, this suggests that a

<sup>19</sup>Suppose that a terrorist attack occurs tomorrow and that doubles news consumption throughout the entire US. In this case, both external viewership and regional viewership would increase by 100% leading to a peer effect estimate of 1.

<sup>20</sup>If region  $i$  is New York City, the highest viewership region in the data, then this automatically implies that NYC is not contributing to  $\ln V_{-it}$ . In the same vein, if  $i$  is a very minor region, then the high viewership regions will most likely be counted in  $\ln V_{-it}$



1% increase in the total viewership of external regions (with uncorrelated weather) should increase regional viewership by approximately 0.34%. To put it a little more practically: if all other people in regions with uncorrelated weather to you doubled their news consumption, your news consumption would increase by about 34%. Though this coefficient may seem large, we think that it is quite credible. The counterfactual that is being captured in the coefficient is quite extreme, as the viewership of all other regions needs to increase. To put this estimate in a more realistic context, suppose that it is lightly raining 20% of other regions. This increases outside region viewership by 0.012 (based on the first stage estimate) which then boosts regional viewership less than half of a percent. This result almost certainly confirms that regional viewership does indeed have positive spillover effects, generating additional viewership across regions. While the direction of this result agrees with the panel models, the magnitude of the coefficient is substantially different, confirming our concerns about potential endogeneity bias.

### **Regional Heterogeneity**

We next look into how the strength of this peer effect might be heterogenous across regions. Table 1.4 presents the estimated results of equation 1.3 using TWFE in column (1) and IV in column (2).

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

The IV results in column (2) support this result. The point estimate on the top regions

Table 1.4: Regional Heterogeneity

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

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

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

These results suggest that cross-region viewership peer effects have a stronger effect on higher viewership regions. Political homophily might be a factor that explains these results. Since the NYT is generally considered a “center-left” publication, the spillover effect from viewership to smaller regions that tend to be more politically conservative is likely to be smaller. Another possible factor is that the lower viewership regions tend to be demographically older and less likely to obtain news content from the internet compared

with the high viewership regions, which would also work to suppress the magnitude of the estimated peer effect.

### Article-Level Results

Though our results provide strong evidence of social spillovers in news consumption, it is important to more precisely understand how these spillovers are operating. Are people causing their peers to generally read more news? Or are peers reading specific pieces of content that were recommended to them? Table 1.5 presents the results from estimating equation 1.3.3 using both TWFE and IV on the top 1000 most read articles in our dataset.

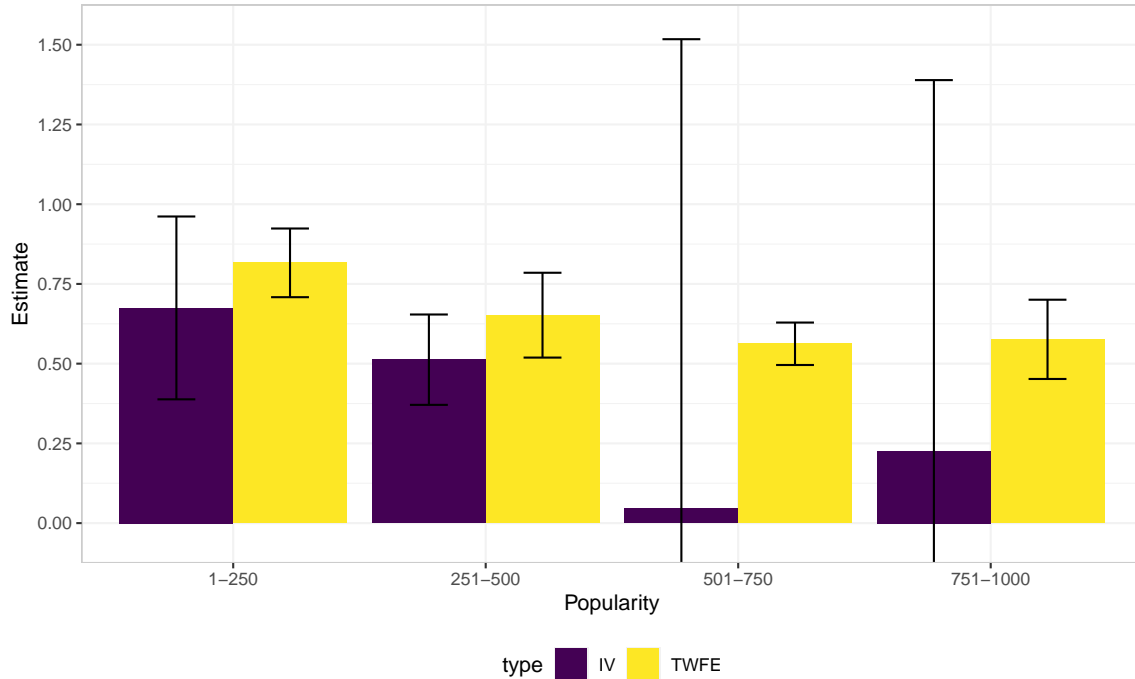
Table 1.5: Article-Level Results

	<i>Dependent variable:</i>	
	$\ln V_{ijt}$	
	(1) TWFE	(2) IV
$\ln V_{-ijt}$	0.705*** (0.040)	0.464** (0.153)
$L_{it}$	0.023*** (0.005)	0.022** (0.007)
$H_{it}$	0.013 (0.008)	0.014 (0.009)
Observations	500,000	500,000
R <sup>2</sup>	0.888	0.861

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

The results in this table mirror the aggregated results. Though the coefficient on  $H_{it}$  is no longer statistically significant, the point estimates are relatively close to those estimated above. However, it is particularly noteworthy that the estimated IV coefficient of the peer effect is actually more positive than in Table 1.3. This coupled with the fact that this analysis is conducted on the subsample of the 1000 most viewed articles seems to support the idea that spillovers in news content consumption occur more at the individual article

Figure 1-5: Heterogeneity in Article-Level Peer Effects



This plot visualizes the coefficient estimates of  $V_{ijt}$ , which denotes the article-level cross-region viewership spillover effect for the top 250, 251-500, 501-750 and 751-1000th most viewed articles using both IV (purple) and TWFE (yellow). 95% Confidence intervals so each of the estimates is plotted as the error bars around each bar.

level than at a general news-reading level. This is further exemplified in Figure 1-5, which plots the estimated peer effect coefficients for the top 250, 251-500, 501-750 and 751-1000th most viewed articles separately. Focusing on the IV estimates (in purple), we see that the estimated spillover effect is especially strong for the top 250 and (to a somewhat lesser degree) 251-500th most viewed articles. Beyond providing additional support for article-level spillovers hypothesis, these results suggest that social spillovers are stronger for more popular content.

### 1.4.2 What is Driving these Peer Effects?

Our results point to positive and significant cross-region peer effects in news consumption. However, it remains an open question whether social media, news aggregators, or some other mechanism is responsible for these peer effects. We investigate this first by comparing

the estimated peer effect on viewership referred from social media sources to viewership referred from search engines and news aggregators. We also examine the degree to which region-to-region Twitter connectivity mediates the strength of the peer effect.

### Viewership Referrals: Social vs Search

Table 1.6 presents the estimation of the equations found in section 1.3.3. Specifically, Columns (1) and (2) report the the estimation of equation 1.5 using TWFE and IV. Similarly, columns (3) and (4) the estimation of equation 1.6, again using IV and TWFE respectively.

Table 1.6: WOM vs Search

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

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

Some clear trends emerge from these results. Both TWFE and IV models show substantially stronger estimated coefficients on content referred from social media than content referred from search engines. In particular, the IV results are quite striking. Looking at column 4, we see that the estimated effect of  $\ln V_{-it}$  on  $\ln V_{it}^{se}$  is -.055. Though this point estimate itself is negative, it is not statistically significantly different from 0. In retrospect, this result may not be too surprising. Although rainfall does have a positive and significant

on viewership, it is still rather unlikely for a small boost in viewership in a number of regions, even if they are the large regions, to dramatically change search or news aggregator ranking results by a significant amount. On the other hand, the estimated effect of  $\ln V_{-it}$  on  $\ln V_{it}^{sm}$  is positive and highly statistically significant at 0.763, meaning that the cross-region effect on content referred from social media is quite strong. We feel these results overall provide powerful corroborating evidence that social media sharing is responsible for the cross region peer effects.

### Social Network Connectivity and Peer Effect Strength

Table 1.7 presents the results of estimating Equation 1.7. We continue to report results using both TWFE and IV in columns (1) and (2) respectively. Looking at the TWFE estimates

Table 1.7: How Social Network Connectivity Mediates Peer Effect Strength

	<i>Dependent variable:</i>	
	$\ln V_{it}$	
	(1) TWFE	(2) IV
$\ln V_{-it}^{sc}$	0.996*** (0.025)	0.669*** (0.066)
$\ln V_{-it}^{mc}$	0.981*** (0.032)	0.652*** (0.096)
$\ln V_{-it}^{wc}$	0.973*** (0.035)	0.403** (0.098)
$L_{it}$	0.020*** (0.003)	0.021*** (0.003)
$H_{it}$	0.021*** (0.006)	0.024*** (0.005)
Observations	20,300	20,300
R <sup>2</sup>	0.986	0.985

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

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

### 1.4.3 Robustness Checks

We perform a series of checks to verify the robustness of our main results. We begin by testing 4 different operationalizations of our rainfall instrument: linear precipitation, quadratic precipitation, linear precipitation with a light rain indicator, and quadratic precipitation with a light rain indicator. We find that all these various operationalizations of rainfall produce extremely strong first-stage F-statistics. More importantly, IV estimates produced by these different operationalizations are both qualitatively and quantitatively similar to estimate seen in column (4) of table 1.3. The exact model specifications and regression results of these alternative operationalizations can all be found in Appendix 1.7.1.

Next, we perform a falsification or placebo test on our weather instrument to check for instrument validity. We accomplish this by randomizing the order of our rainfall instruments since the weather from a random day should not be affecting viewership on a particular date  $t$ . We repeat this process 1000 times to generate a distribution of placebo results. Once randomized, not once out of the 1000 randomizations did the rainfall instruments exceed the strong instrument criteria determined by Stock and Yogo (2002). Accordingly,

the IV estimates produced by these randomized rainfall instruments were only statistically significant at the 5% level in 10 out of 1000 trials (potentially indicating that our standard errors are quite conservative). Moreover, none of the placebo estimates ever even approach the level of significance of real rainfall instruments. Given these results, it is quite unlikely that our estimated results in table 1.3 are driven by luck. Histograms of the first-stage Wald F-statistic and the p-values of the IV estimate can be found in Appendix 1.7.2.

Furthermore, we look into the sensitivity of results regarding the choice of threshold used to exclude regions from  $V_{-it}$ . We re-estimate our main model specification across an entire range of thresholds, from 0.05 to 0.5 (in 0.05 increments) and find the results to be both qualitatively and quantitatively consistent with those presented in table 1.3. These results can be find in Appendix 1.7.3.

As mentioned in section 1.3.3, we adapted the standard linear-in-means model to better suit our setting. As a robustness check, we estimate the cross-region peer effect using the traditional formulation of the linear-in-means model using both TWFE and IV. These estimates are qualitatively similar to the results we presented earlier in the paper. Here, we continue to find evidence of endogeneity bias in the TWFE estimates since they substantially differ from the IV estimates. Similarly, the IV estimates are positive and highly statistically significant implying positive spillovers in viewership. Quantitatively however, the point estimates themselves are significantly smaller than those estimated using equation 1.1. This is likely driven by the large degree of regional heterogeneity and the functional form assumptions of the linear-in-means model. NYC's spillover effect is much stronger relative to those of smaller regions. But, due to the assumptions of the specification, the contribution of NYC is averaged with the contributions of many more smaller regions, driving the estimated coefficient lower since the regional viewership distribution is very heavy tailed. The exact specification and regression results can be found in Appendix 1.7.5.

We further investigate if time-series dependencies are potentially biasing or driving our results. For instance, ongoing stories may be causing autocorrelation in viewership. Alternatively, there are temporal correlations in weather that may causing an exclusion restriction violation. To address these concerns, we estimated a set of model specifications that included lagged terms of the main dependent variable, regional weather, and outside-



region weather. We found that including these lagged terms produced cross-region spillover estimates qualitatively and quantitatively similar to our results above. All the AR model specifications and regression results can be found in Appendix [1.7.6](#).

We also look into concerns that our results may suffer from attenuation bias due to measurement error in our IP-derived geolocation data. Theoretically, our model specification is robust to this type of bias, since our endogenous variable of interest  $V_{-it}$  covers nearly the entire United States, and therefore only requires IP-geolocations to be accurate at that level. Furthermore, IV approaches are also generally considered as a standard solution to attenuation bias. We check this empirically by running our model only on desktop-based geolocation data which tends to be more reliable than mobile geolocation data. This analysis produced qualitatively and quantitatively similar estimates to those presented above and can be found in Appendix [1.7.7](#).

We also recheck all of our main results using different cutoff thresholds for our hierarchical clustering algorithm. Due to the nature of algorithm, a lower threshold will lead to a greater number of smaller clusters, where similarity within each cluster is higher. We find that the choice of threshold does not significantly impact our results qualitatively or quantitatively. These results can be found in in Appendix [1.7.8](#).

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

## 1.5 Discussion

### 1.5.1 Managerial Implications

Our work has broad implications for news organizations and content producers more generally. Though news viewership among peers is naturally correlated, our work provides causal evidence of positive social spillovers in online news consumption. Establishing causality is key, as our results indicate that even for information goods, peer consumption generally does not substitute individual consumption. As such, news content producers organized around maximizing viewership and monetizing through ads may indeed be viable.

Moreover, our work suggests that social spillovers are stronger for more popular content. While generically, it is quite difficult to predict in advance what pieces of content are likely to garner high viewership, there are at some obvious exceptions to the rule. For instance, news viewership will always be high around major events like terrorist attacks, global pandemics, or presidential elections. In light of our findings, it may be prudent for news organizations to try and encourage as much viewership as possible during such events by temporarily increasing their paywall limit or removing it entirely. Alternatively, they could invest in increasing exposure during such periods by paying for more advertisements.

In addition, our work finds that social media, as opposed to search engines or news aggregators, are a more significant driver of the spillover effects. To that end, interventions aimed at increasing social media sharing may prove profitable. For instance, content producers can adopt sharing encouragement interventions by more actively asking their consumers to share their content or even potentially offering incentives to share. On the other hand, the interventions like search engine or news aggregator optimization, may not be as effective to help your content “go viral.”

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

of our approach is that it can provide causal estimates only requiring geolocated data from the content producer itself.

## **1.5.2 Limitations and Extensions**

Though our research finds credible causal evidence of positive social spillovers in news consumption, there are some limitations to our work. First, there remains much work to be done to further understand the dynamics of these spillover effects. To start, the aggregate nature of analysis does not provide much insight to individual level peer effects, which potentially are key to informing optimal seeding policies. In addition, there are many dynamic substitution and complementarity effects that we do not consider. For instance, certain types of news content may be highly substitutable, if multiple articles all provide similar coverage about a single event. In this case, reading one article about the event might reduce the likelihood of reading other articles about the same event. Alternatively, other types of news might be more complementary like reviews or opinion pieces which generally have a wider variety of viewpoints. Moreover, there may also be similarly complex interactions with radio and cable news as well.

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

Lastly, while our work provides strong causal evidence of a positive effect of social media on news viewership, it is indirect. While we can say with some confidence that social media increases news consumption, our analysis does not provide much information about the magnitude of such effects. Regarding our social media and search engine referral analysis specifically, we don't account for potentially important cross-channel complementary or substitution effects. Beyond this, there is also still much to be understood about the role of different social media platforms or different types of word-of-mouth. Facebook, Twitter, Reddit, etc. all make different platform design choices. It would be interesting to understand what those platform-level differences imply about online content and business strategies designed to maximize viewership. Overall, there is still much that needs to be understood about the relationship between social media and online content.

### **1.5.3 Conclusion**

Publishers' strategies in the digital age depend critically on social spillover effects in content consumption. Unfortunately, obtaining causal estimates of such effects is quite difficult due to the "reflection problem." We overcome this challenge by leveraging exogenous variation in regional weather to identify positive and significant cross-region social spillovers in online news consumption. Additionally, our work suggests that this effect increases with the popularity of the content. We also find that social media sharing is the primary driver of these cross-region peer effects. Specifically, cross-region peer effects in viewership are stronger when referred by social media than when referred by search engines. Furthermore, social network connectivity mediates the strength of these peer effects: regions that are strongly-connected on social media exhibit more positive and significant peer effects than regions that are weakly-connected on social media. We also find these effects are more pronounced for more populous regions. Taken together, these results suggest that social media sharing drives online news consumption. Our results and the methods used to estimate them can help content producers make better decisions about how and when to incentivize social media use to drive viewership, revenues, and profit.

## 1.6 Data Processing Procedures

### 1.6.1 NYT and GHCN Data Processing

We are very careful in handling the NYT data since it potentially contains fairly sensitive information. When parsing the dataset, we are careful to avoid any personal identifying information. The only fields that we look at are time of access, the url accessed, type of content, the derived geolocation, and the referrer URL. Given the geographic nature of our identification strategy, any events where geolocation data is not sufficiently precise ("city-level") or missing are excluded from the analysis. In addition, 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 these cities, we use the Google Maps API to obtain geographic coordinate data.

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

## 1.6.2 Determining the Weather of Cities

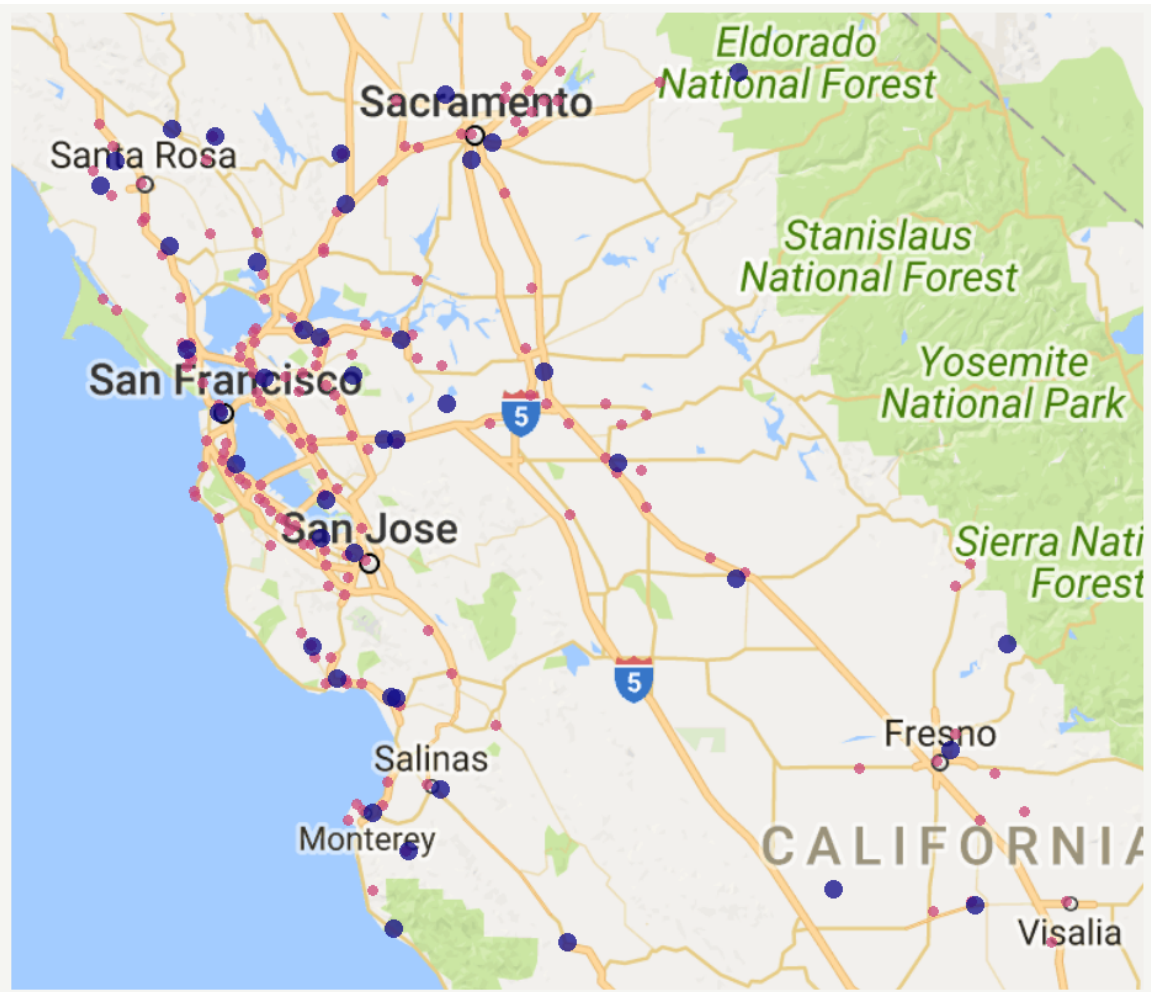
In order for our identification strategy to work, we need to determine the amount of rainfall in each location in our dataset. This turns out to be relatively complicated. For some cases, there's only a single choice for a nearby weather station. However, for many others, there are a number of nearby weather stations that we can use as potential rainfall measurements. For example, looking at Northern California in figure 1-6, 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 considerably 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 weather stations within 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.

Figure 1-6: Northern California Cities and Weather Stations



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

### 1.6.3 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:

$$D_{ij} = (1 - |\rho_{ij}|) + G_{ij}/500$$

Here, the dissimilarity between two cities  $i$  and  $j$  is equal to one minus the absolute Pearson correlation in rainfall between  $i$  and  $j$  ( $1 - |\rho_{ij}|$ ) 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 differ-



ences 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 1-7.

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").

## 1.6.4 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

Figure 1-7: 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.

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 out converted all non-ASCII text. We then used some basic regular expressions to further normalize the location text. After this, we ran the cleaned text through the Google Maps API to recover geographic coordinates and location “type”. Google Maps has several different location classifications, for example, a “sublocality” generally refers official sub-city areas like the borough of Brooklyn, a “locality” generally denotes a city or town, an “administrative\_area\_level\_2” indicates a county, and an “administrative\_area\_level\_1” designates a state (in the United States). There are also some unofficial types of importance, namely, “colloquial\_area” which refers to the area of land that might make up a large metropolitan area like the SF Bay Area, the Tri-State Area, or Greater Los Angeles. We keep users with self-reported locations that have Google Maps types of “neighborhood”, “sublocality”, “locality” and “colloquial\_area”. We specifically discard self-reported locations that has the Google Maps type of “route”, even though “routes” are even more granular than localities. The problem with routes is the vast majority of “routes” are actually self-reported locations like “Cloud 9”—which the Google Maps API will return a result for “Cloud 9 Inn”—or “Nowhere”—which may be some local bar. Moreover, people generally don’t self-report their own location with that degree of specificity (for the most part, the highest degree of specificity of commonly reported is at the level of “Williamsburg, Brooklyn” or “Midtown Manhattan” which both count as neighborhoods).

Using the geographic coordinates, we determine the closest city in the NYT dataset to each of the self-reported user locations. This way we can get region assignments for each of the included self-reported locations. In total, these self-reported locations are assigned to 174 of our regions. We then check the the ratio between between a region’s total viewership and tweets and retweets and exclude regions with ratios in top and bottom 5% of the ratio distribution. We further restrict our analysis to the top 100 remaining regions. Since

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

Using Tweepy, we access the Twitter API to obtain the followers (over 200000) of these 10000 users. Again making use of the Tweepy, we obtain the self-reported locations of the followers. Naturally, the follower self-reported locations have all the problems we described above. Hence, we use the same procedure as above to determine which region these follower accounts belong too. We then use this information to build the region-to-region directed network. Each directed edge  $e_{ij}$  in this network represents the the number of accounts in  $j$  that follow accounts in city  $i$ . Naturally, to account for the stratified sampling approach, we multiply the number of follower links appropriately (for example, if region  $i$  has 1000 accounts, we would then multiply  $e_{ij}$  by 10). Lastly, for each region  $i$  we classify the remaining 99 regions  $j$  as into 33 “strongly”, 33 “moderately”, 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 1-3 in the main paper.

### 1.6.5 Operationalizing Rainfall with a 2-stage Greedy Grid Search

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

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

$$\ln V_{it} = \gamma_1 L_{it} + \alpha_i + \tau_t + \epsilon_{it}$$

where  $L_{it} = 1(R_{it} > r_1)$  against the “restricted model” of:

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

for all possible values of  $r_1$  starting from 0 going to 50 incremented by 0.01. We selected the value of  $r_1$  that maximized the F-statistic of our test. In the second stage stage of our approach, we performed an F-test comparing the updated “full model”:

$$\ln V_{it} = \gamma_1 L_{it} + \gamma_2 H_{it}, \alpha_i + \tau_t + \epsilon_{it}$$

where  $H_{it} = 1(R_{it} > r_2)$  against the updated “restricted model” of:

$$\ln V_{it} = \gamma_1 L_{it} + \alpha_i + \tau_t + \epsilon_{it}$$

for all possible values of  $r_2$  starting from 0 going to 50 incremented by 0.01. Again, we selected the value of  $r_2$  that maximized the F-statistic of our test.

## 1.7 Robustness Checks

### 1.7.1 Alternate Weather Instrument Specifications

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

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 R_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.8)$$

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 R_{it} + \phi_2 R_{it}^2 + \alpha_i + \tau_t + \epsilon_{it} \quad (1.9)$$

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 R_{it} + \gamma_1 L_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.10)$$

$$\ln V_{it} = \beta \ln V_{-it} + \phi_1 R_{it} + \phi_2 R_{it}^2 + \gamma_1 L_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.11)$$

Equation 1.10 (linear precipitation) replaces the rainfall indicators with a linear term for precipitation in region  $i$  on date  $t$  ( $R_{it}$ ). Equation 1.9 (quadratic precipitation) simply adds

an additional quadratic precipitation term ( $R_{it}^2$ ). Equation 1.10 (linear precipitation and rainfall indicator) combines both the linear precipitation term with binary indicator for rainfall. Lastly, equation 1.11 (quadratic precipitation and rainfall indicator) simply adds a quadratic precipitation term to the prior equation. The coefficient estimates of these specifications are presented in table 1.8 where columns (1) through (4) refer to equations 1.8 through 1.11 respectively.

Table 1.8: Alternative Weather IV

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

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

Naturally, these different rainfall operationalizations imply different first stages as well. The modified first stage specifications as presented below:

$$\ln V_{-it} = \psi_1 R_{-it} + \omega_1 R_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \quad (1.12)$$

$$\ln V_{-it} = \psi_1 R_{-it} + \psi_2 R_{-it}^2 + \omega_1 R_{it} + \omega_2 R_{it}^2 + \alpha_{-i} + \tau_t + \nu_{-it} \quad (1.13)$$

$$\ln V_{-it} = \psi_1 R_{-it} + \psi_3 R_{it} + \zeta_1 L_{-it} + \eta_1 L_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \quad (1.14)$$

$$\ln V_{-it} = \psi_1 R_{-it} + \psi_2 R_{-it}^2 + \omega_1 R_{it} + \omega_2 R_{it}^2 + \zeta_1 L_{-it} + \eta_1 L_{it} + \alpha_{-i} + \tau_t + \nu_{-it} \quad (1.15)$$

The first stage regression estimates can be found in table 1.9 where columns (1) through (4) refer to equations 1.12 through 1.15 respectively.

Table 1.9: Alternative Weather First Stage

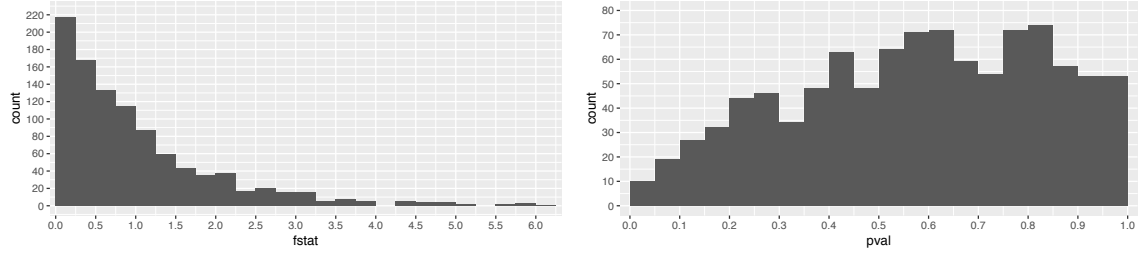
	<i>Dependent variable:</i>			
	$\ln V_{-it}$			
	(1) TWFE	(2) IV	(3) TWFE	(4) IV
$R_{-it}$	0.008*** (0.00004)	-0.010*** (0.0002)	0.006*** (0.0001)	-0.022*** (0.0003)
$R^2_{-it}$		0.002*** (0.00003)		0.003*** (0.00003)
$L_{-it}$			0.046*** (0.001)	0.120*** (0.001)
$R_{it}$	0.0001** (0.00004)	0.0002***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

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

## 1.7.2 Placebo Testing the Weather

We generate a placebo draw of our data by just randomizing the order of the rainfall indicators. We are careful to ensure that the pair of indicators are not broken up. We use this procedure to generate 1000 placebo draws of our data. We estimate equation 1.1 using IV for each one of these draws. We report a histogram of the estimated first stage Wald F-statistic using these 1000 placebo draws in figure 1-8(a). We also report a histogram of

Figure 1-8: Placebo Histograms



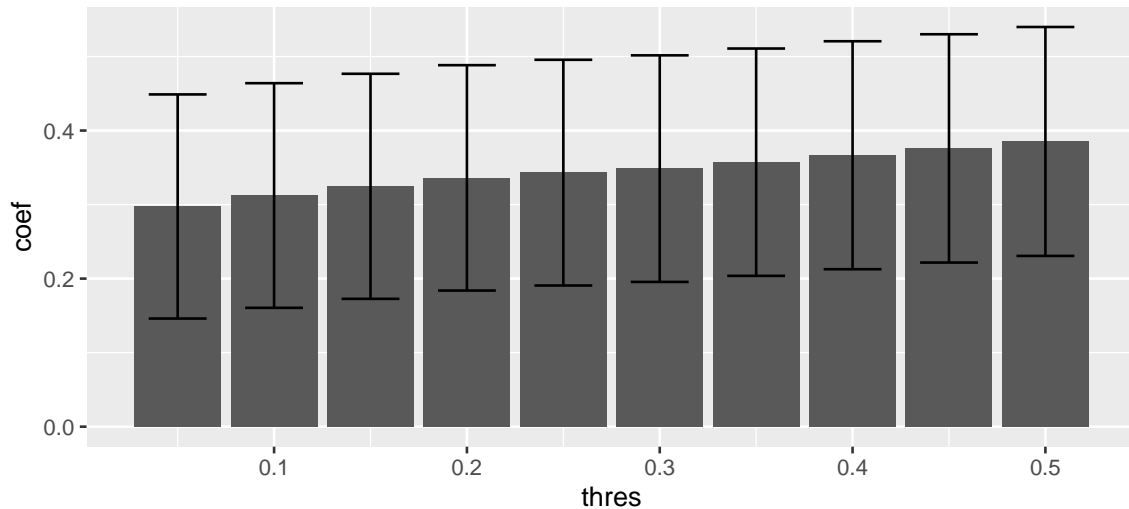
(a) Placebo First Stage Wald F-statistic Histogram    (b) Placebo IV Estimate P-value Histogram

the p-value IV coefficient estimate of  $\ln V_{-it}$  of these draws in figure 1-8(b).

### 1.7.3 Precipitation Correlation Threshold Sensitivity Analysis

Recall that in the main text, we report results where regions with correlated precipitation coefficient of 0.25 are excluded from  $V_{-it}$ ,  $L_{-it}$ , and  $H_{-it}$ . In this section, we examine the robustness of our results to different choices of this threshold. We plot the only the IV estimated coefficients on the spillover effect for thresholds ranging from 0.05 to 0.5 in 0.05 increments in figure 1-9.

Figure 1-9: Threshold Sensitivity



This figure the plots the IV estimated coefficient on  $V_{-it}$  for weather thresholds from 0.05 to .5 incremented by 0.05. The 95% confidence interenal computed using region and date clustered standard errors is also shown for each estimate.



Here we see that our results are robust across a wide range of threshold choices as the estimates generated from different thresholds are all quite close quantitatively. We do note that there is extremely consistent slight positive trend in the threshold. However, this trend should not be surprising: as the threshold increases, more regions are included into  $V_{-it}$ . Moreover, the newly added regions are likely to be geographically closer and therefore more densely connected relative to regions with less weather correlation. As such, the total amount of influence in  $V_{-it}$  is increasing as the threshold increases, leading to more positive coefficient estimates.

### 1.7.4 Including Weather Related Content

In the main text, we present results where we excluded content with weather-related content tags. Below we present the results from estimating our main model specification Equation 1.1 without excluding these views from our analysis. Results from using TWFE and IV are presented in Table 1.10. As can be seen in the table, the estimated coefficients are all almost identical to the ones presented in Table 1.3. As such, we believe that there is very little bias induced by the some content potentially being based on the weather.

### 1.7.5 Linear-in-Means Specification

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

$$\ln V_{it} = \beta \left( \frac{1}{|U_i|} \sum_{j \in U_i} \ln V_{jt} \right) + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.16)$$

The only difference between this specification and our main specification given by equation 1.1 is the main independent variable of interest. For the linear-in-means specification, we use  $\left( \frac{1}{|U_i|} \sum_{j \in U_i} \ln V_{jt} \right)$ , the average log viewership of the regions in  $U_i$ . We estimate this specification using both TWFE and IV. The results are presented in table 1.11.

Table 1.10: With Weather Related Content

<i>Dependent variable:</i>		
$\ln V_{it}$		
	(1) TWFE	(2) IV
$\ln V_{-it}$	1.051*** (0.010)	0.351*** (0.079)
$L_{it}$	0.020*** (0.003)	0.021*** (0.003)
$H_{it}$	0.017*** (0.004)	0.018*** (0.004)
Observations	101,500	101,500
R <sup>2</sup>	0.981	0.981

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

### 1.7.6 Auto-regressive Models

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

$$\ln V_{it} = \beta \ln V_{it} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \sum_k \left( \lambda_1 \ln V_{i(t-k)} + \phi_{1k} L_{i(t-k)} + \phi_{2k} H_{i(t-k)} + \pi_{1k} L_{-i(t-k)} + \pi_{2k} L_{-i(t-k)} \right) + \epsilon_{it} \quad (1.17)$$

This model adds various lagged terms to our main specification given in equation 1.1, indexed by the term  $k$ .  $\ln V_{i(t-k)}$  are autoregressive terms of the dependent variable,  $L_{i(t-k)}$  and  $H_{i(t-k)}$  are lagged terms of regional rain, and  $L_{-i(t-k)}$  and  $H_{-i(t-k)}$  are lagged terms of outside-region rain. We include the lagged terms of outside region rain to test if there any temporal spillovers in weather that might be biasing our results. We present results for up to  $k = 3$  in table 1.12.

For both TWFE and IV estimates, we see that all 3 autoregressive model specifications

Table 1.11: Linear-in-Means Specification

	<i>Dependent variable:</i>	
	ln $V_{it}$	
	(1) TWFE	(2) IV
$\frac{1}{ U_i } \sum_{j \in U_i} \ln V_{jt}$	0.0003*** (0.00000)	0.0001*** (0.00002)
$L_{it}$	0.020*** (0.003)	0.021*** (0.003)
$H_{it}$	0.016*** (0.004)	0.018*** (0.004)
Observations	101,500	101,500
R <sup>2</sup>	0.981	0.981

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

produce nearly identical estimates to the results presented in columns (3) and (4) of table 1.3. Again, these results do seem to suggest the TWFE estimates may be still biased, given that they are consistently and substantially more positive than IV estimates across all our estimated results. Moreover, we see that the IV coefficients are qualitatively similar even when including lagged values of outside-region weather. Hence, this suggests that temporal spillovers in weather are not biasing our estimates.

Table 1.12: Estimation of Autoregressive Models

95

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

## 1.7.7 Desktop Only Results

To address potential bias that may result from measurement error in our IP-geolocation data, we estimate the following model:

$$\ln V_{it}^{desktop} = \beta \ln V_{-it}^{desktop} + \gamma_1 L_{it} + \gamma_2 H_{it} + \alpha_i + \tau_t + \epsilon_{it} \quad (1.18)$$

The only difference between this specification and equation 1.1 is that the region and peer-region viewership variables only consider traffic from desktop devices which are generally more reliable when it comes to IP-based geolocation data. We present the results in table 1.13 below:

Table 1.13: Only Desktop Viewership

<i>Dependent variable:</i>		
$\ln V_{it}^{desktop}$		
	(1) TWFE	(2) IV
$\ln V_{-it}^{desktop}$	1.081*** (0.0009)	0.385*** (0.0064)
$L_{it}$	0.025*** (0.003)	0.025*** (0.003)
$H_{it}$	0.021*** (0.004)	0.023*** (0.004)
Observations	101,500	101,500
$R^2$	0.985	0.984

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

Here we see results that are qualitatively and quantitatively similar to our main results. Perhaps the most significant difference is rainfall estimates are slightly more positive. This should be expected as desktop viewership is likely to be more impacted by the rain compared with mobile viewership.

### 1.7.8 How does the Hierarchical Clustering Threshold Affect Estimates?

In this section, we examine how sensitive our results were to our choice of hierarchical clustering cutoff threshold. We re-estimated equation 1.1 using both TWFE and IV for many different threshold options. These are presented in table 1.14.

Table 1.14: Clustering Cutoff and Estimated Peer Effects

Thres.	N	TWFE	IV	Thres.	N	TWFE	IV
.250	971	1.044*** (0.009)	0.264*** (0.063)	.650	323	1.050*** (0.013)	0.241* (0.099)
.275	927	1.047*** (0.009)	0.312*** (0.064)	.675	299	1.051*** (0.013)	0.367*** (0.098)
.300	874	1.049*** (0.009)	0.299*** (0.067)	.700	280	1.052*** (0.013)	0.386*** (0.092)
.325	823	1.053*** (0.009)	0.366*** (0.066)	.725	257	1.049*** (0.013)	0.332** (0.096)
.350	776	1.055*** (0.009)	0.306*** (0.069)	.750	231	1.048*** (0.014)	0.294* (0.118)
.375	734	1.051*** (0.009)	0.299*** (0.073)	.775	211	1.044*** (0.015)	0.396** (0.115)
.400	698	1.050*** (0.010)	0.134 (0.086)	.800	186	1.043*** (0.015)	0.384** (0.126)
.425	661	1.055*** (0.010)	0.205* (0.084)	.825	170	1.046*** (0.015)	0.387* (0.147)
.450	624	1.053*** (0.010)	0.278** (0.080)	.850	157	1.043*** (0.015)	0.384* (0.149)
.475	591	1.053*** (0.010)	0.327*** (0.077)	.875	144	1.046*** (0.016)	0.378* (0.156)
.500	542	1.051*** (0.011)	0.340*** (0.083)	.900	133	1.044*** (0.017)	0.401* (0.166)
.525	503	1.052*** (0.011)	0.350*** (0.078)	.925	127	1.045*** (0.017)	0.356 (0.187)
.550	466	1.055*** (0.012)	0.308*** (0.084)	.950	115	1.041*** (0.017)	-0.112 (0.312)
.575	421	1.052*** (0.012)	0.380*** (0.074)	.975	107	1.046*** (0.018)	0.095 (0.294)
.600	386	1.052*** (0.013)	0.398*** (0.078)	1.000	103	1.046*** (0.019)	0.318 (0.224)
.625	355	1.050*** (0.012)	0.227* (0.091)				

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



# Chapter 2

## Is Paid Advertising Cannibalizing Organic App Installs?

### 2.1 Introduction

Mobile apps are an increasingly important aspect of people’s lives. According to the “US Time Spent with Mobile 2019” Report (eMarketer 2019), for the first time, US adults spent more time on their mobile devices than watching TV—with over 90% of that time dedicated to mobile apps. In a similar vein, consumer spending on mobile apps has also increased, reaching \$120 billion globally in 2019, more than double the amount spent in 2016<sup>1</sup> (App Annie 2020). As a consequence of this expansion of consumer attention and spending, competition in the mobile app market is quite fierce: there are millions of apps available on the Google Play and Apple App stores with thousands more being released each day.

Given such a crowded market and limited consumer attention, it should not be surprising that mobile app developers<sup>2</sup> consider marketing critical to their success. According to a recent survey conducted by The Manifest, 98% of businesses have a documented app marketing strategy and over half dedicate at least 30% of their app development budget

---

<sup>1</sup>These are gross spending numbers, inclusive of percentage taken by the app stores. However, payments for goods and services made on apps like Airbnb, Amazon, Uber, Lyft, etc. are not counted in these totals.

<sup>2</sup>While there are differences between mobile app developers and mobile app publishers, we will use the term “developers” to refer to both. The term “publishers” will generally refer to advertising publishers such as Google, Facebook, Vungle, etc.

to marketing<sup>3</sup>. In fact, the growth of mobile app install advertising has even outpaced the growth of consumer spending, more than doubling in just 2 years (\$27.1 billion in 2017 to \$57.8 billion in 2019). Moreover, future growth is projected to remain quite high, once again doubling by 2023, with total mobile app install ad spend projected at over \$118 billion<sup>4</sup>.

As app install marketing spend continues to grow, one increasingly important question is the degree to which paid advertising is cannibalizing organic installs. Though mobile app developers are certain that advertising does drive installs, there are concerns that reported metrics may be overstating advertising effectiveness. For example, in 2017, P&G’s Chief Brand Officer Marc Pritchard has threatened to cut off his company’s ad spend if publishers fail to address his concerns surrounding transparency and fraud<sup>5</sup>. In fact, prior academic research has shown that paid search advertising generally substitutes for organic traffic at least to some degree. In the most extreme case, Blake et al. (2015) found that branded search ad traffic was almost entirely recovered from organic search. Precisely measuring this substitution effect in the context of mobile app install ads has major implications for developers. Given such high market competition and the importance of marketing in user acquisition, even minor optimizations in advertising spend can drastically impact firm outcomes.

In this study, we analyze an advertising spend shutoff “experiment<sup>6</sup>” conducted by a major US-based mobile game developer we will refer to as GameSpace. Using a difference-in-differences (DiD) approach, we surprisingly find that when advertising is halted, organic installs drop by about 24%. We further replicate this result with a regression discontinuity in time (RDiT) approach which finds that the spend shutoff decreased decreased organic installs between 18% and 30%, depending on the bandwidth and model specification chosen. To try and provide some more granular insights, we also run several fixed-effects panel models. We find that every \$100 is positively associated with 4 organic installs and 34.4

---

<sup>3</sup>“How to Measure Your Mobile App Marketing.” <https://bit.ly/2Zxn1zt>. Accessed May 2020

<sup>4</sup>“App Install Ad Spend 2019-2022.” <https://bit.ly/2MhDgcX>. Accessed May 2020.

<sup>5</sup>“Digital Advertising Is Facing Its Ultimate Moment of Truth, and Billions of Dollars Are at Stake.” <https://bit.ly/2ZFFBGP>. Accessed May 2020.

<sup>6</sup>It is important to note that the term “experiment” is used in the general sense of “taking a course of action and observing the the eventual outcome” rather than the formal sense of “randomized assignment of a specific intervention.”



paid installs. These regressions also uncover evidence of temporal spillovers in paid installs (but not organic installs): every \$100 of yesterday’s spend is associated with an additional 3.4 paid installs today. Although these panel regressions lack an explicit identification strategy (besides selection on fixed effects), the estimates themselves are remarkably consistent quantitatively with our DiD and RDiT findings.

Our work contributes to the growing literature on digital advertising effectiveness. While prior work has found heterogeneous results, one consistent conclusion is that paid advertising is not quite as efficient as the direct numbers suggest due to substitution with organic channels. Our paper is the first study, to the best of our knowledge, that finds the entirely opposite effect: namely, that advertising, on average, is generating additional “organic” conversions. Moreover, our work has immediate managerial implications. Given that mobile app install advertising is 10.5% more effective than the reporting numbers suggest, mobile app publishers may be systematically under-investing in their marketing budgets.

## **2.2 Related Literature**

### **2.2.1 Advertising Shutoff Experiments**

In their pioneering study, Blake et al. (2015) (henceforth BNT) produced strong evidence that paid search advertising substitutes for organic traffic. Using a series of large scale field experiments conducted at eBay, they showed that when branded search ads were suspended, most of the lost advertising traffic simply shifted to organic search traffic. Although they also found that non-branded search ads had a positive effect on new or infrequent consumers, average returns were ultimately negative since most ads were delivered to frequent consumers who—at least in the short run—were not affected by ads.

Coviello et al. (2017) were concerned with the generalizability of BNT’s findings given that eBay was particularly well-known at the time. They essentially replicated BNT’s experiment at Edmunds.com (instead of eBay) and found drastically different results: less than half of paid search traffic was recovered through organic search. Though these re-

sults indicate that paid search may provide a positive return, they still point to a significant substitution effect between paid advertising and organic traffic.

Golden and Horton (2017) and Simonov and Hill (2019) further confirm Coviello et al. (2017)'s results. Although both studies are primarily focused on understanding how a firm's paid search advertising impacts their competitors outcomes, their "direct" results show that paid search is indeed effective, but that it crowd's out organic traffic to some degree. In the case of Golden and Horton (2017), they they noted that paid advertising is approximately 63% efficient. Similarly, Simonov and Hill (2019) found that a paid search ad cannibalized about 37.8% of a brand's organic search traffic.

Our work builds on this existing literature in a number of ways. First, our work is not limited to paid search advertising. While paid search is still the most dominant channel in digital advertising, other channels continue to gain market share each year, reaching 61% of all digital ad spend in 2019 Marin Software (2019). Second, unlike prior work, our results uncover evidence that paid advertising can complement organic channels.

## **2.2.2 Mobile Advertising**

As the importance of mobile as media channel has increased over the past decade, academic interest in mobile advertising has correspondingly intensified. Due to the unique GPS data provided by mobile devices (Shankar and Balasubramanian 2009), many have focused on quantifying the value of location-based targeting. Early work (Ghose et al. 2013, Luo et al. 2014, Fang et al. 2015, Fong et al. 2015) demonstrated the importance of geographic proximity in mobile promotion effectiveness. In particular, Luo et al. (2014) found significant interaction effects between location and temporal targeting: while same-day mobile promotions were most effective for proximal consumers, prior-day promotions were most effective for non-proximal consumers. More recent work by Ghose et al. (2019) illustrated the potential gains of trajectory-based targeting—where targeting is based on trajectories in consumers' movement patterns rather than just current location.

Some have leveraged the rich location data to understand how environmental factors impact mobile ad effectiveness. For instance, Andrews et al. (2016) showed that the purchase

rate of consumers in physically crowded subway trains was substantially higher relative to uncrowded trains. In a different vein, Li et al. (2017) confirmed the moderating effect of the weather: when it is bright and sunny outside, people are more likely and more quickly to respond to mobile promotions; rainy weather, on the other hand, decreases both response rate and speed.

Naturally, not all mobile advertising research relates to location targeting. Bart et al. (2014) investigated what types of products are most suitable for mobile display advertising. Using a large scale randomized field experiment, they determined that mobile display campaigns tended to benefit “higher involvement” and “utilitarian” products. Zhang et al. (2019) displayed the effectiveness of personalized promotion strategies in combating attrition and low engagement in mobile apps. Others have focused specifically on mobile in-app advertisements. Ghose and Han (2014) found that showing in-app ads tended to decrease app demand by approximately the same amount as an 8% price increase. On the other hand, Rafieian and Yoganarasimhan (2020) explored the interplay between in-app behavioral targeting and privacy. Though they found that while efficient targeting does increase total surplus, the ad publisher can potentially receive less total revenue since targeting “thins out the market.”

Though the literature on mobile advertising continues to grow, there has been surprisingly little work dedicated to mobile app install advertising. Most of the existing work is focused on improvements in various machine learning algorithms that power the bidding and consumer targeting systems (Ma et al. 2016, Bhamidipati et al. 2017, Sahu et al. 2018). Our work, to the best of our knowledge is one of the first empirical papers to investigate app install advertising effectiveness.

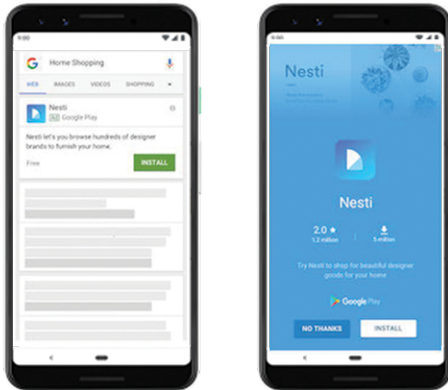
## **2.3 Data**

Mobile app install ads are advertisements that are designed to drive installs of a mobile app. Although they can appear across the the entire spectrum of digital channels (search, social media, display, in-app, video, etc.), they generally link to an app’s listing in an App Store to allow consumers to directly install the app from the ad. Moreover, they are generally

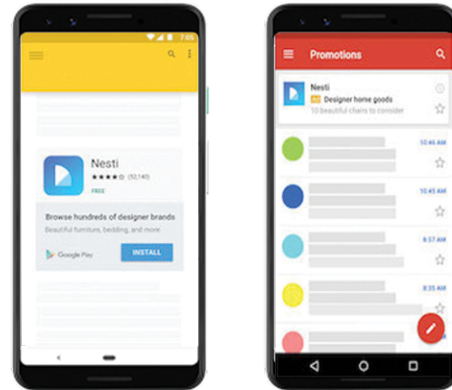
mobile-only as to make the app installation process as frictionless as possible. Figure 2-1 illustrates four examples of such ads.

Figure 2-1: Examples of Mobile App Install Ads

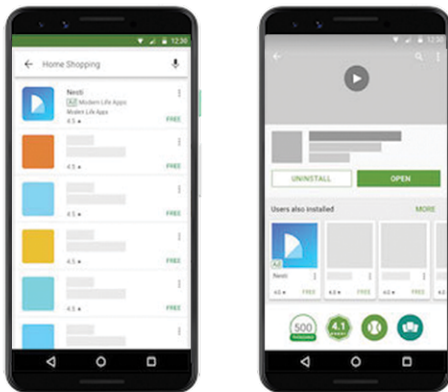
GOOGLE SEARCH NETWORK



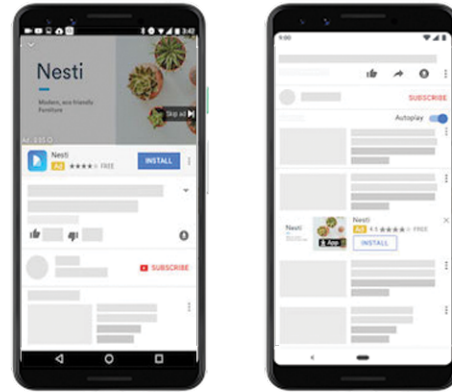
GOOGLE DISPLAY NETWORK



GOOGLE PLAY STORE



YOUTUBE



Our analysis is conducted on the historical advertising campaign data of GameSpace. This data was provided to us through our collaboration with AdTech, a US-based startup that manages and optimizes digital advertising spend on behalf of GameSpace and other clients. This data tracks the daily spend, impressions, clicks, and installs for all the digital advertising campaigns set up by GameSpace across 85 different ad publishers between for 500 days between 2018 and 2019 (exact dates are omitted due to privacy concerns). The install numbers are provided via a third-party tracker that utilizes last touch attribution. Organic install numbers, in particular, are generated simply by counting the number of

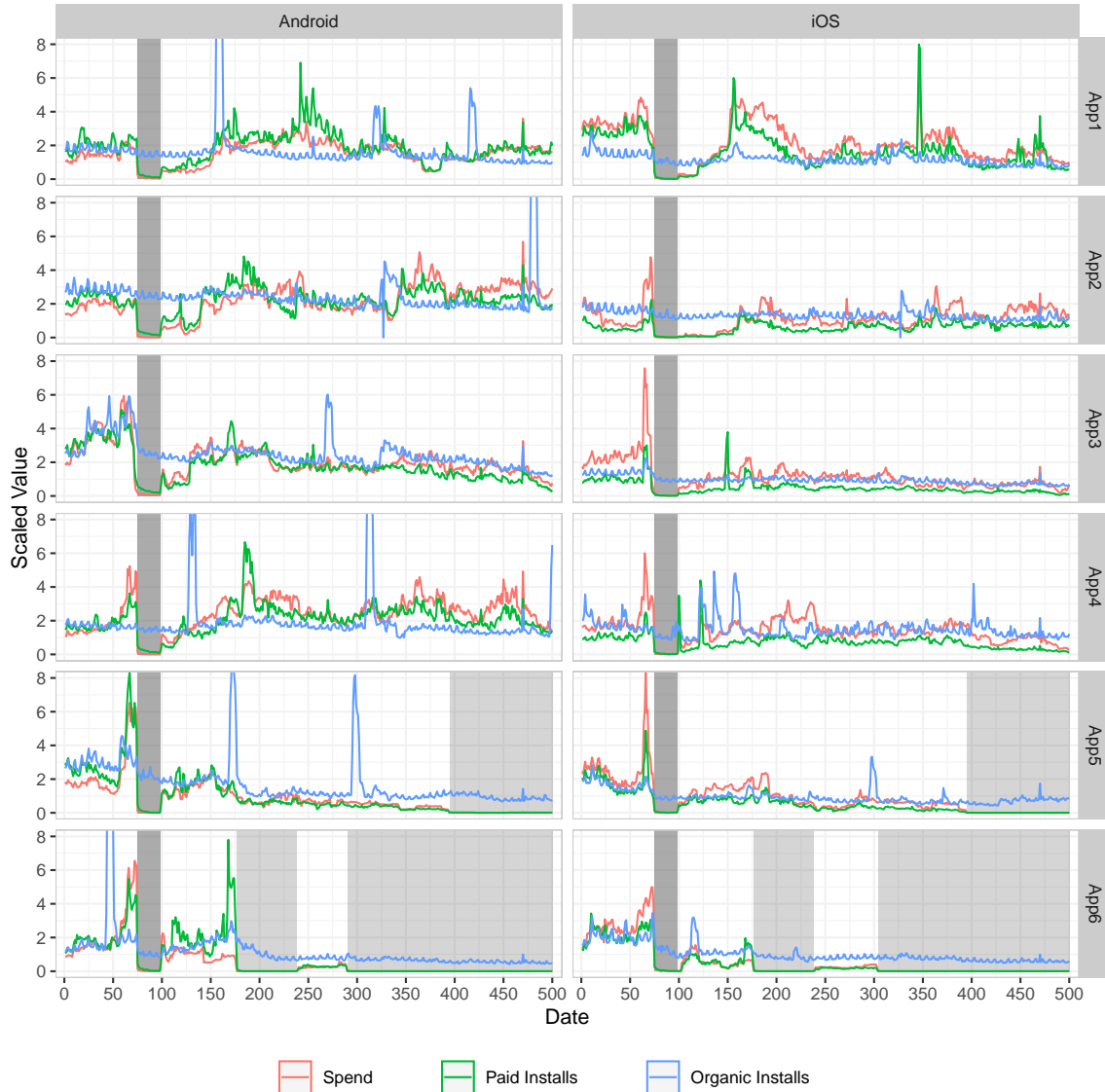
unattributed installs per day. Overall, organic installs account for just over 40% of all installs. To try and give a sense of the scope of this data, over 100 billion impressions were served worldwide during our time period.

For our analysis, we aggregate this historical campaign data by simply by summing installs and spend for all campaigns for each app and operating system (OS) combination at a daily level (so each observation of the data is uniquely identified by an app, OS, and date combination). We restrict our analysis to 6 particular mobile apps out of GameSpace's larger portfolio, leading to a dataset with 6000 observations ( $6 \text{ apps} \times 2 \text{ OS} \times 500 \text{ dates}$ ). We focus specifically on these 6 since they had the most complete and active advertising activity throughout our data. Furthermore, all these apps are relatively mature, with the youngest being released over 6 months prior to the beginning of our observation period, allowing us to avoid any inflation in organic install numbers due to media mentions. Despite our criteria for selecting these apps, it is worth noting that there is a fair bit heterogeneity with the most popular app garnering around 4 times the spend and installs of the least popular app.

To provide a sense how our data track over time, we plot the scaled values for spend (red), paid installs (green), and organic installs (blue) for the 12 different app-OS combinations in Figure 2-2. Scaled values are computed by dividing each series by the standard deviation of that metric for that particular app. This means that scale is preserved between device OS (e.g. App1 Android spend and App1 iOS spend are divided by the same number), but will differ across metric and app (e.g. App1 iOS spend, App1 iOS paid installs, and App2 iOS spend are all divided by different numbers). As can be seen in this Figure, there are several instances where organic installs are unusually high. These spikes refer to periods where an app was featured in some way on a particular app store. To avoid confounding that might result from coordination with this featuring, we simply remove these dates from analysis leaving us with a dataset with 5829 observations.

The key to our identification strategies is the spend shutoff “experiment” (similar to the branded search shutoff experiment of BNT) conducted between days 75 and 98 (highlighted in dark grey in Figure 2-2). in Figure 2-3, we zoom in on this time period to give a better sense of the underlying dynamics. To be more precise, GameSpace actually stopped adding

Figure 2-2: Scaled Time Series of Spend, Paid Installs, and Organic Installs

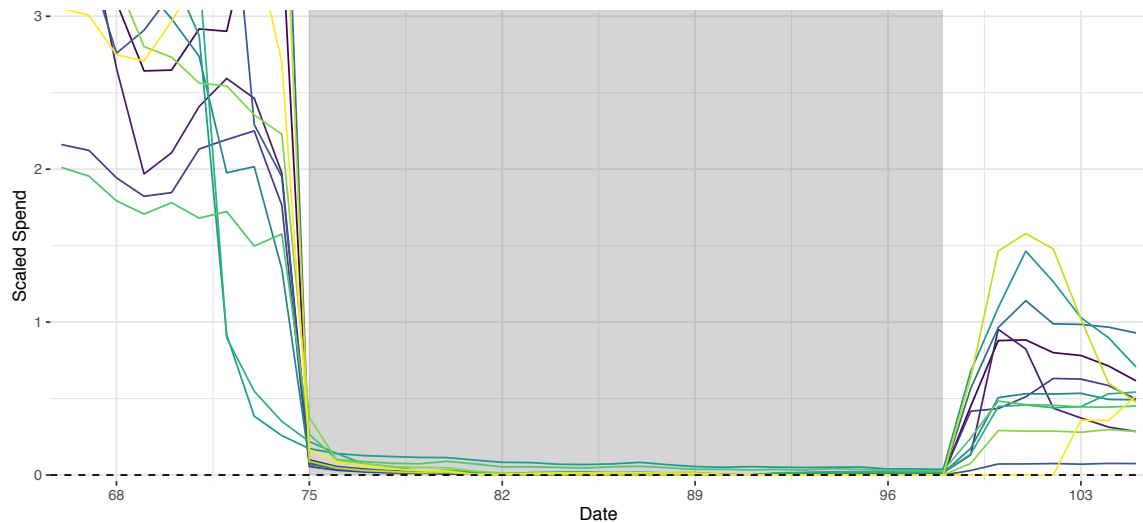


This Figure plots the scaled time series values for spend (red), paid installs (green), and organic installs (blue) faceted by app and OS. Each row denotes a different app and each column denotes a different OS. Scaled values are produced by dividing the app-OS-date summed numbers by the standard deviation of that metric across both OSs. This means that lines of different colors are scaled differently; lines of the same color across the same column but different rows are scaled differently; but lines of the same color across the same row are scaled identically. In this Figure, there are several instances of unusually high organic installs. These spikes refer to periods of time where an app was somehow featured in the respective app store.

additional spend to their accounts several days earlier. However, as many campaigns still had positive balance, it took several more days for the remaining budget to be used up. In

some cases, some apps never entirely had their advertising turned off entirely—looking closely at Figure 2-3 we can see some series never quite reach 0 (black dashed line)—though for all intents and purposes, it is virtually non-existent. One additional thing to note is the heightened amount of spending, at least for app-OS pairs in the weeks leading up to the global shutoff. This was simply due to the fact that GameSpace simply moved its spending forward, using what it would have spent during the shutoff in the weeks before.

Figure 2-3: Scaled Spend, Days 67 - 105



This Figure overlays the scaled spend, computed exactly the same as above, for all of the 12 different app-OS combinations we analyze in this study. The plot specifically zooms in on the shutoff period, which is highlighted in grey.

Unlike the other apps, GameSpace eventually decided to discontinue advertising for App5 and App6 highlighted in light grey in Figure 2-2. For App5, this occurred on day 395 and continued until the end of the observation period. For App6, this started day 177 and also continued to end of the observation period with the exception of days 239 - 289 where there was a very minor amount of ad spending.

## 2.4 Empirical Strategy and Model Specifications

In this paper, we employ 3 main empirical methods: difference-in-difference estimation (DiD), regression discontinuity in time (RDiT), and fixed effects panel regressions. We de-

scribe the details of each method and the corresponding model specifications in 3 different subsections below.

## 2.4.1 Difference-in-difference

In the context of digital advertising, DiD is typically not an appropriate empirical strategy to use. First, there isn't necessarily a natural choice of a discrete "treatment" to estimate. Moreover, even if there were, the key identifying assumption of parallel trends is nearly impossible to observe or verify since the underlying amount of spend between different units is always changing. However, in our case, GameSpace's global spend shutoff addresses both of these issues.

Naturally, the obvious choice for a "treatment" here is the an ad spend shutoff. Unfortunately, DiD cannot be used to estimate the impact of the global shutoff, as that impacted all the apps. However, it can be used to estimate the effect the shutoff for App5 and App6 as they discontinued advertising unlike the other 4 apps. Hence, we take App5-Android, App5-iOS, App6-Android, and App6-iOS as our 4 "treatment" units and the remaining 8 app-OS combinations as our control units. So how exactly can we verify parallel trends? Here the global spend shutoff comes into play. During this period, the underlying spending dynamics across all app-OS units is essentially the same. Hence, we can look specifically to this period in order to verify if parallel trends holds for organic installs or not.

To perform our DiD estimation, we employ the following model specification detailed in Equation 2.1 below:

$$\log(Y_{ijt}) = \delta D_{ijt} + \alpha_{ij} + \tau_t + \epsilon_{ijt} \quad (2.1)$$

For our study, we look at 3 main outcomes for  $Y_{ijt}$ : organic installs, paid installs, and total installs. While we are primarily interested in the number of organic installs, we still want to estimate the effect of the shutoff on paid and total installs to provide additional context. We log transform this variable as it allows us to interpret our estimated coefficients multiplicatively rather than additively, which is more appropriate given that there is significant heterogeneity (in terms of both spend and installs) across the various apps. Our treatment  $D_{ijt}$  is simply an indicator for whether spend was shutoff for app  $i$  on OS  $j$  on date  $t$ .  $\delta$  is



the main parameter of interest here denoting the the “treatment effect” of the spend shutoff. To be more specific, the counterfactual being estimated here is the relative difference in organic installs under no spend when compared to “normal” spend for the treated units or the relative treatment effect on the treated (RTT). Next, since separate app and OS fixed effects are not necessarily (or likely to be) linearly additive, we instead employ a joint app-OS fixed effect  $\alpha_{ij}$  which helps control for any time-invariant unobservables for each specific app-OS combination. Next, we account for any seasonality or time trends using time fixed effects  $\tau_t$ . Lastly,  $\epsilon_{ijt}$  denotes the error term.

## 2.4.2 Regression Discontinuity in Time

RDiT is an adaption of popular regression discontinuity (RD) framework that is wildly used in the social sciences. The key difference for RDiT is that employs time as the running variable and uses some treatment date as a threshold. Although not quite as robust as standard RD designs (see Hausman and Rapson 2018 for more details), RDiTs have seen a fair amount of use for policy evaluation, especially in the context of environmental regulation.

In our context, our choice for the threshold is naturally the start of the shutoff period April 4, 2018. While the ad spend reactivation theoretically serves as an additional discontinuity, we decided to exclude it from our analysis since the reactivation was far more gradual as can be seen in Figure 2-3. We specifically estimate 2 different specifications for our RDiT: local constant and local linear described by Equations 2.2 and 2.3 respectively:

$$\log(Y_{ijt}) = \delta D_{ijt} + \alpha_{ij} + \omega_t + \epsilon_{ijt} \quad (2.2)$$

$$\log(Y_{ijt}) = \delta D_{ijt} + \gamma_1(t - c) + \gamma_2 D_{ijt}(t - c) + \alpha_{ij} + \omega_t + \epsilon_{ijt} \quad (2.3)$$

$Y_{ijt}$ ,  $D_{ijt}$ ,  $\alpha_{ij}$ , and  $\epsilon_{ijt}$  are as they were above. While  $\delta$  technically has a slightly different interpretation here, it is essentially representing the same quantity as in Equation 2.1. Notice however, that neither of these equations contain time-fixed effects  $\tau_t$ . As time has become our running variable, including time-fixed effects would result in perfect multicollinearity. Instead, we opt to include a set of day-of-week fixed effects, denoted by  $\omega_t$

to account for weekly cyclicalities. For Equation 2.3 specifically,  $(t - c)$  simply denotes the difference in days between the current date  $t$  and the cutoff of date 75  $c$ .  $\gamma_1$  and  $\gamma_2$  simply capture the slopes on each side of the discontinuity, but ultimately aren't of any interest to us.

We purposefully limit ourselves to the simpler RD specifications since the more complex specifications run the risk of overfitting. This is especially problematic since the decrease in spending, while rather drastic, is still not perfectly sharp which “stretches out” the discontinuity. As such, higher order polynomial terms are more likely to be capturing this underlying shift in spending, rather than a true nonlinearity. For instance, as seen in Figure 2-3, the two series corresponding to App3 already attained a much lower level of spend 3 days before the cutoff compared to the other apps. Though this is the most prominent example, many other apps already had declining ad spend in the days before  $c$ . Furthermore, there is a risk of persistence effects: installing an app today due to an ad exposure yesterday or the day before. While such effects will eventually decay, this will also have a similar effect of stretching out the the discontinuity.

### 2.4.3 Fixed Effects Linear Panel Regression

Although our DiD and RDiT specification above specifications allow for identification of causal effects, one major limitation is that the “treatment” being evaluated is quite blunt. Practically speaking, firms would tweak and optimize their ad spend rather than simply shutting it off entirely. To try and understand how continuous spend interacts with organic installs, we use the following panel regression specifications detailed in Equations 2.4:

$$Y_{ijt} = \beta \text{spend}_{ijt} + \pi_{ijt} + \epsilon_{ijt} \quad (2.4)$$

Here, we change our main variable of interest to  $\text{spend}_{ijt}$ , the amount of dollars spent advertising app  $i$  for OS  $j$  on date  $t$  across all publishers and campaigns. For this specification, we do not log transform any of our variables because we think the underlying relationship between spending and installs (organic, paid, or total), is likely better approximated linearly, rather than multiplicatively or as an elasticity. It is important to note that we

avoid using date-fixed effects here since date-level shocks are likely multiplicative rather than additive<sup>7</sup>. Instead, we employ a special app-OS-time fixed effect  $\pi_{ijt}$  that consists of the interactions between the app-OS fixed effects  $\alpha_{ij}$  and day-of-week fixed effects  $\omega_t$  as well as the interactions between app-OS fixed effects  $\alpha_{ij}$  and week-fixed-effects  $\psi_t$ . More formally:  $\pi_{ijt} = \alpha_{ij} \times \omega_t + \alpha_{ij} \times \psi_t$ .

## 2.5 Results

Like above, this section is organized into three separate subsections each corresponding to our 3 estimation methods. In Section 2.5.1 we test the parallel trends assumption and report the results of our DiD estimation (Equation 2.1). In Section 2.5.2, we report the our RDiT estimates (Equations 2.2 and 2.3) for several different choices of bandwidth. Lastly, in Section 2.5.3, we report our panel regression estimates (Equation 2.4). To account for unobserved events like app or OS updates, we report all our results with app and OS 2-way clustered standard errors. Our analysis was conducted entirely in R (R Core Team 2019) and all plots were created using the `ggplot2` package (Wickham 2016a). Linear models and Poisson models were fitted using the `lfe` (Gaure 2013) and `fixest` (Bergé 2018) packages respectively.

### 2.5.1 Difference-in-Difference

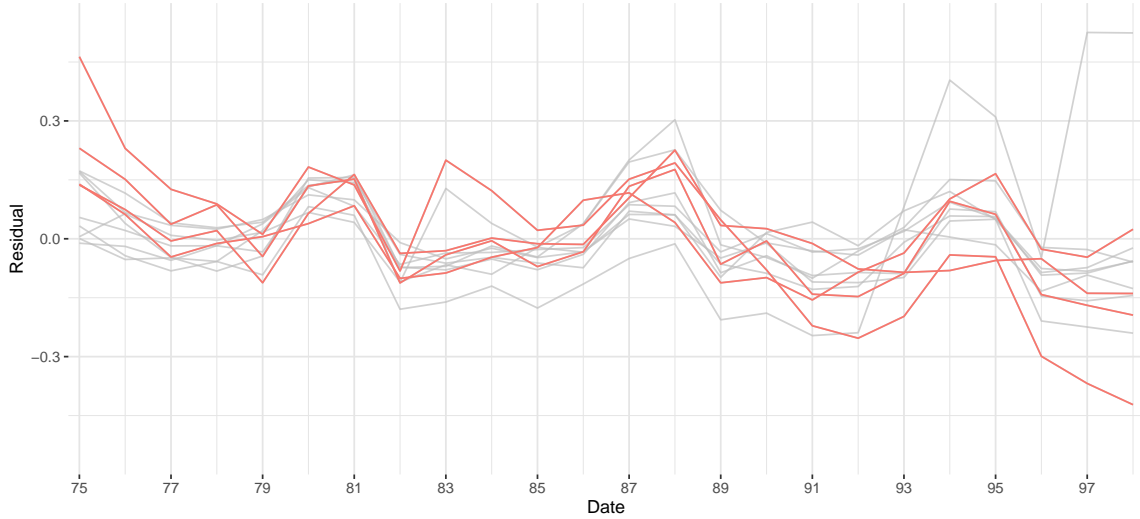
We begin this section by first determining whether the parallel trends assumption holds. In Figure 2-4, we plot the residuals of log organic installs after partialling out (Frisch and Waugh 1933, Lovell 1963) the app-OS fixed effects during the shutoff period for each app-OS combination.

Looking at this plot, there doesn't seem to be any systematic difference between the trends of the "treated" (in red) and control units (in grey). We more formally verify this by

---

<sup>7</sup>To elaborate a bit, suppose that there are 2 apps: AppA which is incredibly popular and AppB which is not. Consider that app installs are significantly higher on weekends compared to weekdays. In a level-level model, date fixed effects impose the assumption that the weekend would increase both apps' installs by the same number, clearly unreasonably given AppA's much greater popularity. Rather, such date-level shocks are better approximated multiplicatively, i.e. both apps receive 15% more installs during the weekend.

Figure 2-4: Parallel Trends



This Figure plots the residuals of regressing  $\log(Y_{ijt})$  on the app-OS fixed effects  $\alpha_{ij}$  over the time span of the spend shutoff. The time series of the 4 “treated” units are highlighted in red.

conducting a Wald Test of the full model:

$$\log(Y_{ijt}) = \alpha_{ij} + \tau_t + \sum_t \lambda_t(T_{ij} \times \tau_t) + \epsilon_{ijt}$$

against the restricted model of:

$$\log(Y_{ijt}) = \alpha_{ij} + \tau_t + \epsilon_{ijt}$$

where  $T_{ij}$  is simply an indicator for whether the app-OS unit  $ij$  is considered a treated unit or not. Here we find an F-stat of 1.1417 which corresponds to a p-value of 0.3002 indicating that all the  $\lambda_t$  parameters are not statistically significant<sup>8</sup>. As the  $\lambda_t$  parameters are meant to capture the differences between treated and control units on date  $t$ , their joint insignificance is strong evidence indicating that parallel trends does indeed hold.

Next, we move onto the results of our DiD estimation. We fit the specification given by Equation 2.1 for our 3 outcome variables of organic installs, paid installs, and total installs.

<sup>8</sup>This F-stat is produced assuming homoskedastic errors, which provides the most conservative test since we want to fail to reject the null. We recomputed the test with both heteroskedastic and clustered errors which yielded even less significant F-stats.

As a robustness check, we also estimate a Poisson regression version of our specification. We report these results in Table 2.1.

Table 2.1: Difference-in-Difference Estimates

	<i>Dependent variable:</i>					
	Organic Installs		Paid Installs		Total Installs	
	(1) OLS	(2) Poisson	(3) OLS	(4) Poisson	(5) OLS	(6) Poisson
$D_{ijt}$	-0.279** (0.086)	-0.365*** (0.090)	-5.508*** (0.424)	-5.568*** (0.475)	-1.164*** (0.214)	-1.476*** (0.234)
Adj. R <sup>2</sup>	0.920	0.932	0.926	0.794	0.868	0.844
Obs	5829	5829	5829	5829	5829	5829

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001. Cluster-robust standard errors are reported. Adj. Pseudo-R<sup>2</sup> is reported for Poisson models.

Contrary to our initial expectations, we found that the advertising suspension had a highly statistically significant negative effect ( $p < 0.002$ ) on organic installs. Taking the point estimate as given, this implies that apps will lose approximately 24.3% of their organic installs compared to “normal” levels of ad spend. This result is rather shocking as all the previous academic research has provided strong evidence showing the substitution between paid and organic channels. To check if everything was working as intended, we also looked at the results for paid installs and total installs. In fact, the point estimate for paid installs of -5.508 indicates that paid installs dropped by 99.6%, almost exactly what we would expect. As for total installs, the point estimate of -1.164 corresponds to a drop of 68.8% which is remarkably close to what the other coefficients would suggest. Organic and paid and organic installs make up approximately 40% and 60% of total installs respectively. Multiplying each of these shares by the estimated impact of the ad shutoff leads to a projected drop of 69.5% of total installs, less than 1% off our estimate. These results are all qualitatively and quantitatively similar when estimating a Poisson model. If anything, the Poisson estimates indicate even stronger effects of the ad suspension.

## 2.5.2 Regression Discontinuity in Time

We begin this section by plotting the regression fit from both Equations 2.2 and 2.3 in Figure 2-5 for several different choices of the bandwidth: 3, 7, 11, 15, 19, and 23 for each of the 3 outcomes. The fit produced from different bandwidths are represented with different colors, starting from dark purple (representing 3) to yellow (representing 23). The points on the plot are formed using the residuals of  $\log(Y_{ijt})$  after partialling out the app-OS fixed effects ( $\alpha_{ij}$ ) and day-of-week fixed effects ( $\omega_t$ ). We also report the estimated discontinuities and app and OS 2-way clustered standard errors for all these different choices of models and bandwidths in Table 2.2.

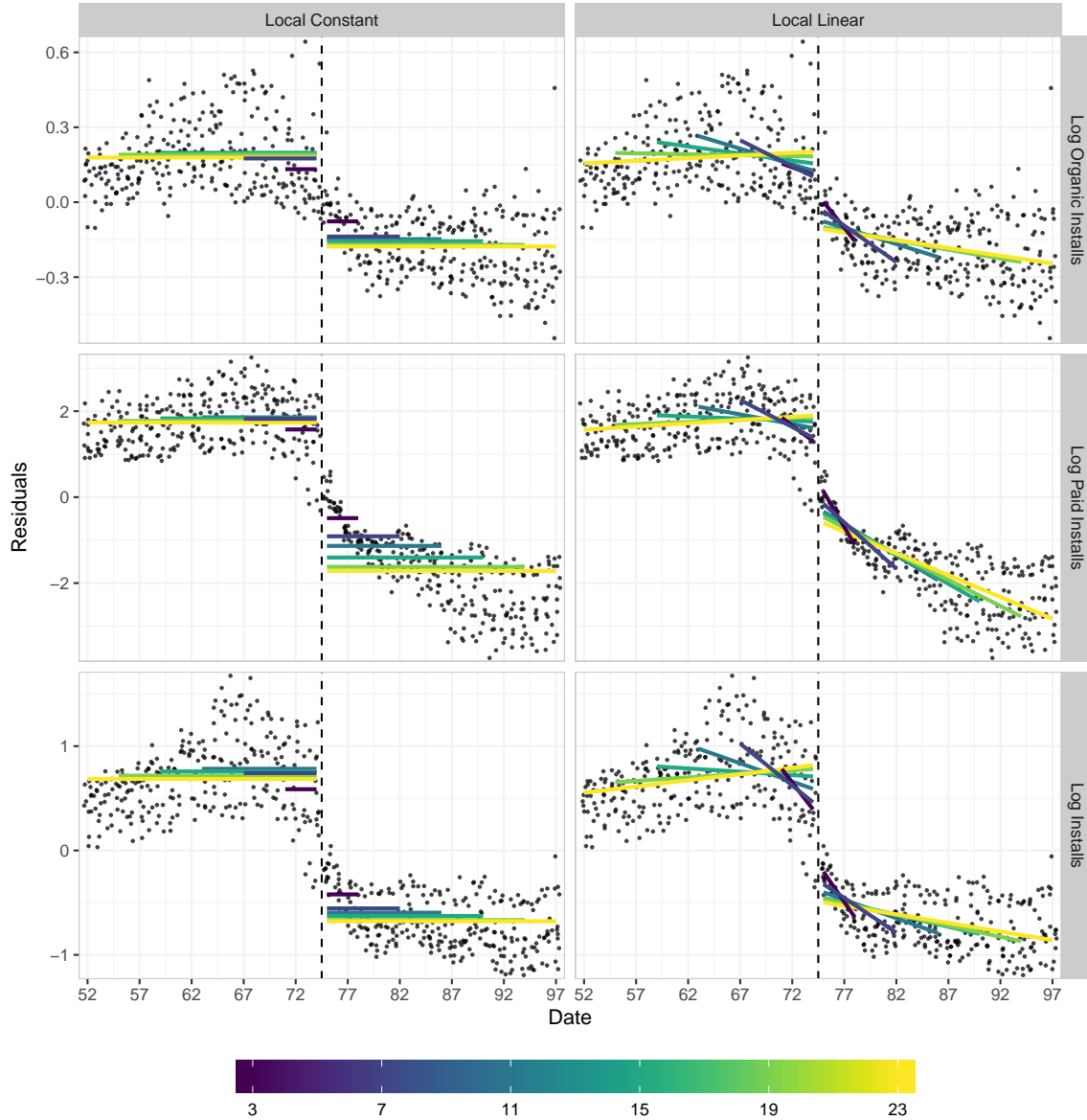
Table 2.2: RDiT Estimates

		<i>Dependent variable:</i>					
		Organic Installs		Paid Installs		Total Installs	
Model		(1) L. Constant	(2) L. Linear	(3) L. Constant	(4) L. Linear	(5) L. Constant	(6) L. Linear
OLS-3		-0.209** (0.066)	-0.075 (0.064)	-2.064*** (0.256)	-0.872*** (0.252)	-1.008*** (0.113)	-0.461*** (0.185)
Poisson-3		-0.189** (0.065)	-0.099 (0.057)	-2.039*** (0.216)	-1.263*** (0.186)	-1.017*** (0.107)	-0.645*** (0.156)
OLS-7		-0.313*** (0.069)	-0.113 (0.063)	-2.733*** (0.366)	-1.368*** (0.196)	-1.298*** (0.125)	-0.711*** (0.136)
Poisson-7		-0.275*** (0.071)	-0.118 (0.062)	-2.536*** (0.354)	-1.442*** (0.138)	-1.254*** (0.136)	-0.768*** (0.105)
OLS-11		-0.344*** (0.067)	-0.191** (0.069)	-2.991*** (0.431)	-1.851*** (0.215)	-1.379*** (0.136)	-0.956*** (0.122)
Poisson-11		-0.303*** (0.066)	-0.174* (0.072)	-2.748*** (0.424)	-1.836*** (0.163)	-1.334*** (0.149)	-0.975*** (0.114)
OLS-15		-0.354*** (0.063)	-0.249*** (0.071)	-3.237*** (0.469)	-2.070*** (0.249)	-1.385*** (0.139)	-1.151*** (0.121)
Poisson-15		-0.312*** (0.062)	-0.224** (0.074)	-2.864*** (0.3782)	-2.041*** (0.195)	-1.334*** (0.143)	-1.136*** (0.131)
OLS-19		-0.361*** (0.062)	-0.281*** (0.068)	-3.393*** (0.490)	-2.272*** (0.296)	-1.384*** (0.130)	-1.238*** (0.127)
Poisson-19		-0.317*** (0.061)	-0.247*** (0.068)	-2.933*** (0.396)	-2.199*** (0.218)	-1.325*** (0.141)	-1.217*** (0.137)
OLS-23		-0.356*** (0.065)	-0.309*** (0.065)	-3.452*** (0.500)	-2.459*** (0.326)	-1.364*** (0.126)	-1.311*** (0.136)
Poisson-23		-0.316*** (0.061)	-0.268*** (0.067)	-2.974*** (0.401)	-2.282*** (0.235)	-1.314*** (0.134)	-1.264*** (0.146)

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001. Cluster-robust standard errors are reported. Adj. Pseudo-R<sup>2</sup> is reported for Poisson models.

Beginning with the results on organic installs, regardless of our choice of bandwidth

Figure 2-5: Regression Discontinuity in Time



This figure plots the residuals after partialing out the app-OS fixed effects ( $\alpha_{ij}$ ) and day-of-week fixed effects ( $\omega_t$ ) for the period 23 days prior to the global shutoff until the end of the shutoff from log organic installs (top), log paid installs (middle), and log installs (bottom). A tiny bit of random noise is added to each point for visualization purposes. The scatterplot is duplicated so that the fitted values of the local constant model (left) can be plotted separately from the fitted values of the local linear model (right). Different lines denote fitted values produced from different choices for the RD bandwidth, ranging from 3 (dark purple) to 23 (yellow).

or model specification, the estimated discontinuity is always negative and generally highly statistically significant, confirming our DiD results above. In the case of the Local Constant

specification, the estimated discontinuity becomes rather stable at approximately -0.35 (or approximately -0.31 using Poisson estimates) after bandwidth reaches 11 days before and after the cutoff. On the other hand, for the local linear specification, the estimated discontinuity is grows increasingly negative as we increase the bandwidth. The results for paid installs and total installs are also qualitatively similar to our DiD results above.

Typically, smaller bandwidths are supposed to provide less biased estimates of the local average treatment effect (LATE) of interest. However, there are some concerns with following this general rule in our context. Recall that our treatment is a blunt approximation of the underlying effect of advertising spend which first ramps up and ramps down in the weeks prior to the shutoff. This can be seen clear as the the slope of the local linear regression in the pre-shutoff period changes from negative to positive as the bandwidth increases. As such, the smallest bandwidths here aren't necessarily going to provide the best estimate of the counterfactuals that we are interested in. Moreover, as we mentioned before, there are likely to be some persistence effects of advertising in the previous days that takes some time to fully decay. We can see some evidence by looking at slope of the local linear regression on the shutoff side of the discontinuity. Regardless of choice of bandwidth, the slope is always negative. Although spend was still decreasing during the shutoff period, by the time the shutoff started, spend was already at such low levels further decreases were barely perceptible, even after zooming in in Figure 2-3. In contrast, the decline immediately following the cutoff as seen in Figure 2-5 is much more significant and thereby unlikely to driven by the nearly flat decreases in spending.

As such, we consider the medium-bandwidth Local Constant results to be most credible estimates of a shutoff effect. Taking the point estimates as given, this suggests that shutting off spend decreases organic installs by about 29.5%. Recall however, that GameSpace shifted much of spend forward, thereby inflating the amount of spending that occurred prior to the shutoff. As such, we find this result very much in line with our prior DiD results as we would naturally expect going from abnormally high spend to no spend is going to produce a larger effect than going from "normal spend" to no spend.



### 2.5.3 Fixed Effect Panel Regressions

As mentioned earlier, both DiD and RDiT use an extremely blunt measure of spend that only allows us to estimate very extreme counterfactuals. Although our panel regression results do not have an explicit identification strategy (aside from assuming that fixed effects manage to control for all the unobserved confounders), the correlational estimates we provide may still be quite useful. We present the results of our panel regressions in Table 2.3. To improve readability, coefficients and standard errors have been multiplied by 100. As such, the coefficients can be interpreted as organic/paid/total installs per \$100 dollars. Beyond just the specification given by Equation 2.4, we also fit additional models that include lagged spending as covariates to investigate potential temporal spillovers.

Table 2.3: Panel Regressions

	<i>Dependent variable:</i>								
	Organic Installs			Paid Installs			Total Installs		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$\text{spend}_{ijt}$	3.912*** (0.502)	4.012*** (1.032)	4.021*** (1.003)	36.74*** (3.854)	34.39*** (3.667)	34.39*** (3.635)	40.65*** (4.347)	38.40*** (3.237)	38.40*** (4.544)
$\text{spend}_{ij(t-1)}$	–	–0.147 (1.100)	–0.195 (0.776)	–	3.384*** (0.473)	3.359*** (0.925)	–	3.237*** (0.159)	3.164*** (1.206)
$\text{spend}_{ij(t-2)}$	–	–	0.059 (0.438)	–	–	0.032 (0.997)	–	–	0.092 (1.024)
Adj. R <sup>2</sup>	0.975	0.975	0.975	0.958	0.958	0.958	0.971	0.971	0.971
Obs	5829	5817	5805	5829	5817	5805	5829	5817	5805

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001. Cluster-robust standard errors are reported. Adj. Pseudo-R<sup>2</sup> is reported for Poisson models.

Given our prior results, we are not surprised to find that advertising spend has a statistically significant positive relationship with organic installs. Based on our results in columns (2) and (3), there seems to be very little evidence to suggest that advertising spend impacts organic installs in future periods, as neither of the lagged terms are statistically significant. Though not displayed in the table, we ran models with up to 28 lags, with similar results: a statistically significant same day coefficient of approximately 4 and lagged coefficients statistically indistinguishable from 0. As such, our results indicate that every \$100 spent on advertising is associated with approximately 4 additional organic installs. Remarkably, this number is extremely consistent with our RDiT findings. Taking the difference in spend be-

tween the pre-shutoff period and the shutoff period and then multiplying by 0.04 produces numbers in the ballpark of a 29.5% reduction in organic installs.

Moving over to the paid installs results, we naturally see a much larger effect on spend. Unlike the organic installs results, columns (5) and (6) suggest that advertising does impact next day paid installs, as well as same day paid installs. Again, we ran models with up to 28 lags and the first lagged term stays statistically significant and quantitatively similar across all the models. Again the interpretation of these coefficients are quite straightforward: every \$100 dollars spent on advertising is associated with approximately 34.4 paid installs on the same day and 3.4 paid installs the next day. Overall, our results suggest that may be receiving approximately 10.5% more installs than anticipated.

## 2.6 Discussion

Despite the importance of advertising to mobile app publishers, not much is actually known about the optimization mobile app install ad budgets. Although prior research has shown that paid channels substitute for organic ones, the effect can vary quite wildly. Given the overall competitiveness of the mobile app market, precisely determining the degree to which paid advertising is cannibalizing organic installs is an incredibly important question for ad spend optimization. To answer this question, we study an ad spend shutoff “experiment” conducted by a major mobile game developer. Using a variety of empirical methods, we find no evidence of cannibalization. Rather, our results indicate that advertising is actually boosting the number of organic installs. Incredibly, our results are both qualitatively and quantitatively consistent across all of our different empirical strategies.

One possible reason that our results are so different may be due to the fact that prior experiments have been focused on search advertising. Since users are searching for specific keywords, they are much more likely to be aware of the site that is being advertised. In our context, advertising occurs across the entire spectrum of digital channels including search, social media, display, video, and in-app (though largely mobile rather than desktop). Furthermore, the mobile app market has an incredibly wide selection. Given that people have limited attention, these mobile app install ads could be generating awareness which even-

tually leads to an organic install. If this is the case, it would be somewhat consistent with prior findings. While Blake et al. (2015) found very large substitution effects in general, they noted a positive impact of non-branded advertising on new and infrequent consumers. In a similar vein, both Coviello et al. (2017) and Golden and Horton (2017) found significantly smaller crowding out effects of branded advertising, potentially since the firms they were analyzing were not quite as well-known as eBay.

Although we are quite confident in our results, our study is not without limitations. First, the “experiment” at the heart of our analysis is not a true experiment. While GameSpace did not strategically choose a date to conduct the experiment, there nonetheless may be something special about the shutoff period that could be driving our estimates. If this were the case however, it is worth noting that only the RDiT results would be confounded. For our DiD results, the shutoff period is only being used to verify parallel trends. The DiD estimates are computed from the future shutdown periods of App5 and App6, and would not be impacted by a hypothetical shutoff period confounder unless it was somehow causing parallel trends to only hold during the shutoff. Another limitation is with the aggregate nature of our analysis. Although we have daily campaign level data from over 85 different publishers, we are unable to identify campaign or publisher-level heterogeneity since the spend shutoff applied across all campaigns of all publishers. Though we could simply try to attribute based on publisher share, such projections may be a quite misleading. For example, App Store search ads may substitute for organic installs similar to how paid search ads substitute for organic clicks. Lastly, the results from our study may not readily generalize. Our data comes from one the largest mobile gaming developers and advertising effects may be quite different for different types of apps or developers of different sizes.

Despite these limitations, we believe that our study makes several novel contributions. To start, our study is, to the best of our knowledge, the first advertising spend shutoff that uncovers the possibility for complementary rather than substitution between paid and organic channels. Furthermore, we are one the first papers to examine mobile app install advertising, a \$57.8 billion and rapidly growing market. Beyond this, our work has immediate managerial implications. Our results show that paid advertising is about 10.5% more efficient than the tracked numbers would suggest. As such, mobile app developers that have

optimized their spend just based on reported metrics may potentially be under-investing in their marketing.

## **Chapter 3**

# **The Interdependence of Regional COVID-19 Reopenings in the United States**

### **3.1 Introduction**

Countries around the world imposed strict limits on human mobility by banning gatherings, closing down non-essential businesses, and instituting shelter-in-place policies in response to the COVID-19 pandemic. These measures effectively reduced the spread of the virus Flaxman et al. (2020), Hsiang et al. (2020), preventing the rise of new cases and their associated morbidity and mortality. However, after shutting down and sheltering-in-place, many countries have since lifted lockdown orders and are beginning to reopen to businesses, schools, and travel. While there is abundant research on the efficacy of shutting down, there unfortunately exists little quantitative evidence on the impact of reopening or the factors that contribute to making it safe and effective. Although many reopenings seem to have proceeded safely, the United States serves as a cautionary tale. After reopenings began, several COVID-19 hotspots emerged in the U.S., causing local governments to reimpose social distancing measures like bar closures and limits on gatherings. As spikes in new cases continue to reemerge in these American hotspots, it is critical we understand

how reopenings contribute to their resurgence, especially given the relatively long incubation period and asymptomatic spread of COVID-19, which creates longer lead times in assessing the eventual impacts of policy decisions.

Population-scale digital trace data Buckee et al. (2020) has been useful for studying the impacts of social distancing policies and what makes them successful Oliver et al. (2020). Researchers have shown, for example, that demographics attributes Olsen and Hjorth (2020), political partisanship Painter and Qiu (2020), Allcott et al. (2020), broadband access Chiou and Tucker (2020), belief in science Brzezinski et al. (2020), and information exposure Simonov et al. (2020), Ash et al. (2020) moderate compliance with social distancing policies. Holtz et al Holtz et al. (2020) find strong evidence of cross-county spillovers from shelter-in-place policies, underscoring the importance of governmental coordination to reduce a potential “loss from anarchy” in piecemeal implementations of closure policies across regions. Unfortunately, Holtz et al only analyze data from March and April, before shelter-in-place policies began to lift, and little empirical research investigates the effects of subsequent reopening policies across regions.

If reopening creates substantial mobility and exhibits strong spillover effects, countries that reopen without national coordination could face significant difficulty in controlling the resurgent spread of the coronavirus. If some regions reopen while others remain closed, travel and social spillovers could create mobility effects in neighboring and distant states that reduce the effectiveness of isolated regional closures. If pandemic resurgence creates an uncoordinated oscillation between reopening and closure in different regions at different times, the coronavirus could ricochet from region to region, making the pandemic hard to control and dramatically increasing the economic and public health costs of the crisis.

We therefore combined data on the mobility of over 22 million mobile devices, daily data on state-level closure and reopening policies, social media connections among 220 million U.S. Facebook users, temperature and precipitation from 62,000 weather stations and county-level census data to measure the direct impact of a focal state’s COVID-19 closure and reopening policies on its own population’s mobility patterns; the spillover effects of other states’ closure and reopening policies on a focal state’s mobility patterns; and the mediation of these effects by endogenous peer behavior across state and county borders

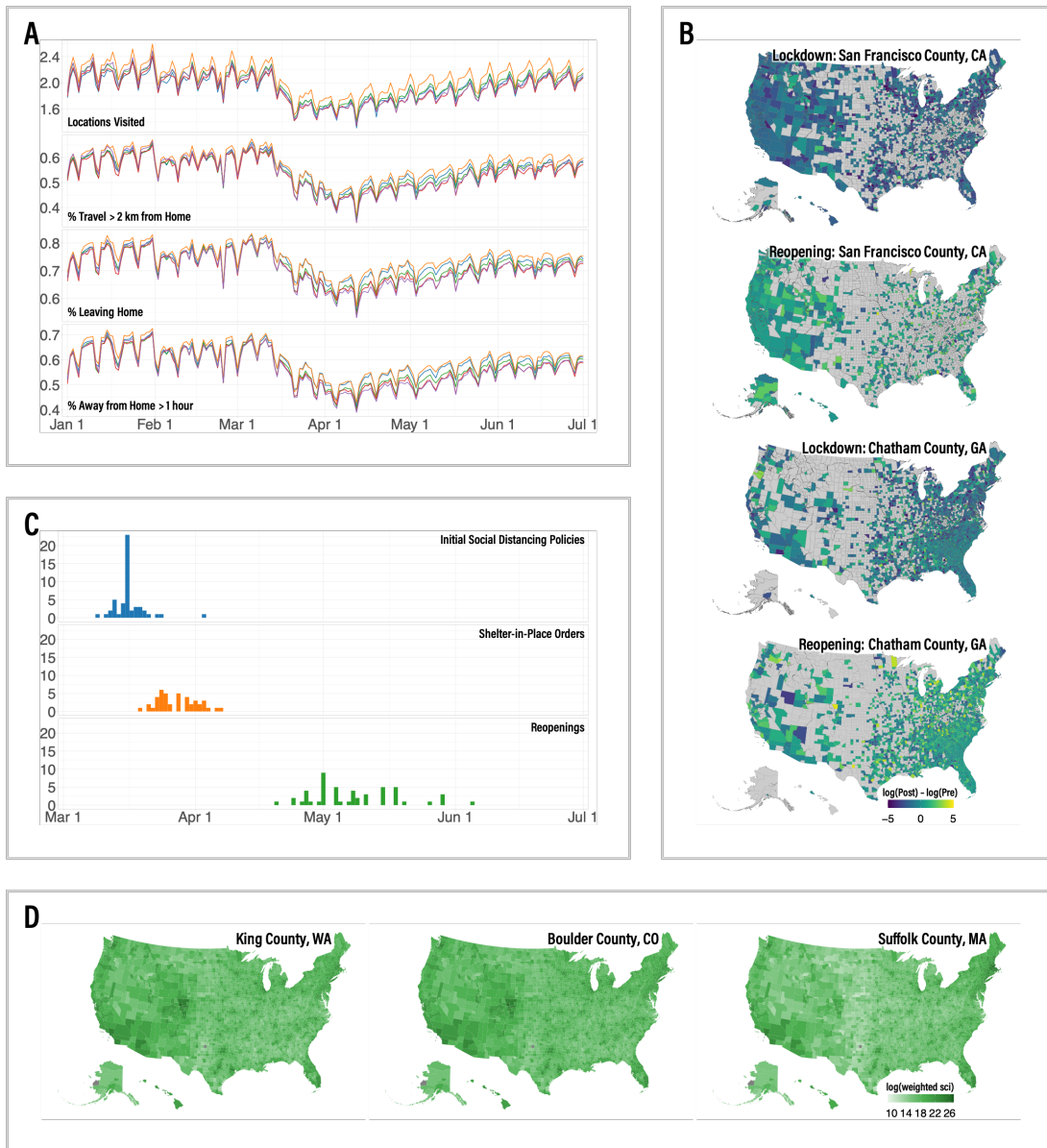
from January 1, 2020 to June 30, 2020. We further investigated the impacts of both origin- and destination-county closure and reopening policies on cross-county mobility across directed, dyadic pairs of counties, capturing the travel related spillover effects created by uncoordinated policies implemented across states and counties.

To construct daily measures of human mobility, we leverage differentially private, aggregated data on over 22 million mobile devices provided by Safegraph. For each county, we track the daily number of locations visited by mobile devices, the proportion of devices that travel less than 2km away from home, the proportion of devices leaving home, and the proportion of devices that spend more than 1 hour away from home (Fig. 3-1A). To construct measures of county-to-county travel, we track the both number of devices moving from an origin county to a destination county (Fig. 3-1B) and the proportion of origin devices that move from an origin county to a destination county for each origin-destination county pair. We limit our analysis to the 2683 counties with a daily mean device count of at least 500 to minimize measurement error induced by the Laplacian noise introduced by Safegraph’s differential privacy algorithm.

We analyze state-level data on closure and reopening policies from the COVID-19 US State Policy (CUSP) Database Raifman et al. (2020). Due to the different approaches taken by various states across the closure and reopening policy space, we simplify our analysis to three consolidated “policy periods:” the “initial policy period” (*ip*), which covers the period from when a state implements its first closure policy of any kind until it implements a stay-at-home order; the “stay-at-home period” (*sh*), which covers the duration of a statewide stay-at-home order or until the state starts to reopen; and the “reopening period” (*ro*), which starts after a state begins its reopening plan (Fig. 3-1C).

To construct peer policy and mobility measures, we rely on Facebook’s “Social Connectedness Index” (SCI) Bailey et al. (2018), which provides a measure of the intensity of Facebook connectedness between geographic locations, generated from an anonymized snapshot of the entire Facebook friendship network in the U.S. (Fig. 3-1D). These data are further supplemented with temperature and precipitation data from the Global Historical Climatology Network (GHCN) Menne et al. (2012) and county-level census population counts.

Figure 3-1: Visualizations of Data



(A) shows the time series trends from Jan 1, 2020 until June 30, 2020 of the following mobility outcomes: number of locations visited per device, the proportion of devices traveling more than 2 km, the proportion of devices leaving home, and the proportion of devices that are away from home for more than 1 hour. Different colors correspond to averages across clusters of 10 states grouped by the time they started reopening. (B) shows some examples of the difference in travel to a destination county for the 3 weeks before and after a shelter-in-place order is implemented or reopening starts. (C) plots the count of states that enter into a particular policy period on each day. (D) displays some examples of population-weighted social connectedness index used to construct socially weighted measures of alter state policies and peer behavior.



## 3.2 Model Specifications

We begin our empirical analysis with a difference-in-differences (DiD) estimation—an approach widely used across economics, political science, and public health for policy or program evaluation. The base model is specified as follows:

$$\log(Y_{it}) = \delta_{(ip)}D_{it}^{(ip)} + \delta_{(sh)}D_{it}^{(sh)} + \delta_{(ro)}D_{it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it}, \quad (3.1)$$

where  $\log(Y_{it})$  denotes the log transformed mobility outcomes (e.g. the number of locations visited or the fraction of devices leaving home). The policy variables,  $D_{it}^{(ip)}$ ,  $D_{it}^{(sh)}$ , and  $D_{it}^{(ro)}$  are binary indicators that take the value 1 once county  $i$  is subject to a statewide closure policy of any sort (ip), a stay-at-home order (sh), and reopening (ro) respectively.  $f(W_{it})$  flexibly controls for the potential non-linear effects of weather using a “double machine learning” approach Chernozhukov et al. (2018), while  $\alpha_i$  and  $\tau_t$  denote a set of county and time fixed effects and  $\epsilon_{it}$  denotes the error term.

We extend this base specification to capture spillover effects with the following specifications:

$$\begin{aligned} \log(Y_{it}) = & \delta_{(ip)}D_{it}^{(ip)} + \delta_{(sh)}D_{it}^{(sh)} + \delta_{(ro)}D_{it}^{(ro)} + \\ & \gamma_{(ip)}D_{-it}^{(ip)} + \gamma_{(sh)}D_{-it}^{(sh)} + \gamma_{(ro)}D_{-it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it} \end{aligned} \quad (3.2)$$

$$\begin{aligned} \log(Y_{it}) = & \beta \log(Y_{-it}) + \delta_{(ip)}D_{it}^{(ip)} + \delta_{(sh)}D_{it}^{(sh)} + \delta_{(ro)}D_{it}^{(ro)} + \\ & \gamma_{(ip)}D_{-it}^{(ip)} + \gamma_{(sh)}D_{-it}^{(sh)} + \gamma_{(ro)}D_{-it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it}, \end{aligned} \quad (3.3)$$

where  $D_{-it}^{(ip)}$ ,  $D_{-it}^{(sh)}$ , and  $D_{-it}^{(ro)}$  denote the socially weighted average of alter states’ policies, weighted by Facebook connectedness. The cross-state policy spillovers are captured by the terms  $\gamma_{(ip)}$ ,  $\gamma_{(sh)}$ , and  $\gamma_{(ro)}$  respectively. Endogenous peer behavior is captured by  $\log(Y_{-it})$ , which is the log transformed socially weighted average of alter states’ mobility behavior. As estimation of peer effects is generically confounded in observational data Manski (1993), we employ an instrumental variables (IV) approach where we leverage alter county weather as a source of exogenous variation to properly identify the endogenous peer effect  $\beta$  Coviello et al. (2014b), Aral and Nicolaides (2017b), Aral and Zhao (2020), Holtz

et al. (2020).

To measure the impact of policy on cross-county mobility, we employ the following specifications:

$$\log(Y_{o \rightarrow d,t}) = \sum_m \lambda_m D_{ot}^m + \sum_n \psi_n D_{dt}^n + \alpha_{o \rightarrow d} + \tau_t + \epsilon_{o \rightarrow d,t} \quad (3.4)$$

$$\log(Y_{o \rightarrow d,t}) = \sum_m \lambda_m D_{ot}^m + \sum_n \psi_n D_{dt}^n + \sum_m \sum_n \pi_{m,n} (D_{ot}^m * D_{dt}^n) + \alpha_{o \rightarrow d} + \tau_t + \epsilon_{o \rightarrow d,t} \quad (3.5)$$

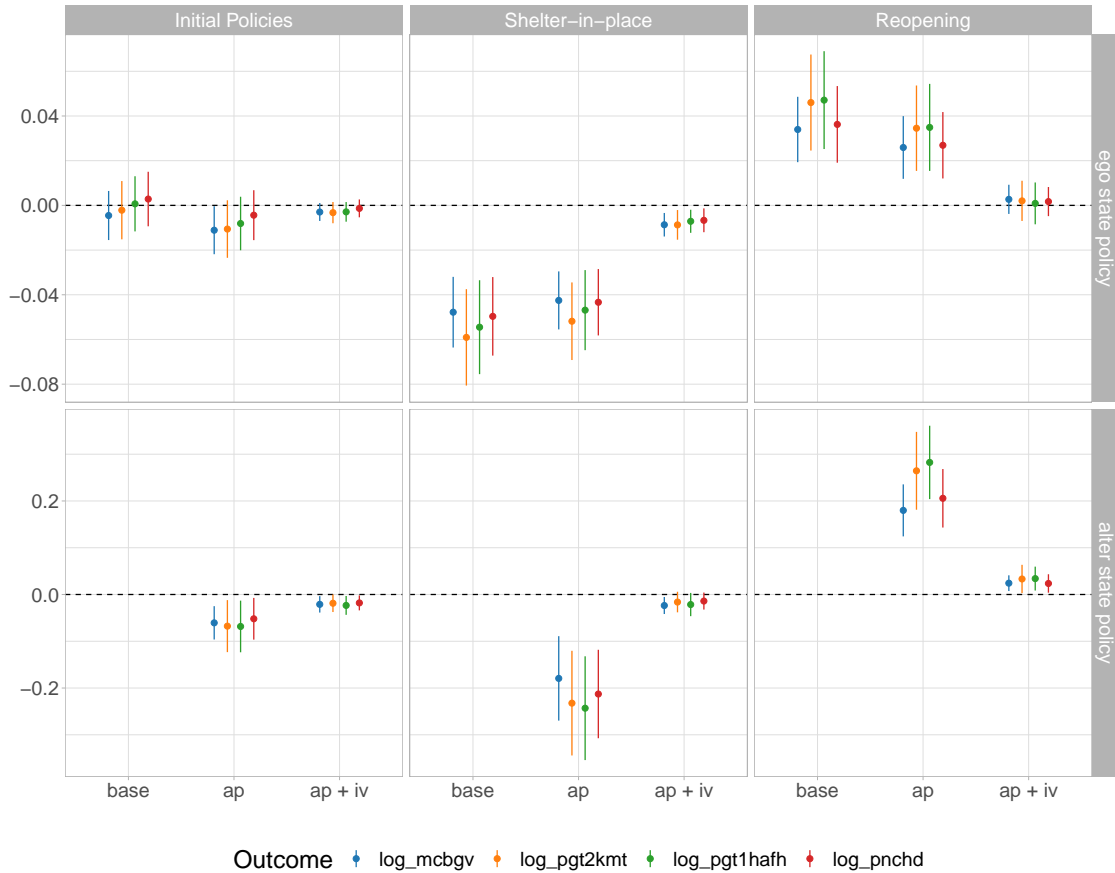
Here,  $\log(Y_{o \rightarrow d,t})$  refers to the log transformed cross-state mobility from an origin county  $o$  moving to a destination county in a different state  $d$  on date  $t$ . Origin and destination policies are denoted by  $D_{ot}^m$  and  $D_{dt}^n$  respectively, where  $m, n \in \{(ip), (sh), (ro)\}$ ;  $\alpha_{o \rightarrow d}$  and  $\tau_t$  correspond to directed dyad and time fixed effects; and  $\epsilon_{o \rightarrow d,t}$  represents the error term. Equation 3.4 models the impacts of origin and destination policy linearly whereas Equation 3.5 includes all possible interactions between origin and destination policies.

### 3.3 Results

Results from our base model indicate that statewide shelter-in-place orders reduced mobility within a state by 5-6% on average (Fig 3-2). Once reopened, mobility increased by 3-5%, returning to levels statistically indistinguishable from pre-pandemic levels. When accounting for alter state policy spillovers, these ego state policy estimates, while lower, are not significantly different from the base model estimates. Consistent with Holtz et al. (2020), we also find strong evidence of spillover effects in social distancing policies and reopenings. Specifically, our estimates indicate that once all alter states begin implementing some kind of social distancing policy, ego county mobility drops by 4-6%. If all alter states impose a lockdown, then mobility drops by an additional 15-20%. However, after all alter states begin reopening, an ego county's mobility increases by 19-32%.

Again consistent with Holtz et al. (2020), we find that endogenous peer behavior mediates the impact of both ego-state and alter-state policy. After accounting for alter state mobility behavior, the coefficient estimates of ego-state closures and reopenings move closer

Figure 3-2: Ego and Alter State Policy Impact on Mobility



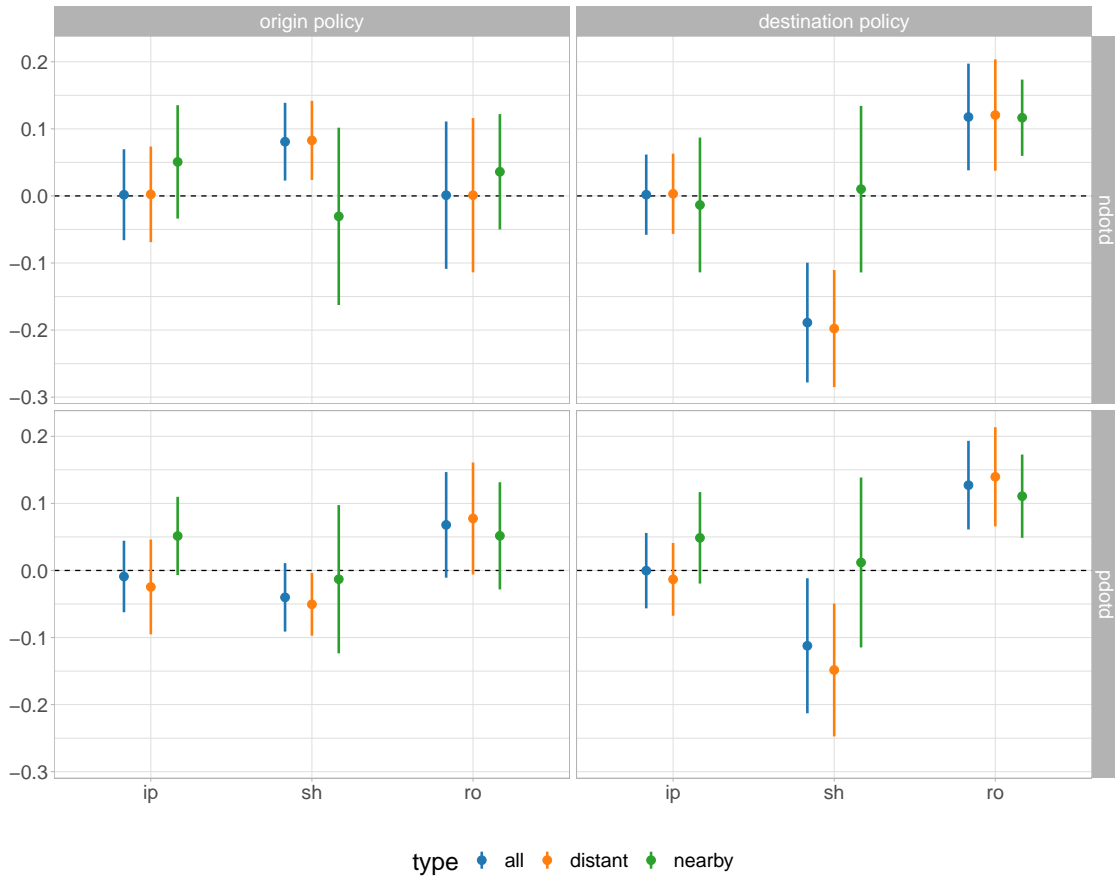
The top row of this Figure plots estimates of the ego state’s policy parameters:  $\delta_{(ip)}$ ,  $\delta_{(sh)}$ , and  $\delta_{(ro)}$  from left to right. Meanwhile, the bottom row corresponds to the alter states’ policy parameters:  $\gamma_{(ip)}$ ,  $\gamma_{(sh)}$ , and  $\gamma_{(ro)}$  again from left to right. Within each column the x-axis denotes the the model specification used to generate the estimates: “base” which corresponds to Eq. 3.1, “ap” which corresponds to Eq. 3.2, and “ap + iv” which corresponds to Eq. 3.3. 95% confidence errors are computed using state-clustered standard errors. Due to the coding of the policy variables, each of these parameters interpreted as the marginal effect of moving into the corresponding policy period. The “base” and “ap” estimates are produced using weighted least squares, with weights determined by county population. The “ap + iv” estimates are produced using two stage weighted least squares (again with county population weights), where alter state endogenous peer behavior is instrumented for with using alter state weather. Different colors correspond to the 4 main mobility outcomes: log\_mcbgv (blue), log\_pgt2kmt (orange), log\_pgt1hafh (green), and log\_pnchd (red).

to 0, though they are not significantly different from estimates produced by Equations 3.1 and 3.2. On the other hand, the changes in the coefficient estimates of alter state policy are much more dramatic, with both the impacts of alter-state closures and reopenings becoming statistically indistinguishable from zero. The estimate of the peer effect coefficient itself is quite significant, ranging from 1.8-2.2, meaning that a 1% increase or decrease in all out-of-state peer mobility causes between a 1.8-2.2% increase or decrease in ego county mobility.

Our cross-state mobility analysis also shows clear evidence of significant cross-state policy spillovers (Figs 3-3). Coefficient estimates produced by Equation 3.4 indicate that while the impacts of origin policies are not statistically distinguishable from 0, destination closures and reopenings produce large effects. Specifically, destination counties under statewide shelter-in-place orders receive 8-14% less cross-state traffic compared to pre-pandemic levels. These estimates also exhibit notable heterogeneity as travel from “distant” counties more than 100km away decreases by 13-18%, while there is no measurable impact on travel from “nearby” counties less than 100km away. As expected, reopenings boost travel to destination counties, by 12-13%, with no detectable differences between nearby or distant counties.

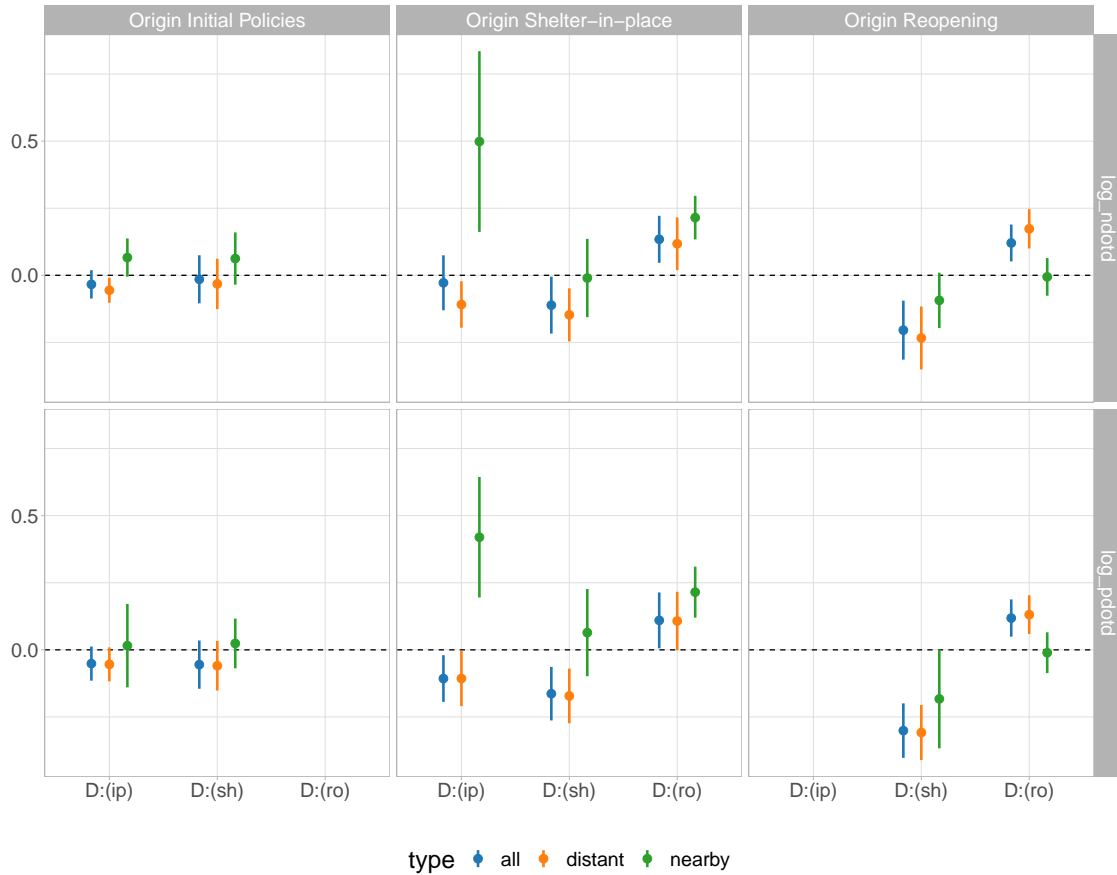
The interactions between origin and destination potentially provide the richest results (Fig 4.). When origin counties are in their initial policy period, rolling out social distancing policies but before shelter-in-place orders are imposed, destination policy does not seem to impact cross state mobility. In contrast, once an origin county is subject to a statewide shelter-in-place order, travel to distant counties decreases by 10% if they’ve only just started to implement social distancing policies. But, travel to nearby counties increases by 52-65% if those counties have not yet implemented a shelter-in-place order. Once destinations implement a shelter-in-place orders, distant cross-state travel decreases by 14-16%, with no detectable effect on nearby cross-state travel. When a destination reopens and an origin is still locked down, people from both nearby and distant counties not in that state flood into the destination, by 11-12% from nearby counties and 24% from distant counties. If the origin has reopened, but the destination has yet to reopen, travel to both nearby and distant destination counties is unsurprisingly depressed by 9-17% and 21-27% respectively.

Figure 3-3: Origin and Destination State Policy Impact on Cross-State Mobility



This figure plots the results of from estimating Equation 3.4. The left column corresponds to the origin policy parameters  $\lambda_{(ip)}$ ,  $\lambda_{(sh)}$ , and  $\lambda_{(ro)}$ , while the the right column corresponds destination policy parameters  $\psi_{(ip)}$ ,  $\psi_{(sh)}$ , and  $\psi_{(ro)}$ . These policies are denoted across the x-axis ip, sh, and ro from left to right. The top row reflects policy effects on the number of devices moving from an origin to a destination (ndotd), estimated using OLS and the bottom row reflects policy effects on the proportion of origin devices moving from an origin to a destination (pdotd) estimated using WLS with weights proportional to origin county population. 3 different coefficients are presented corresponding to estimates using all pairs in blue, distant pairs (> 100km) in orange, and nearby pairs (< 100km) in green. 2-way origin and destination state clustered standard errors are used to compute 95% confidence intervals.

Figure 3-4: Origin and Destination State Policy Interactions



This figure plots the results of from estimating Equation 3.5. Each column corresponds to different origin policy periods: initial policies, shelter-in-place, and reopening from left to right. As with Fig. 3-3, the top row reflects policy effects on the number of devices moving from an origin to a destination (ndotd), estimated using OLS and the bottom row reflects policy effects on the proportion of origin devices moving from an origin to a destination (pdotd) estimated using WLS with weights proportional to origin county population. Different values along the x-axis of each column correspond to different destination policy periods: destination initial policies D:(ip), destination shelter-in-place D:(sh), and destination reopening D:(ro). Within each column, the marginal effects of each destination policy conditional on the origin policy are plotted. 3 different coefficients are presented corresponding to estimates using all pairs in blue, distant pairs (> 100km) in orange, and nearby pairs (< 100km) in green. 2-way origin and destination state clustered standard errors are used to compute 95% confidence intervals.

However, once a destination county reopens there a 14-19% increase in travel from distant origins, though no change in travel from nearby origins.

### 3.4 Conclusion

To our knowledge, our work is the first to study the impacts of reopenings on mobility behavior not only taking into account, but also explicitly estimating on cross-state spillovers and the mediation of cross-state policy effects by peer behaviors. However, this work is not without its limitations. First, there may be concerns about the representativeness of the Safegraph panel. Although the size of panel is sufficiently large to minimize concerns about sampling error, there might be systematic sampling bias as mobile device ownership significantly varies by age and income. While Safegraph has shown their panel to be geographically consistent with the US Census population estimates, it is still not clear whether certain demographics are over- or under-represented as no device-level demographic data is collected by Safegraph. Second, our analysis does not explore the impacts of specific reopening policies (i.e. resuming restaurant dine-in service or lifting gathering restrictions) and rather measures average changes in mobility behavior across entire policy periods. Third, our analysis only captures variation in state-level closures and reopenings and ignores the relatively few instances that local or county level policy may differ from the state. We encourage such analysis in future work because disputes between states and localities may further thwart efforts to reduce mobility and control the pandemic. Fourth, our analysis is restricted to mobility outcomes and purposefully avoids making extrapolations to health outcomes like morbidity and mortality. While it is widely believed that reductions in mobility drive reductions in new infections and their associated deaths (Kraemer et al. (2020)), rigorously establishing the causal chain from cross-state spillovers to infection rates and deaths is beyond the scope of this paper.

Despite these limitations, our work provides critical information for policy makers, especially given the importance of restricting mobility to prevent the spread of COVID-19. Though many countries have seemingly reopened safely, new hotspots may yet emerge, forcing governments to reimpose mobility restrictions. Our results show that it is cru-

cially important to take spillover effects into account when formulating national policy and for national and regional policies to coordinate policies across regions where spillovers are strong. For example, local stay-at-home orders may be far less effective than policy makers would hope when peers' states and counties remain reopened, due to travel and peer influence. Moreover, such closure policies may actual be counterproductive Jia et al. (2020), Chinazzi et al. (2020) as they can encourage those locked down regions to flee to reopened regions, potentially causing new hotspots to emerge. Our data demonstrates that such travel spillovers are not only systematic and predictable, but also large and thus meaningful to public health.

## 3.5 Materials and Methods

### 3.5.1 Data

#### Safegraph Data

Our primary measures of human mobility are constructed from data provided by Safegraph<sup>1</sup>, a San Francisco-based company that sells data related to points of interest that are relevant to businesses. Safegraph collects anonymized human geo-spatial data from a number of partner mobile applications that need to obtain affirmative opt-in consent from device users. We specifically make use of Safegraph's "Social Distancing Metrics" dataset<sup>2</sup> which provides daily measures of mobility behavior aggregated at the census block group level starting from January 1, 2020. To further preserve privacy, Safegraph applies a differential privacy algorithm to all metrics that it computes other than device\_count. We use this data to construct 4 measures of county level mobility and 2 measures of cross-county dyadic travel.

**County Level Mobility Measures** We create the following 4 county-level measures of mobility: mean census block group visted (mcgbv), proportion of devices with greater than 2km traveled (pgt2kmt), proportion of devices spending more than 1 hour away from home

---

<sup>1</sup><https://www.safegraph.com>

<sup>2</sup><https://docs.safegraph.com/docs/social-distancing-metrics>



(pgt1hafh) and proportion of not completely home devices (pnchd). A device's "home" location is assigned by determining its common nighttime location across a period of 6 weeks at an approximately 153 by 153m granularity. We construct mcgbv by first summing across the number of non home census block groups in the destination\_cbgs field. We aggregate this count to the county level and simply divide by the count-level sum of device\_count. To build pgt2kmt, we first appropriately sum across the bucketed\_distance\_traveled field at the census block group level, aggregate to the county level, and then divide by device\_count. pgt1hafh is constructed in a similar manner, except that we begin by first appropriately summing across the bucketed\_home\_dwelling\_time field. Lastly to generate pnchd, we simply take 1 minus the county level aggregates of completely\_home\_device\_count divided by device\_count. In our analysis we use the log transformed values of these different measures: log\_mcgbv, log\_pgt2kmt, log\_pgt1hafh, and log\_pnchd.

**Cross-county Dyadic Travel** We create the following 2 measures of cross-county dyadic travel: number of devices moving from origin to destination ndotd and proportion of origin devices moving from origin to destination pdotd. To construct both these measures, we first build the a directed dyad list by unrolling the destination\_cbgs field. We then aggregate to the origin county / destination county for each day to build ndotd. To get to pdotd, we simply divide ndotd by each day's county-level device\_count. This data is quite sparse as there is little travel between most county pairs on most days. For our analysis, we limit ourselves to directed dyads with at least some travel between them for each day in our dataset.

### **COVID-19 US State Policy Database (CUSP)**

Our policy data comes from the COVID-19 US State Policy Database Raifman et al. (2020) assembled by researchers at the Boston University School of Public Health. This database tracks all state-wide wide directives and mandates, but not recommendations. It keeps track of state policies like gathering bans, entertainment closures, business closures, shelter-in-place orders, reopenings, and more for all 50 states plus Washington DC. As of this writing, the latest update to the database was made on Aug. 5, 2020. As mentioned in the main paper, we avoid quantifying the impact of each policy individually, as there is simply not

enough data to generate reliable estimates for such a high-dimensional policy space. We instead consolidate our analysis down to 3 main policy periods: the period from the first statewide social distancing policy of any kind until a shelter-in-place order takes effect or the “initial policies” period (*ip*); the period in which a shelter-in-place order is in effect (or until reopening begins) or the “stay home” period (*sh*); and the period after reopening begins or the “reopening” period (*ro*).

### **Facebook Social Connectedness Index (SCI)**

The Social Connectedness Index (described in more detail in Bailey et al. (2018)) is provided to us as a part of Facebook’s Data for Good<sup>3</sup> Initiative. The Social Connectedness Index provides a measure of the connectedness between two counties (or NUTS3 regions outside the US) through friendship ties on Facebook. It is constructed from an anonymized snapshot of the global Facebook friendship graph of over 2.45 billion users. Specifically, the sci between two counties is computed as:

$$sci_{ij} = \frac{fb\_connections_{ij}}{fb\_users_i \times fb\_users_j} \quad (3.6)$$

The numerator,  $fb\_connections_{ij}$  is just the number of friendship ties that are empirically observed between users in county  $i$  and  $j$ , while the denominator is simply the product between the number of Facebook users that reside in county  $i$  ( $fb\_users_i$ ) and county  $j$  ( $fb\_users_j$ ). Therefore  $sci_{ij}$  can be interpreted as the probability that a friendship link exists between a random user that resides in  $i$  and a random user that resides in  $j$ . However, the Social Connectedness Index does not directly report this probability, and instead reports a `scaled_sci` measure that is directly proportional.

The version of the SCI we use for our analysis is based on snapshot taken on December 31, 2019. Rather than using the `scaled_sci` measure directly, we instead weight it by the population of the friend county (using 2018 estimates from the US Census) to capture the relative differences in the number of ties coming from alter counties. For example, the `scaled_sci` between New York City and Boulder, Colorado is relatively low due to the

---

<sup>3</sup><https://dataforgood.fb.com/>

population of Facebook users in NYC. However, the number of friendship links between the two counties is relatively high, again due to the fact of NYC's population.

### **Weather Data**

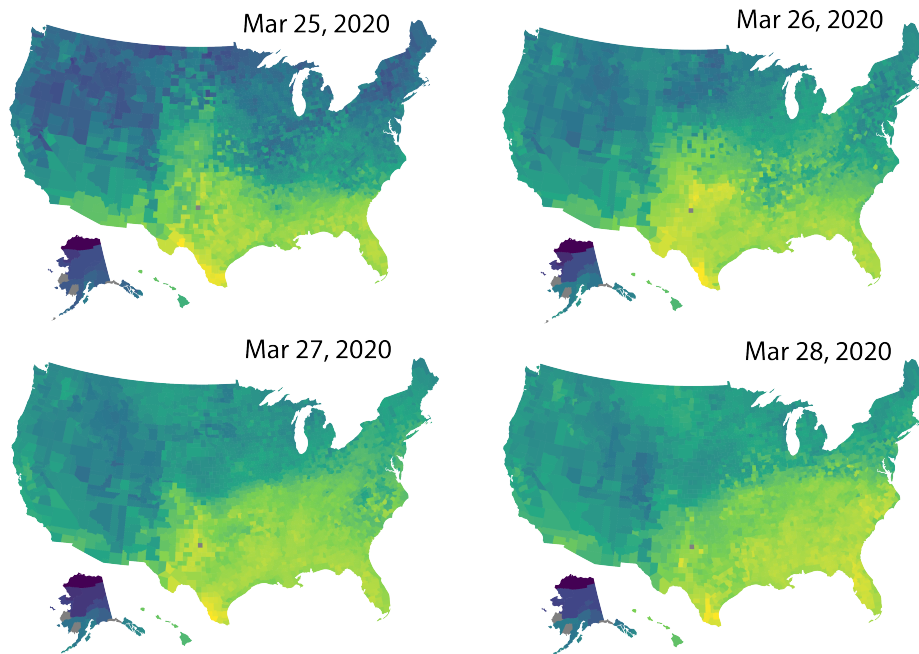
We use weather data from the National Oceanic and Atmospheric Association's (NOAA) Global Historical Climatology Network (GHCN). This data records daily observations of maximum temperature, precipitation, and other weather metrics for roughly 62,000 weather stations in the United States (more details can be found in Menne et al. (2012)). In order to construct measures of county level-weather, we begin by filtering out any weather stations that are missing maximum temperature or precipitation measures entirely. We use the geographic coordinates of each weather station, along with shapefiles specifying county borders to determine which weather stations are contained in which counties. For counties that contain three or more weather stations, we simply generate county precipitation and max temperature by its weather stations.

However, out of 3,233 counties in the US, 243 contain no weather stations and 967 have fewer than three weather stations. For each of these counties, we assign the nearest three stations within 100 kilometers of the county's centroid. To generate county-level measures we again just take the average of these assigned weather stations. Though this procedure assigns weather stations to every county, there may still be some missing values for either precipitation or max temperature. For these remaining county-day values, we find the nearest 3 weather stations within 100km without missing data on that particular day and take the average to fill in the missing data. This procedure achieves 99.9% coverage of county-days in our sample. We provide some visualizations of county-level maximum temperature and precipitation across the United States in Figures 3-5 and 3-6.

### **Difference-in-Difference Models**

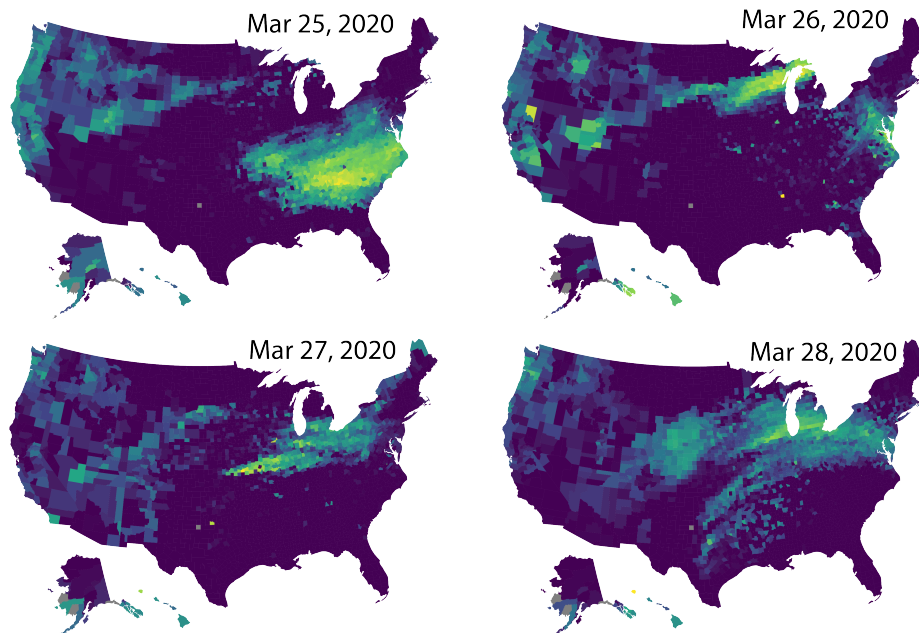
As mentioned in the main text, our base empirical approach to measuring the causal effects of COVID-19 policies and reopenings is difference-in-differences (DiD), much like Allcott et al. (2020), Painter and Qiu (2020), Chiou and Tucker (2020). The most basic version of requires multiple observations over time of at two groups, where a "treatment" group

Figure 3-5: County-Level Max Temperature across Four Consecutive Days



The maximum daily temperature (in degrees Celsius) at the county level over four consecutive days. The brighter color indicates higher maximum temperature.

Figure 3-6: County-Level Precipitation across Four Consecutive Days



Daily precipitation (in millimeters) at the county level over four consecutive days. The brighter color indicates higher precipitation.

is exposed to some treatment or intervention at some point but a “control” group does not. DiD works by comparing the change over time in the outcomes for the treatment group, relative to the change over time for the control group. The key assumption of DiD is parallel trends: while there may be substantive differences between treatment and the control groups, had the treatment group not received treatment, then the time series trends in outcomes would be the same.

Base Model For our analysis, our base model is a commonly used adaptation of the basic DiD model that simply employs unit and time fixed effects that allows for arbitrary or staggered variation in treatment timing across different units<sup>4</sup>. Specifically:

$$\log(Y_{it}) = \delta_{(ip)}D_{it}^{(ip)} + \delta_{(sh)}D_{it}^{(sh)} + \delta_{(ro)}D_{it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it}$$

Here,  $\log(Y_{it})$  refers one of the four aforementioned (Section 3.5.1) mobility outcomes,  $\log\_mcbgv$ ,  $\log\_pgt2kmt$ ,  $\log\_pgt1hafh$ , and  $\log\_pnchd$ , for county  $i$  on date  $t$ .  $D_{it}^{(ip)}$  is a binary indicator that takes the value of 1 once county  $i$ 's state has adopted some kind of social distancing policy. Similarly,  $D_{it}^{(sh)}$  takes the value on 1 once  $i$  is subject to a shelter-in-place order while  $D_{it}^{(ro)}$  switches to 1 once  $i$ 's state starts to reopen. Due to the way these binary indicators are coded (once they switch to 1, they do not switch back to 0), the associated parameters  $\delta_{(ip)}$ ,  $\delta_{(sh)}$ , and  $\delta_{(ro)}$  capture the marginal or additive effects conditional on the previous policy. More concretely,  $\delta_{(ip)}$  captures the average difference in mobility between pre-pandemic levels and while counties have just started implementing social distancing policies;  $\delta_{(sh)}$  is then the average difference in mobility during the “stay home” period compared to the  $\delta_{(ip)}$ ; and  $\delta_{(ro)}$  is then difference in mobility compared to  $\delta_{(ip)} + \delta_{(sh)}$ . Put differently, to estimate the differences in mobility from pre-pandemic levels and the reopening, we would need to sum  $\delta_{(ip)}$ ,  $\delta_{(sh)}$ , and  $\delta_{(ro)}$ .  $f(W_{it})$  is a term that captures the effect of local weather, which may be highly nonlinear, using a “double machine learning” (DML) approach Chernozhukov et al. (2018). This procedure is explained in much greater detail in Section 3.5.4.  $\alpha_i$  and  $\tau_t$  denote a set of county and time fixed effects respectively, while  $\epsilon_{it}$  represents the error term.

---

<sup>4</sup>Goodman-Bacon (2018) has shown that the estimand of such staggered DiD models decomposes into a weighted average of all possible two-group/two-period DiD estimators in the data.

## Alters Policies (AP) Model

In addition to parallel trends, DiD also assumes that the stable unit treatment value assumption (SUTVA) holds. Put more plainly, SUTVA simply states that the effect of treatment does not “spillover” to the control groups. However, Holtz et al. (2020) finds strong evidence for spillover effects in social distancing policy. We extend our base model to account such effects as follows:

$$\begin{aligned} \log(Y_{it}) = & \delta_{(ip)} D_{it}^{(ip)} + \delta_{(sh)} D_{it}^{(sh)} + \delta_{(ro)} D_{it}^{(ro)} + \\ & \gamma_{(ip)} D_{-it}^{(ip)} + \gamma_{(sh)} D_{-it}^{(sh)} + \gamma_{(ro)} D_{-it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it} \end{aligned}$$

The key differences here are the inclusion of  $D_{-it}^{(ip)}$ ,  $D_{-it}^{(sh)}$ , and  $D_{-it}^{(ro)}$  which denote the socially weighted average of alter state policy period indicators. More formally:

$$\begin{aligned} D_{-it}^{(ip)} &= \sum_j \omega_{j \rightarrow i} D_{-it}^{(ip)} \\ D_{-it}^{(sh)} &= \sum_j \omega_{j \rightarrow i} D_{-it}^{(sh)} \\ D_{-it}^{(ro)} &= \sum_j \omega_{j \rightarrow i} D_{-it}^{(ro)} \end{aligned}$$

Since we are focused on differences in state policy, we purposefully set weights  $\omega_{j \rightarrow i}$  to be equal to 0 if  $i$  and  $j$  belong to the same state. For counties belonging to different states however, the weights are defined as follows:

$$\omega_{j \rightarrow i} = \frac{\text{scaled\_sci}_{ij} * n_j}{\sum_k \text{scaled\_sci}_{ik} * n_k} : \text{state}_i \neq \text{state}_j, \text{state}_i \neq \text{state}_k$$

Where  $n_j$  is the 2018 US Census estimated population of county  $j$ . Similar to the  $\delta$  parameters,  $\gamma_{(ip)}$ ,  $\gamma_{(sh)}$ , and  $\gamma_{(ro)}$  are also considered marginal or additive effects. One key difference however is that the alter policy variables  $D_{-it}^{(ip)}$ ,  $D_{-it}^{(sh)}$ , and  $D_{-it}^{(ro)}$  are not binary indicators, meaning that each  $\gamma$  is interpreted as the marginal effect if all other states move onto that particular policy period.

## Pre-Trends

Before moving on to our results, we first show evidence that parallel trends is satisfied. First in Figure 1A of the main text, it can be seen that the time series of the various state quintile groups by reopening data generally all follow the same trend, especially in the pre-pandemic period from Jan. 1, 2020 to Feb. 29, 2020. In Fig. 3-7, we plot the average residuals of our 4 main mobility dependent variables after partialing out county and date fixed effects focusing specifically on the pre-pandemic period of our data. Looking at the residuals, it is difficult to discern any systematic trend amongst the different groups. In fact, each series looks essentially like a random walk further supporting the assumption of identical pre-trends.

Figure 3-7: DiD Pre-Period Residuals



The average residuals across counties grouped by state-level reopening date quintiles after partialing out county and date fixed effects from Jan 1, 2020. to Feb. 29, 2020.

## Results

As the population sizes of various counties are quite heterogeneous, we estimate Equations 3.1, 3.2, and 3.3 using weighted least squares, where observations are weighted according to their populations. This then allows us to interpret our results as averages across human mobility rather than averages across county mobility<sup>5</sup>. The results of estimating Equations 3.1 and 3.2 are presented in Tables 3.1 and 3.2 respectively.

Table 3.1: Base Model Results

	<i>Dependent variable:</i>			
	log_mcbgv (1)	log_pgt2kmt (2)	log_pgt1hafh (3)	log_pnchd (4)
Ego State Initial Policies	-0.005 (0.006)	-0.002 (0.007)	0.001 (0.006)	0.003 (0.006)
Ego State Shelter-in-place	-0.048*** (0.008)	-0.059*** (0.011)	-0.055*** (0.011)	-0.050*** (0.009)
Ego State Reopening	0.034*** (0.007)	0.046*** (0.011)	0.047*** (0.011)	0.036*** (0.009)
Observations	470,106	470,106	470,106	470,106
R <sup>2</sup>	0.049	0.043	0.043	0.048
Adjusted R <sup>2</sup>	0.043	0.037	0.037	0.042

*Note:* State-Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Consistent with previous work, both models indicate significant decreases in mobility during statewide shelter-in-place orders. Specifically, both models' point estimates indicate mobility decreases of 4-6%, relative to both the initial policy period (though this basically extends to pre-pandemic levels given that ego state initial policy coefficients are generally not statistically significant and close to 0). However, as we should expect, once a state begins reopening, mobility starts to increase. Again both models' here are quite quantita-

<sup>5</sup>Consider the following example of 2 counties, one with 1000 people and one with 9000 people. Suppose that a shelter-in-place order reduces mobility of county 1 by 10% and county 2 by 20% then the unweighted regression produce a shelter-in-place impact of 15%. In contrast, the weighted regression would produce a shelter-in-place impact of 19% which corresponds to the average decrease in mobility across the population.



Table 3.2: Alters' Policies Model Results

	<i>Dependent variable:</i>			
	log_mcbgv (1)	log_pgt2kmt (2)	log_pgt1hafh (3)	log_pnchd (4)
Ego State Initial Policies	-0.011** (0.005)	-0.011 (0.006)	-0.008 (0.006)	-0.004 (0.006)
Ego State Shelter-in-place	-0.043*** (0.007)	-0.052*** (0.009)	-0.047*** (0.009)	-0.043*** (0.007)
Ego State Reopenings	0.026*** (0.007)	0.035*** (0.010)	0.035*** (0.010)	0.027*** (0.008)
Alter States Initial Policies	-0.061*** (0.018)	-0.068** (0.028)	-0.068** (0.028)	-0.052** (0.023)
Alter States Shelter-in-place	-0.179*** (0.046)	-0.232*** (0.057)	-0.243*** (0.056)	-0.213*** (0.048)
Alter States Reopenings	0.180*** (0.028)	0.264*** (0.042)	0.282*** (0.040)	0.206*** (0.032)
Observations	470,106	470,106	470,106	470,106
R <sup>2</sup>	0.087	0.085	0.090	0.095
Adjusted R <sup>2</sup>	0.082	0.080	0.085	0.090

*Note:* State-Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

tively consistent, indicating a 3-5% increases in mobility. Overall, our estimates indicate that once reopened, mobility is slightly depressed compared to pre-pandemic levels (computed by summing the coefficients across all 3 policy periods), by 1-1.5% according to Table 3.1 and 2-2.5% according to 3.2<sup>6</sup>. Consistent with Holtz et al. (2020), our results indicate significant cross-state policy spillovers. Our estimates suggest that once all alter states initiate social distancing, county mobility drops by 4-5%. If all other states implement a shelter-in-place order, mobility will drop even further, by an additional 16-22%. In contrast, alter states reopenings have a generally larger but opposite effect, increasing county-level mobility by 20-32%.

### 3.5.2 Instrumental Variables Model

In addition to uncovering strong policy spillover effects, Holtz et al. (2020) showed that the spillovers were largely mediated by endogenous peer behavior. To investigate if such behavior extends to reopenings we estimate the following specification:

$$\log(Y_{it}) = \beta \log(Y_{-it}) + \delta_{(ip)} D_{it}^{(ip)} + \delta_{(sh)} D_{it}^{(sh)} + \delta_{(ro)} D_{it}^{(ro)} + \gamma_{(ip)} D_{-it}^{(ip)} + \gamma_{(sh)} D_{-it}^{(sh)} + \gamma_{(ro)} D_{-it}^{(ro)} + f(W_{it}) + \alpha_i + \tau_t + \epsilon_{it}$$

Compared to Equation 3.2 above, the only difference is the inclusion of  $\log(Y_{-it})$  or the log transformed weighted average of alter states' mobility outcomes. As with the alter-policy variables above,  $Y_{-it} = \sum_j \sum_j \omega_{j \rightarrow i} Y_{jt}$ . Despite the seemingly minor addition, our DiD framework is theoretically insufficient at producing causal estimates of endogenous peer behavior due to challenges posed by simultaneity (aka the "reflection problem" Manski (1993)), correlated exposure to unobserved confounding factors, and homophily McPherson et al. (2001).

To address this issue, we shift our approach to instrumental variables (IV), an approach widely used in social sciences to address endogeneity concerns. To provide a basic overview of how IV functions, consider the following simple scenario where  $Y = \beta_0 + \beta X + \epsilon$ , but  $X$  is correlated with the error term  $\epsilon$ . In such a setting, simply regress-

---

<sup>6</sup>The base model estimates are not statistically distinguishable from pre-pandemic levels, but the AP model estimates are.

ing  $Y$  on  $X$  will produce a biased estimate of  $\beta$ . A third variable  $Z$  can be considered an “instrument” for  $X$  if it meaningfully impacts  $X$  and is (conditionally) uncorrelated with the error term  $\epsilon$ . These two restrictions are known as relevancy and exclusion respectively. If  $Z$  does indeed qualify as an instrument, consistent estimates of  $\beta$  can be recovered via a 2-stage least squares (2SLS) procedure where  $X$  is first regressed on  $Z$  to produce fitted values  $\hat{X} = E[X|Z]$  and then  $Y$  is regressed on these fitted values of  $X$ .

In our case, our IV strategy exploits exogenous shocks alters’ mobility behavior stemming from variation in alters’ weather. Similar weather IV approaches have been used to measure emotional contagion Coviello et al. (2014b), peer effects in exercise behavior Aral and Nicolaides (2017b), and social spillovers in online news consumption Aral and Zhao (2020). Most relevantly, Holtz et al. (2020) also leveraged weather instruments to estimate the impact of endogenous peer behavior.

## Weather Instruments and First Stage

We construct our instruments from the county-level weather dataset that we constructed described in Section 3.5.1. In order to construct our alter county instruments, we first we first construct a sequence of county-level indicator variables that take a value of 1 if the amount of rainfall in county  $i$  on date  $t$  falls within or exceeds a specific precipitation decile, conditional on non-zero precipitation<sup>7</sup>. We generate a similar sequence of indicator variables for maximum temperature as well. To avoid perfect multicollinearity, we remove only first max temperature decile indicator. We need not remove the first precipitation decile as these deciles are computed only for non-zero precipitation, meaning that “no precipitation” functions as the base case. We then construct 19 alter-state weather measures again by taking the socially weighted averages of each of 10 precipitation and 9 max temperature deciles<sup>8</sup> to form the alter state weather instruments ( $W_{-it} = V_{-it}^{\text{prcp},1}, \dots, Q_{-it}^{\text{prcp},10}, Q_{-it}^{\text{tmax},2}, \dots, Q_{-it}^{\text{tmax},10}$ ).

<sup>7</sup>For example  $Q_{it}^{\text{prcp},1} = \mathbf{1}(\text{prcp}_{it} > q) : q = \arg_x Pr(\text{prcp}_{it} \geq x) = 0$ ,  $Q_{it}^{\text{prcp},2} = \mathbf{1}(\text{prcp}_{it} \geq q) : q = \arg_x Pr(\text{prcp}_{it} \geq x) = 0.5$ , etc. It is also worth noting that this construction means that if  $V_{it}^{\text{prcp},k} = 1$ , then  $V_{it}^{\text{prcp},j} = 1 : j < k$ .

<sup>8</sup>More formally:  $Q_{-it}^{\text{prcp},k} = \sum_j \omega_{j \rightarrow i} * Q_{jt}^{\text{prcp},k}$  and  $Q_{-it}^{\text{tmax},k} = \sum_j w_{ij} * Q_{jt}^{\text{tmax},k}$ .

This leads to the following first-stage specification:

$$\log(Y_{-it}) = \delta_{(ip)}^{fs} D_{it}^{(ip)} + \delta_{(sh)}^{fs} D_{it}^{(sh)} + \delta_{(ro)}^{fs} D_{it}^{(ro)} + \gamma_{(ip)}^{fs} D_{-it}^{(ip)} + \gamma_{(sh)}^{fs} D_{-it}^{(sh)} + \gamma_{(ro)}^{fs} D_{-it}^{(ro)} + \sum_{d=1}^{10} (\zeta_d^{\text{prcp}} Q_{-it}^{\text{prcp},d} + \zeta_d^{\text{tmax}} Q_{-it}^{\text{tmax},d}) + \alpha_{-i} + \tau_t + \nu_{-it} \quad (3.7)$$

In theory, we should be able to use these alter-state weather variables to instrument for alter-state peer behavior. However, common major concern with weather instruments is that geographically proximate locations tend to have similar weather. Theoretically, this should not pose an issue: even if “alters’ weather” is highly correlated with “own weather,” it should be conditionally ignorable so long as the effects of “own weather” are controlled for. However, this can be quite challenging due to the potential nonlinearities and interactions in the impact of weather. For instance, the likelihood of going outside is going to change much more going from 0mm to 1mm of rain relative to going from 20mm to 21mm. In a similar vein, the impact of rain is likely to be very different if it is a comfortable day outside than if it is cold and dreary. Such complexities may therefore cause a “technical” violation of conditional ignorability since alters’ weather may be providing additional information about own weather that cannot be captured linearly. As such, we adopt a flexible DML procedure to model the impact of weather that we explain in greater detail in Section 3.5.4.

## Results

The results of estimating Equation 3.3 can be found in Table 3.3. Included in this table are the first-stage F-statistics testing the relevancy of the instruments. As can be seen in the table, the F-stats range from 60-70 indicating that we do not have a weak instruments problem. We also estimate Equation 3.3 using limited information maximum likelihood (LIML). These results can be found in Table 3.4.

Across our four main outcomes and both 2SLS and LIML results, several clear trends emerge. First, Ego State shelter-in-place still has a statistically significant negative effect on mobility. While these estimates are smaller in magnitude across the board, they are not quite statistically distinguishable from the estimates found in Table 3.2. Second, again consistent with Holtz et al. (2020), we find strong evidence of endogenous peer effects. Our

Table 3.3: IV 2SLS Results

	<i>Dependent variable:</i>			
	log_mcbgv	log_pgt2kmt	log_pgt1hafh	log_pnchd
	(1)	(2)	(3)	(4)
Ego State Initial Policies	-0.005 (0.006)	-0.002 (0.008)	0.0003 (0.007)	0.002 (0.006)
Ego State Shelter-in-Place	-0.032*** (0.005)	-0.036*** (0.007)	-0.030*** (0.006)	-0.028*** (0.005)
Ego State Reopening	0.014** (0.007)	0.014 (0.010)	0.013 (0.010)	0.012* (0.007)
Alter States Initial Policies	-0.037** (0.018)	-0.032 (0.025)	-0.041* (0.023)	-0.037* (0.020)
Alter States Shelter-in-Place	0.018 (0.025)	0.045 (0.035)	0.048 (0.036)	0.029 (0.027)
Alter States Reopening	0.039** (0.018)	0.031 (0.033)	0.031 (0.030)	0.029 (0.023)
Endogenous Alter States Behavior	1.835*** (0.190)	2.228*** (0.235)	2.193*** (0.217)	2.073*** (0.193)
First-Stage F	71.879	60.895	62.580	72.430
Observations	470,106	470,106	470,106	470,106
R <sup>2</sup>	0.476	0.402	0.467	0.492
Adjusted R <sup>2</sup>	0.473	0.398	0.464	0.489

*Note:* State-Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Table 3.4: IV LIML Results

	<i>Dependent variable:</i>			
	log_mcbgv (1)	log_pgt2kmt (2)	log_pgt1hafh (3)	log_pnchd (4)
Ego State Initial Policies	-0.002 (0.006)	-0.001 (0.008)	0.002 (0.007)	0.004 (0.006)
Ego State Shelter-in-Place	-0.027*** (0.005)	-0.032*** (0.008)	-0.025*** (0.007)	-0.024*** (0.006)
Ego State Reopening	0.009 (0.007)	0.009 (0.010)	0.007 (0.010)	0.007 (0.007)
Alter States Initial Policies	-0.026 (0.019)	-0.024 (0.027)	-0.034 (0.025)	-0.032 (0.021)
Alter States Shelter-in-Place	0.103*** (0.020)	0.110*** (0.025)	0.122*** (0.028)	0.102*** (0.023)
Alter States Reopening	-0.023 (0.018)	-0.024 (0.031)	-0.033 (0.025)	-0.025 (0.022)
Endogenous Alter States Behavior	2.632*** (0.022)	2.747*** (0.023)	2.751*** (0.021)	2.705*** (0.021)
Observations	470,106	470,106	470,106	470,106
R <sup>2</sup>	0.516	0.414	0.484	0.516
Adjusted R <sup>2</sup>	0.513	0.410	0.481	0.513

*Note:* State-Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

estimates indicate that a 1% increase or drop in mobility by all peers in different states will cause mobility in an ego county to increase or drop by approximately 2% using the 2SLS estimate. The LIML estimates are even more extreme, indicating that a 1% change in alter state peer behavior we lead to 2.7% in ego county mobility behavior. As with Holtz et al. (2020), once we introduce endogenous peer behavior into our model, the alter state policy coefficients move significantly closer to 0. In the case of the 2SLS results, most of these coefficients are no longer statistically different from 0.

### 3.5.3 Dyad-level Difference-in-differences

In this section, we explore the potential impacts of both origin and destination policies on cross-state travel. As our policy variation is at the state level, we specifically focus on cross-state county-pairs. Here we use the following 2 model specifications:

$$\log(Y_{o \rightarrow d,t}) = \sum_m \lambda_m D_{ot}^m + \sum_n \psi_n D_{dt}^n + \alpha_{o \rightarrow d} + \tau_t + \epsilon_{o \rightarrow d,t}$$

$$\log(Y_{o \rightarrow d,t}) = \sum_m \lambda_m D_{ot}^m + \sum_n \psi_n D_{dt}^n + \sum_m \sum_n \pi_{m,n} (D_{ot}^m * D_{dt}^n) + \alpha_{o \rightarrow d} + \tau_t + \epsilon_{o \rightarrow d,t}$$

$\log(Y_{o \rightarrow d,t})$  refers to the one of our log-transformed cross-county mobility metrics (described in Section 3.5.1 above) based on the number of devices identified with a home in an origin county  $o$  and stopping for at least one minute in a destination county  $d$ .  $D_{ot}^m$  and  $D_{dt}^n$  denote origin and destination policies respectively, where  $m, n \in \{(ip), (sh), (ro)\}$ . As with Equation 3.1 above, these policy are binary indicators that flip to 1 once the corresponding policy period begins. This means that for Equation 3.4 the associated parameters  $\lambda_{(ip)}$ ,  $\lambda_{(sh)}$ ,  $\lambda_{(ro)}$ ,  $\gamma_{(ip)}$ ,  $\gamma_{(sh)}$ , and  $\gamma_{(ro)}$  are interpreted as the marginal effect, but only within each parameter family<sup>9</sup>.

In Equation 3.5, which models all potential interactions between origin and destination policies, the  $\lambda$ s capture the marginal effects of origin policy if the destination is in the pre-policy period. Likewise, the  $\gamma$ s capture the marginal effects of destination policy if the origin is in the pre-policy period. The interaction parameters  $\pi_{m,n}$  are then the additional

---

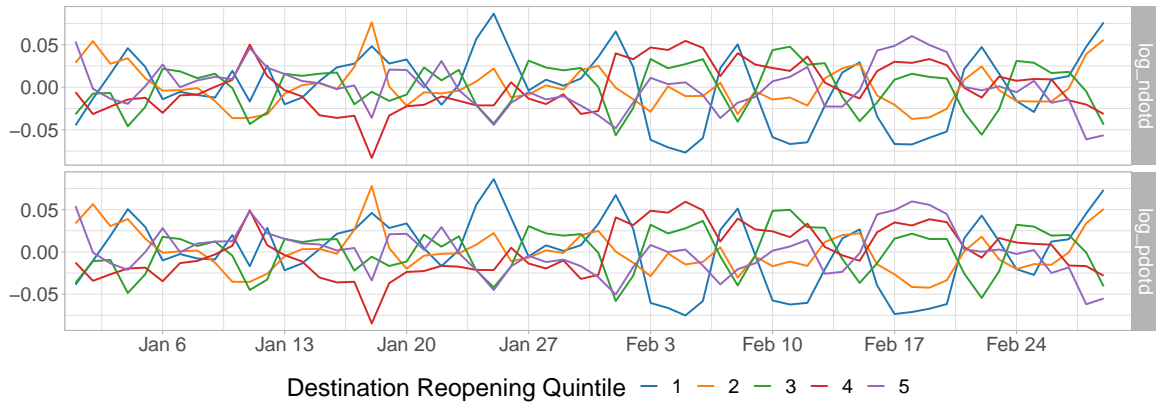
<sup>9</sup>That is to say each successive  $\lambda$  is marginal to only the previous  $\lambda$ 's and each successive  $\gamma$  is marginal to the previous  $\gamma$ s

marginal effect above and beyond the sum of preceding parameters. That is,  $\pi_{(sh),(rop)}$  is the additional marginal effect when compared to  $\lambda_{(ip)} + \lambda_{(sh)} + \gamma_{(ip)} + \gamma_{(sh)} + \gamma_{(ro)} + \pi_{(ip),(ip)} + \pi_{(ip),(sh)} + \pi_{(ip),(ro)} + \pi_{(sh),(ip)} + \pi_{(sh),(sh)}$ . Lastly,  $\alpha_{o \rightarrow d}$  and  $\tau_t$  denote directed dyad and time fixed effects, and  $\epsilon_{o \rightarrow d,t}$  captures the error term.

## Pre Trends

As with Section 3.5.1, our analysis is based on a difference-in-differences approach. Naturally, this means that it is important to verify that there aren't any systematic differences in pre-trends. In Figure 3-8, we plot the average residuals of our 2 cross-state mobility variables after partialing out dyad and date fixed effects for the period between Jan. 1, 2020 and Feb. 29, 2020. as with above, it is difficult to find any systematic trend amongst the different groups organized around destination county's statewide reopening start. Again, each series looks a like a mean 0 random walk suggesting that parallel trends does indeed hold.

Figure 3-8: Dyad Pre-Period Residuals



The average residuals across dyads grouped by destination state-level reopening date quintiles after partialing out dyad and date fixed effects from Jan 1, 2020. to Feb. 29, 2020.

## Results

The results from estimating Equations 3.4 and 3.5 are displayed in Tables 3.5 and 3.6. While  $\log\_ndotd$  is estimated using OLS,  $\log\_pdotd$  is estimated using WLS where weights are proportional to origin county population.



Table 3.5: Dyadic Travel Results

	<i>Dependent variable:</i>					
	log_ndotd			log_pdotd		
	(1) All	(2) Nearby	(3) Distant	(4) All	(5) Nearby	(6) Distant
Origin Initial Policies	-0.009 (0.027)	0.012 (0.025)	-0.016 (0.035)	0.002 (0.034)	0.085 (0.051)	0.0005 (0.035)
Origin Shelter-in-Place	-0.040 (0.026)	-0.117*** (0.034)	-0.012 (0.026)	0.081*** (0.029)	-0.137*** (0.045)	0.090*** (0.029)
Origin Reopening	0.068* (0.040)	0.106*** (0.036)	0.065 (0.044)	0.001 (0.055)	0.090** (0.044)	0.002 (0.061)
Destination Initial Policies	-0.0003 (0.028)	0.010 (0.020)	-0.004 (0.029)	0.002 (0.030)	0.036 (0.030)	0.002 (0.031)
Destination Shelter-in-Place	-0.112** (0.051)	-0.095*** (0.032)	-0.110* (0.057)	-0.189*** (0.045)	-0.061 (0.046)	-0.193*** (0.045)
Destination Reopening	0.127*** (0.033)	0.165*** (0.025)	0.130*** (0.032)	0.118*** (0.040)	0.185*** (0.029)	0.124*** (0.042)
Observations	3,107,468	708,708	2,398,760	3,107,468	708,708	2,398,760
R <sup>2</sup>	0.764	0.905	0.664	0.853	0.928	0.819
Adjusted R <sup>2</sup>	0.762	0.905	0.662	0.853	0.927	0.818

Note: 2-way Origin State and Destination State Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

### 3.5.4 Double Machine Learning Weather Controls

To flexibly control for the impact of weather, we employ a “double machine learning” (DML) procedure Chernozhukov et al. (2018). This approach is designed to estimate and draw inferences on a low-dimensional parameter in the presence of high-dimensional nuisance parameters. Consider the following “canonical example” from Chernozhukov et al. (2018) which we reproduce here:

$$Y = D\theta_0 + g_0(Z) + U, \quad E[U|D, Z] = 0$$

$$D = m_0(Z) + V, \quad E[V|Z] = 0$$

$Y$  denotes the outcome,  $D$  is a policy or treatment variable,  $\theta_0$  is the low-dimensional parameter of interest,  $Z$  is a high-dimensional vector of covariates ( $g_0(Z)$  can be considered to be the high-dimensional nuisance parameter), and  $U$  and  $V$  are the errors. The basic intuition behind DML is that  $g_0(\cdot)$  and  $m_0(\cdot)$  can be estimated using non-parametric sta-

Table 3.6: Dyadic Travel With Interactions

	<i>Dependent variable:</i>					
	log(ndotd)			log(pdotd)		
	(1) All	(2) Nearby	(3) Distant	(4) All	(5) Nearby	(6) Distant
Origin Pre-Policies × Destination Initial Policies	0.038 (0.034)	0.034 (0.026)	0.032 (0.034)	0.055* (0.030)	0.053 (0.034)	0.055* (0.030)
Origin Pre-Policies × Destination Shelter-in-Place	-0.135* (0.070)	-0.059 (0.042)	-0.140 (0.085)	-0.201*** (0.030)	-0.132** (0.050)	-0.196*** (0.035)
Origin Pre-Policies × Destination Reopening	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Origin Initial Policies × Destination Pre-Policies	0.024 (0.029)	0.041 (0.027)	0.013 (0.039)	0.058 (0.042)	0.100* (0.053)	0.058 (0.044)
Origin Initial Policies × Destination Initial Policies	-0.072** (0.027)	-0.054** (0.026)	-0.067** (0.030)	-0.106*** (0.026)	-0.048 (0.031)	-0.108*** (0.028)
Origin Initial Policies × Destination Shelter-in-Place	0.119** (0.054)	0.024 (0.058)	0.141** (0.070)	0.148*** (0.041)	0.121** (0.049)	0.143*** (0.043)
Origin Initial Policies × Destination Reopening	0.068 (0.052)	0.154*** (0.046)	0.062 (0.050)	0.141** (0.068)	0.267*** (0.084)	0.147** (0.071)
Origin Shelter-in-Place × Destination Pre-Policies	-0.006 (0.035)	-0.364*** (0.115)	0.058 (0.043)	0.192*** (0.067)	-0.399*** (0.075)	0.193*** (0.071)
Origin Shelter-in-Place × Destination Initial Policies	0.004 (0.041)	0.290*** (0.102)	-0.029 (0.043)	-0.057 (0.058)	0.293*** (0.100)	-0.048 (0.059)
Origin Shelter-in-Place × Destination Shelter-in-Place	-0.094** (0.038)	-0.110** (0.045)	-0.098** (0.037)	-0.111*** (0.036)	-0.060 (0.050)	-0.111*** (0.039)
Origin Shelter-in-Place × Destination Reopening	0.064 (0.041)	0.070** (0.034)	0.065 (0.042)	-0.032 (0.053)	0.003 (0.061)	-0.032 (0.054)
Origin Reopening × Destination Pre-Policies	0.175** (0.069)	0.143** (0.058)	0.198** (0.079)	0.124 (0.101)	0.256** (0.122)	0.129 (0.109)
Origin Reopening × Destination Initial Policies	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Origin Reopening × Destination Shelter-in-Place	-0.110** (0.043)	0.019 (0.039)	-0.147*** (0.048)	-0.142** (0.067)	-0.075 (0.103)	-0.147** (0.069)
Origin Reopening × Destination Reopening	-0.010 (0.045)	-0.097** (0.038)	0.003 (0.052)	0.008 (0.042)	-0.164** (0.062)	0.009 (0.045)
Observations	3,107,468	708,708	2,398,760	3,107,468	708,708	2,398,760
R <sup>2</sup>	0.764	0.905	0.665	0.854	0.928	0.819
Adjusted R <sup>2</sup>	0.763	0.905	0.663	0.853	0.928	0.818

Note: 2-way Origin State and Destination State Clustered Standard Errors are reported. \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

tistical methods (aka machine learning) and then “partialed out” Frisch and Waugh (1933) from both  $Y$  and  $D$ . Then one simply regresses the residuals of the dependent variable on the residuals of the treatment variable in order to estimate  $\theta_0$ . In order to provide guarantees that key moment conditions are satisfied, the machine learning predictions needs to be orthogonalized which can be achieved via sample splitting. As such, the general double ML algorithm is as follows:

1. Split the dataset into  $K$  equal size partitions or “folds.” Let  $F_k, F_k^c : k \in 1, \dots, K$  denote each fold and its complement.<sup>10</sup>
2. Estimate  $g_0$  and  $m_0$  with some non-parametric statistical model of choice using only the observations in  $B_1^c$
3. Form residuals  $\tilde{Y} := Y - \hat{g}_0(Z)$  and  $\tilde{D} := D - \hat{m}_0(Z)$  only on observations in  $F_1$ .
4. Regress  $\tilde{Y}$  on  $\tilde{D}$  to obtain an estimate of  $\theta_0$ . Overall, this estimate can be thought of as function of  $F_1$  and  $F_1^c$ :  $\hat{\theta}_0(F_1, F_1^c)$ .
5. Repeat steps 2-4 for the the remaining  $K - 1$  folds
6. Form the final estimate of  $\theta_0$  by averaging across all estimates:  $\hat{\theta}_0^* = \frac{1}{K} \sum_k \hat{\theta}_0(F_k, F_k^c)$

In our case, we consider a county’s own weather to be the high-dimensional nuisance parameter, as we are not principally interested in identifying the effect of own weather on social distancing behavior. We use gradient boosted decision trees via XGBoost Chen and Guestrin (2016), a state-of-art machine learning algorithm, to estimate  $f(\cdot)$  in Equations 3.1, 3.2, and 3.3, as well as the effect of weather on any of the other variables included in our models. XGBoost is an ensemble method that works by fitting a series of forward stage-wise decision trees aimed to minimizing a specified loss function. To give a general idea of the basic procedure:

1. Fit an initial decision tree  $T_1$  that minimizes  $E [(Y - T_1(X))^2]$ , where  $Y$  is the outcome and  $X$  are the covariates or features.

---

<sup>10</sup>Suppose a dataset has 100 observations and is split into 5 block.  $B_1$  consists of observations 1-20 and  $F_1^c := F_2, F_3, F_4, F_5$  consists of the remaining observations 21-80.

2. Each successive tree is then fitted on the residuals of the previous state<sup>11</sup>:

$$T_n = \arg \min_T E[(Y - \sum_{i=1}^{n-1} T_i(X) - T(X))^2]$$

In order to prevent overfitting, this iterative process is stopped once out-of-sample predictive performance starts to decline.

As with many other machine learning algorithms, there are a number of hyperparameters that control this estimation procedure of XGBoost. In particular, we adjust:

- `tree_depth`: Controls the depth that each tree-based model is allowed to grow to. The deeper the tree, the more complex the model.
- `eta`: Controls the “learning rate” or step size of each model. One way to think of this parameter is as a form of regularization on each model step in order to prevent overfitting.
- `nrounds`: The maximum number of stages the fitting process is allowed to continue on for.

We fix `tree_depth` to 2 and `eta` = 0.5, but allow `nrounds` to run up to a maximum of 100. Then, for each individual variable, the optimal number of rounds (given our choice of `tree_depth` and `eta`) is determined via a cross-validation procedure for each variable individually<sup>12</sup>. Once the optimal `nrounds` is determined, we form the residuals for all our dependent variables and covariates by first partialing out the set of fixed effects and then following the DML approach described above.

### 3.5.5 Software

Data processing, analysis, and plotting was conducted in R Team et al. (2013) and Python Rossum (1995). `pandas` McKinney et al. (2010), `jsonlite` Ooms (2014), and various tidyverse libraries Wickham et al. (2019)—`dplyr`, `lubridate`, `readr`, `stringr`, `tidyr`, etc.—were used to

---

<sup>11</sup>To be more precise, the degree to which each successive tree contributes to the ensemble can be controlled via tuning hyperparameter called a learning rate. We provide a little bit more detail on this below.

<sup>12</sup>We note that it would be more optimal to do an exhaustive grid search across the entire hyperparameter space for each individual variable that needs to have the effect of weather partialled out. However, such a grid search would be extremely computationally expensive and would only yield very minor improvements in predictive accuracy.

process and prepare the data for analysis. Regression analysis was performed using lfe Gaure (2013). Other statistical analysis were conducted using xgboost Chen et al. (2019). doMC Analytics (2014) was used to parallelize computation. Tables were created using the stargazer package Hlavac (2015). Plots were generated using ggplot2 Wickham (2016b), viridis, ggsci, and urbnmapr.



# Bibliography

- Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, David Y Yang. 2020. Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. Tech. Rep. 26946, National Bureau of Economic Research.
- Analytics, Revolution. 2014. domc: Foreach parallel adaptor for the multicore package. *R package version 1*(3).
- Andrews, Michelle, Xueming Luo, Zheng Fang, Anindya Ghose. 2016. Mobile ad effectiveness: Hyper-contextual targeting with crowdedness. *Marketing Science* **35**(2) 218–233.
- App Annie. 2020. The state of mobile 2020. Tech. rep., App Annie.
- Aral, Sinan. 2013. To go from big data to big insight, start with a visual. *Harvard Business Review* .
- Aral, Sinan, Lev Muchnik, Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51) 21544–21549.
- Aral, Sinan, Christos Nicolaides. 2017a. Exercise contagion in a global social network. *Nature Communications* **8**(14753). doi:10.1038/ncomms14753.
- Aral, Sinan, Christos Nicolaides. 2017b. Exercise contagion in a global social network. *Nature communications* **8**(1) 1–8.
- Aral, Sinan, Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science* **337**(6092) 337–341. doi:10.1126/science.1215842.
- Aral, Sinan, Michael Zhao. 2020. Social spillovers in online news consumption. *Available at SSRN 3328864* .
- Ash, Elliott, Sergio Galletta, Dominik Hangartner, Yotam Margalit, Matteo Pinna. 2020. The effect of fox news on health behavior during covid-19. *Available at SSRN 3636762* .
- Athey, Susan, Markus M. Mobius, Jenő Pal. 2017. The impact of aggregators on internet news consumption .
- Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, Arlene Wong. 2018. Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives* **32**(3) 259–80.
- Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo, Matthew O. Jackson. 2013. The diffusion of microfinance. *Science* **341**(6144). doi:10.1126/science.1236498.
- Bart, Yakov, Andrew T Stephen, Miklos Sarvary. 2014. Which products are best suited to mobile advertising? a field study of mobile display advertising effects on consumer attitudes and intentions. *Journal of Marketing Research* **51**(3) 270–285. doi:10.1509/jmr.13.0503.
- Barthel, Michale. 2017. Newspapers fact sheet. Tech. rep., Pew Research Center.

- Bergé, Laurent. 2018. Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers* (13).
- Bhamidipati, Narayan, Ravi Kant, Shaunak Mishra. 2017. A large scale prediction engine for app install clicks and conversions. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 167–175.
- Blake, Thomas, Chris Nosko, Steven Tadelis. 2015. Consumer heterogeneity and paid search effectiveness: A large-scale field experiment. *Econometrica* **83**(1) 155–174.
- Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, James H. Fowler. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* **489**(7415). doi:10.1038/nature11421.
- Bramoullé, Yann, Habiba Djebbari, Bernard Fortin. 2009. Identification of peer effects through social networks. *Journal of Econometrics* **150**(1) 41–55.
- Brock, William A., Steven N. Durlauf. 2001. Interactions-based models. Elsevier, 3297 – 3380. doi:https://doi.org/10.1016/S1573-4412(01)05007-3.
- Brzezinski, Adam, Valentin Kecht, David Van Dijke, Austin L Wright. 2020. Belief in science influences physical distancing in response to covid-19 lockdown policies. *University of Chicago, Becker Friedman Institute for Economics Working Paper* (2020-56).
- Buckee, Caroline O, Satchit Balsari, Jennifer Chan, Mercè Crosas, Francesca Dominici, Urs Gasser, Yonatan H Grad, Bryan Grenfell, M Elizabeth Halloran, Moritz UG Kraemer, et al. 2020. Aggregated mobility data could help fight COVID-19. *Science* .
- Calzada, Joan, Ricard Gil. 2016. What do news aggregators do? evidence from google news in spain and germany .
- Chen, Tianqi, Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.
- Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng, Yutian Li. 2019. *xgboost: Extreme Gradient Boosting*. R package version 0.90.0.1.
- Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins. 2018. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* **21**(1).
- Chinazzi, Matteo, Jessica T Davis, Marco Ajelli, Corrado Gioannini, Maria Litvinova, Stefano Merler, Ana Pastore y Piontti, Kunpeng Mu, Luca Rossi, Kaiyuan Sun, et al. 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**(6489) 395–400.
- Chiou, Lesley, Catherine Tucker. 2017. Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy* **26**(4) 782–805. doi:10.1111/jems.12207.
- Chiou, Lesley, Catherine Tucker. 2020. Social distancing, internet access and inequality. Tech. Rep. 26982, National Bureau of Economic Research.
- Coleman, James S. 1968. Equality of educational opportunity. *Integrated Education* **6**(5) 19–28.
- Coviello, Lorenzo, Uri Gneezy, Lorenz Goette. 2017. A large-scale field experiment to evaluate the effectiveness of paid search advertising .
- Coviello, Lorenzo, Yunkyu Sohn, Adam D. I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, James H. Fowler. 2014a. Detecting emotional contagion in massive social networks. *PloS one* **9**(3) e90315.



- Coviello, Lorenzo, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, James H Fowler. 2014b. Detecting emotional contagion in massive social networks. *PloS one* **9**(3).
- Dellarocas, Chrysanthos, Juliana Sutanto, Mihai Calin, Elia Palme. 2016. Attention allocation in information-rich environments: The case of news aggregators. *Management Science* **62**(9) 2543–2562. doi:10.1287/mnsc.2015.2237.
- Durlauf, Steven N., Hisatoshi Tanaka. 2008. Understanding regression versus variance tests for social interactions. *Economic Inquiry* **46**(1) 25–28. doi:10.1111/j.1465-7295.2007.00076.x.
- eMarketer. 2019. Global digital ad spending 2019. Tech. rep., eMarketer.
- Fang, Zheng, Bin Gu, Xueming Luo, Yunjie Xu. 2015. Contemporaneous and delayed sales impact of location-based mobile promotions. *Information Systems Research* **26**(3) 552–564.
- Flaxman, Seth, Swapnil Mishra, Axel Gandy, H Juliette T Unwin, Thomas A Mellan, Helen Coupland, Charles Whittaker, Harrison Zhu, Tresnia Berah, Jeffrey W Eaton, et al. 2020. Estimating the effects of non-pharmaceutical interventions on covid-19 in europe. *Nature* 1–5.
- Fong, Nathan M, Zheng Fang, Xueming Luo. 2015. Geo-conquesting: Competitive locational targeting of mobile promotions. *Journal of Marketing Research* **52**(5) 726–735.
- Frisch, Ragnar, Frederick V Waugh. 1933. Partial time regressions as compared with individual trends. *Econometrica* 387–401.
- Gaure, Simen. 2013. lfe: Linear group fixed effects. *The R Journal* **5**(2) 104–117.
- George, Lisa, Joel Waldfogel. 2003. Who affects whom in daily newspaper markets? *Journal of Political Economy* **111**(4) 765–784.
- Ghose, Anindya, Avi Goldfarb, Sang Pil Han. 2013. How is the mobile internet different? search costs and local activities. *Information Systems Research* **24**(3) 613–631.
- Ghose, Anindya, Sang Pil Han. 2011. An empirical analysis of user content generation and usage behavior on the mobile internet. *Management Science* **57**(9) 1671–1691.
- Ghose, Anindya, Sang Pil Han. 2014. Estimating demand for mobile applications in the new economy. *Management Science* **60**(6) 1470–1488.
- Ghose, Anindya, Beibei Li, Siyuan Liu. 2019. Mobile targeting using customer trajectory patterns. *Management Science* **65**(11) 5027–5049.
- Gilchrist, Duncan S., Emily G. Sands. 2016. Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy* **124**(5) 1339–1382. doi:10.1086/688177.
- Golden, Joseph M, John J Horton. 2017. The effects of search advertising on competitors: An experiment before a merger .
- Goodman-Bacon, Andrew. 2018. Difference-in-differences with variation in treatment timing. Tech. rep., National Bureau of Economic Research.
- Hausman, Catherine, David S Rapson. 2018. Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics* **10** 533–552.
- Hlavac, Marek. 2015. Stargazer: Well-formatted regression and summary statistics tables. *R package version* **5**(1).
- Holtz, David, Michael Zhao, Seth G Benzell, Cathy Y Cao, M Amin Rahimian, Jeremy Yang, Jennifer Nancy Lee Allen, Avinash Collis, Alex Vernon Moehring, Tara Sowrirajan, et al. 2020. Interdependence and the cost of uncoordinated responses to covid-19. *Proceedings of the National Academy of Sciences* doi:10.1073/pnas.2009522117.

- Hong, Sounman. 2012. Online news on twitter: Newspapers' social media adoption and their online readership. *Information Economics and Policy* **24**(1) 69–74.
- Hsiang, Solomon, Daniel Allen, Sébastien Annan-Phan, Kendon Bell, Ian Bolliger, Trinetta Chong, Hannah Druckenmiller, Luna Yue Huang, Andrew Hultgren, Emma Krasovich, et al. 2020. The effect of large-scale anti-contagion policies on the covid-19 pandemic. *Nature* 1–9.
- Jia, Jayson S., Xin Lu, Yun Yuan, Ge Xu, Jia Jianmin, Nicholas A. Christakis. 2020. Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature* .
- Kraemer, Moritz UG, Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M Pigott, Louis du Plessis, Nuno R Faria, Ruoran Li, William P Hanage, et al. 2020. The effect of human mobility and control measures on the COVID-19 epidemic in china. *Science* .
- Kramer, Adam D. I., Jamie E. Guillory, Jeffrey T. Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences* **111**(24) 8788–8790. doi:10.1073/pnas.1320040111.
- Li, Chenxi, Xueming Luo, Cheng Zhang, Xiaoyi Wang. 2017. Sunny, rainy, and cloudy with a chance of mobile promotion effectiveness. *Marketing Science* **36**(5) 762–779.
- Lovell, Michael C. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association* **58**(304) 993–1010.
- Luo, Xueming, Michelle Andrews, Zheng Fang, Chee Wei Phang. 2014. Mobile targeting. *Management Science* **60**(7) 1738–1756.
- Ma, Qiang, Shanmugavelayutham Muthukrishnan, Wil Simpson. 2016. App2vec: Vector modeling of mobile apps and applications. *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 599–606.
- Manski, Charles F. 1993. Identification of endogenous social effects: The reflection problem. *The review of economic studies* **60**(3) 531–542.
- Marin Software. 2019. The state of digital advertising 2019. Tech. rep., Marin Software.
- McKinney, Wes, et al. 2010. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, vol. 445. Austin, TX, 51–56.
- McPherson, Miller, Lynn Smith-Lovin, James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* **27**(1) 415–444.
- Menne, Matthew J, Imke Durre, Russell S Vose, Byron E Gleason, Tamara G Houston. 2012. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology* **29**(7) 897–910.
- Moretti, Enrico. 2011. Social learning and peer effects in consumption: Evidence from movie sales. *The Review of Economic Studies* **78**(1) 356–393.
- NOAA. 2018. Noaa national centers for environmental information (ncei) u.s. billion-dollar weather and climate disasters .
- Oliver, Nuria, Bruno Lepri, Harald Sterly, Renaud Lambiotte, Sébastien Delataille, Marco De Nadai, Emmanuel Letouzé, Albert Ali Salah, Richard Benjamins, Ciro Cattuto, et al. 2020. Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle.
- Olsen, Asmus Leth, Frederik Hjorth. 2020. Willingness to distance in the COVID-19 pandemic. Tech. rep., University of Copenhagen.
- Ooms, Jeroen. 2014. The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv preprint arXiv:1403.2805* .

- Painter, Marcus, Tian Qiu. 2020. Political beliefs affect compliance with COVID-19 social distancing orders. Tech. rep. Available at SSRN 3569098.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rafieian, Omid, Hema Yoganarasimhan. 2020. Targeting and privacy in mobile advertising .
- Raifman, Julia, Kristen Nocka, David Jones, Jacob Bor, Sarah Lipson, Jonathan Jay, Megan Cole, Noa Krawczyk, Philip Chan, Sandro Galea, et al. 2020. Covid-19 us state policy database .
- Rossum, Guido. 1995. Python reference manual .
- Sahu, Anit K, Shaunak Mishra, Narayan Bhamidipati. 2018. Managing app install ad campaigns in rtb: A q-learning approach .
- Schwartz, Josh. 2018. What happens when facebook goes down? people read the news .
- Sen, Anaya, Pinar Yildirim. 2015. Clicks bias in editorial decisions: How does popularity shape online news coverage? doi:10.2139/ssrn.2619440.
- Shankar, Venkatesh, Sridhar Balasubramanian. 2009. Mobile marketing: A synthesis and prognosis. *Journal of Interactive Marketing* **23**(2) 118–129.
- Simonov, Andrey, Shawndra Hill. 2019. Competitive advertising on brand search: Traffic stealing and customer selection .
- Simonov, Andrey, Szymon K Sacher, Jean-Pierre H Dubé, Shirsho Biswas. 2020. The persuasive effect of fox news: non-compliance with social distancing during the covid-19 pandemic. Tech. rep., National Bureau of Economic Research.
- Sismeiro, Catarina, Ammara Mahmood. 2018. Competitive vs. complementary effects in online social networks and news consumption: A natural experiment. *Management Science* .
- Stock, James H, Motohiro Yogo. 2002. Testing for weak instruments in linear iv regression.
- Stocking, Galen. 2017. Digital news fact sheet. Tech. rep., Pew Research Center.
- Team, R Core, et al. 2013. R: A language and environment for statistical computing .
- Tucker, Catherine. 2008. Identifying formal and informal influence in technology adoption with network externalities. *Management Science* **54**(12) 2024–2038. doi:10.1287/mnsc.1080.0897.
- Wickham, Hadley. 2016a. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
- Wickham, Hadley. 2016b. *ggplot2: elegant graphics for data analysis*. Springer.
- Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy McGowan, Romain François, Garrett Golemund, Alex Hayes, Lionel Henry, Jim Hester, et al. 2019. Welcome to the tidyverse. *Journal of Open Source Software* **4**(43) 1686.
- Zhang, Xiaoquan Michael, Feng Zhu. 2011. Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review* **101**(4) 1601–15.
- Zhang, Yingjie, Beibei Li, Xueming Luo, Xiaoyi Wang. 2019. Personalized mobile targeting with user engagement stages: Combining a structural hidden markov model and field experiment. *Information Systems Research* **30**(3) 787–804.