Social Media and Newsroom Production Decisions*

Julia Cagé[†1], Nicolas Hervé[2], and Béatrice Mazoyer[2,3]

[1]Sciences Po Paris and CEPR, [2]Institut National de l'Audiovisuel, [3]CentraleSupélec

October 19, 2020

Abstract

Social media affects not only the way we consume news, but also the way news is produced, including by traditional media outlets. In this paper, we study the propagation of information from social media to mainstream media, and investigate whether news editors are influenced in their editorial decisions by stories popularity on social media. To do so, we build a novel dataset including a representative sample of all tweets produced in French between July 2018 and July 2019 (1.8 billion tweets, around 70% of all tweets in French during the period) and the content published online by about 200 mainstream media during the same time period, and develop novel algorithms to identify and link events on social and mainstream media. To isolate the causal impact of popularity, we rely on the structure of the Twitter network and propose a new instrument based on the interaction between measures of user centrality and news pressure at the time of the event. We show that story popularity has a positive effect on media coverage, and that this effect varies depending on media outlets' characteristics. These findings shed a new light on our understanding of how editors decide on the coverage for stories, and question the welfare effects of social media.

Keywords: Internet, Information spreading, Network analysis, Social media, Twitter, Text analysis

JEL No: C31, D85, L14, L15, L82, L86

* We gratefully acknowledge the many helpful comments and suggestions from Dominique Cardon, Edgard Dewitte, Mirko Draca, Michele Fioretti, Malka Guillot, Emeric Henry, Céline Hudelot, Brian Knight, Gregory Martin, Benjamin Marx, Elisa Mougin, Maria Petrova, Zeynep Pehlivan, Andrea Prat, Jesse Shapiro, Clémence Tricaud, and Katia Zhuravskaya. We are grateful to seminar participants at Brown University, the CEPR Virtual IO Seminar (VIOS), the DC Political Economy Webinar, the London School of Economics (CEP/DOM Capabilities, Competition and Innovation Seminar Series), the Medialab, the Oz Virtual Econ Research Seminar (UNSW), and Sciences Po Paris, and to conference participants at the Digital Economics Research Network Workshop. Martin Bretschneider, Michel Gutmann and Auguste Naim provided outstanding research assistance. We gratefully acknowledge financial help from the NET Institute and from the French National Research Agency (ANR-17-CE26-0004). Responsibility for the results presented lies entirely with the authors. An online Appendix with additional empirical material is available here.

†Corresponding author. `julia [dot] cage [at] sciencespo [dot] fr`.

1 Introduction

Nearly three fourth of the American adults use social media, and more than two-thirds report that they get at least some of their news on them (Pew Research Center, 2018, 2019). While there is growing concern about misinformation online and the consumption of fake news (Allcott and Gentzkow, 2017; Allcott et al., 2019), little is known about the impact of social media on the production of news. However, social media not only affect the way we consume news, but also the way news is produced, including by traditional media. First, social media users may report events before traditional media (Sakaki et al., 2010).[1] Second, news editors may use social media as a signal to draw inferences about consumers' preferences. Furthermore, social media compete with mainstream media for consumers' attention, and this may affect publishers' incentives to invest in quality (de Cornière and Sarvary, 2019).

In this article, we investigate how editors decide on the coverage for stories, and in particular the role played by social media in their decision. To do so, we have built a completely new dataset including around 70% of all the tweets in French emitted during an entire year (July 2018 - July 2019)[2] and the content produced online by French-speaking general information media outlets during the same time period. 205 media outlets are included, encompassing newspapers, TV channels, radio stations, pure online media, and the news agency AFP.[3]) Our dataset contains around 1.8 billion tweets as well as 4 million news articles. For each tweet (that can be original tweets but also retweets or quotes of past tweets) as well as each news article, we determine their precise "publication" time. Furthermore, we collect information on the Twitter users, in particular their number of followers and the number of interactions generated by their tweets.

We develop a new detection algorithm to identify news events on Twitter together with an algorithm to identify news events on mainstream media. Then, we propose a "community detection" algorithm to investigate the relationship between social media and mainstream media events, and the propagation of information between Twitter and traditional media. An event here is a cluster of documents (tweets and/or media articles) that discuss the same news story.[4] When an event is covered on both social and mainstream media, we determine

[1] As highlighted by Alan Rusbridger as early as 2010, *"increasingly, news happens first on Twitter."* (...) *"If you're a regular Twitter user, even if you're in the news business and have access to wires, the chances are that you'll check out many rumours of breaking news on Twitter first. There are millions of human monitors out there who will pick up on the smallest things and who have the same instincts as the agencies to be the first with the news. As more people join, the better it will get."* (https://www.theguardian.com/media/2010/nov/19/alan-rusbridger-twitter).

[2] The data collection is based on the language of the tweets (French language vs. other languages) and not on the location of the users. More details are provided in Section 2.1.

[3] Our dataset also includes the content produced online by 10 French-speaking foreign media outlets such as *Le Temps* (Switzerland). More details on the data collection are provided in Section 2.2.

[4] Our algorithms are based on natural language processing methods. As will be described in Section 2.3 below, events are detected by our algorithms using the fact that the documents share sufficient semantic similarity.

the origin of the information, i.e. the first Twitter user who tweets about the event or the first media article that breaks the story. For the subset of events that originate first on social media, we study whether the event popularity on Twitter impacts the coverage that mainstream media devote to this event.

The scale of our dataset (one year of data with several million tweets and news articles) allows us to follow a split-sample approach to relieve concerns about specification search and publication bias (Leamer, 1978, 1983; Glaeser, 2006).[5] Following Fafchamps and Labonne (2016, 2017) and Anderson and Magruder (2017), we first perform our analysis on the July 2018 - September 2018 time period (three months of data). The results of this draft rely on this sub-sample that we use to narrow down the list of hypotheses we wish to test and to specify the research plan that we will pre-register. The final version of the paper will follow the pre-registered plan and perform the empirical analysis on the remainder of the data (October 2018 - July 2019).

The sample we use in this version of the article to delineate our research plan (July 2018-September 2018) includes 417 million tweets and $929,764$ news articles. We identify $5,137$ joint events, i.e. events that are covered both on Twitter and on traditional media. Producing these data is our first contribution. It is to the best of our knowledge the most exhaustive dataset on social media and mainstream media events available to researchers, and the algorithms we develop to analyze these data could be of future use to other researchers studying online information. Our second contribution is descriptive. While there is a growing literature focusing on the propagation of fake news on social media (Vosoughi et al., 2017, 2018), little is known about the propagation of all news between social media and traditional media outlets. In this article, we investigate this propagation and provide new descriptive evidence on the interlinking between social and mainstream media.

Third and most importantly, we open the black box of newsroom production decisions, and investigate the extent to which news editors are influenced in their editorial decisions by stories' popularity on social media. Focusing on the subset of news stories that originate first on Twitter ($4,392$ out of the $5,137$ joint events), we investigate how their popularity affects the coverage that traditional media devote to these stories. The popularity of a story on Twitter is measured before the first media article devoted to the story appears; we use the total number of tweets related to the event, including retweets and quotes. We refer to the author of the first tweet in the event as the seed of the event.

The main empirical challenge here lies in the fact that a story's popularity on Twitter and its media coverage can both be driven by the intrinsic interest of the story, regardless of what happens on social media. Hence, to identify the specific role played by social media, we need to find exogenous sources of variation of a story's popularity on Twitter. To do so,

[5] We thank Jesse Shapiro for suggesting this approach.

we propose a new instrument that relies on the interaction between the seed's centrality in the Twitter network and the "news pressure" at the time of the event.[6] Our identification assumption is that, once we control for the direct effects of centrality and news pressure, as well as for the seed's number of followers, the interaction between the seed's centrality and news pressure should only affect traditional news production through its effect on the tweet's visibility on Twitter.

To measure centrality, in the spirit of an intention-to-treat analysis, we compute a measure of the number of "impressions"[7] generated by the previous tweets of the seed's followers: the higher this number, the higher the potential number of retweets, regardless of the tweet's intrinsic interest. We approximate the potential number of "impressions" by the observed number of interactions (retweets/likes/quotes) generated by the previous tweets of the seed's followers (i.e. by all their tweets before the news event). Importantly here, we use the average number of interactions generated by the tweets of the seed's followers, not by the tweets of the seed itself, given the former is arguably more exogenous. News pressure is measured by the number of interactions generated by all the tweets published in the hour preceding the first tweet in the event. We also control for day-of-the-week fixed effects, month fixed effects, as well as an indicator variable equal to one if the first tweet is tweeted during the night. We perform the analysis both at the event level – considering the overall number of articles published and the number of media outlets covering the event – and at the media outlet level, to investigate heterogeneity depending on the media characteristics and study the mechanisms at play. In the latter case, we control for media outlet fixed effects and cluster the standard errors at the event level.

We show that an increase of $1,000$ in the number of tweets published before the first media article appears is correlated with an increase of 3.2 in the number of media articles published in the event; this increase is partly driven by a higher number of media outlets covering the event (+0.4). This effect is both statistically and economically significant. In terms of magnitude, it implies that a one-percent increase in the number of tweets is associated with a 8.9% increase in the number of articles.

These results are robust to controlling for the endogeneity of the event popularity on Twitter. As expected given the direction of the omitted variable bias, the magnitude of the IV estimates is smaller ($1,000$ additional tweets lead to 2 additional media articles) than the one we obtain with the naive approach. Reassuringly, our IV results are robust to controlling for additional characteristics of the seed of the event, and doing so does not affect the magnitude of the estimates.

[6] We thank Katia Zhuravskaya for suggesting this approach.

[7] The number of impressions is a total tally of all the times the tweet has been seen. Unfortunately, this statistics is not directly available to researchers. But we can approximate this number by using the observed number of interactions (retweets/likes/quotes) generated by the tweet.

To understand the mechanisms behind our findings, we then turn to the media-level analysis and investigate the heterogeneity of our results depending on the characteristics of the media outlets. First, journalists monitor Twitter; the Muck Racks "State of Journalism 2019" report reveals that nearly 60% of reporters turn to digital newspapers or magazines as their first source of news, and 22% check Twitter first. For each media outlet in our sample, we compute information on their social media presence and show that the magnitude of the effect is stronger for the media whose journalists are more present on the social network. However, while this may help to understand why a number of stories emerge first on Twitter and the high reactivity of mainstream media, it does not explain why the intensity of the media coverage (on the intensive margin) also varies with the popularity of a story on Twitter. In the absence of perfect information on consumer preferences, publishers may use Twitter as a signal that allows them to draw inferences about what news consumers are interested in. We investigate whether our results vary depending on the media outlets' business model, in particular whether they use a paywall and their reliance on advertising revenues. In addition, for a subset of the media, we gather information on the size of their newsroom, which provides us with a proxy on their investment in news quality. We also investigate whether there is heterogeneity depending on the offline format of the media (e.g. newspapers vs. TV channels) and on the topic of the event (e.g. sport, international affairs, economics, etc.). While the magnitude of the observed effects varies depending on the type of media – e.g. it is stronger for national daily newspapers than for local newspapers – we find no statistically significant difference depending on the reliance on advertising revenues.

Finally, we discuss the welfare implications of our results. Given our instrumental variable strategy, our causal estimates capture the effects of a variation in popularity that is uncorrelated with the underlying newsworthiness of a story. In other words, our findings suggest that social media may provide a biased signal of what readers want. Turning to the demand for news, we show that the popularity of a news event on Twitter is associated with a higher demand for the news articles published by the media on this event, but that the magnitude of the effect is much smaller than the observed increase in media coverage.

Our results are robust to a variety of estimation procedures, to different measures of centrality, to the use of different samples and to a battery of additional sensitivity tests. In particular, we show that they are robust to dropping the news events whose seed is the Twitter account of either a media outlet or a journalist, as well as the events broken by seeds who broke more than one event during our time period, to avoid capturing a celebrity bias as well as tweets by influencers.

Literature review This paper contributes to the growing literature on the impact of the introduction of new media technologies on political participation, government accountability and electoral outcomes (see among others Gentzkow et al. (2011); Snyder and Stromberg (2010); Cagé (2020) on newspapers; Strömberg (2004) on radio; Gentzkow (2006); Angelucci and Cagé (2019); Angelucci et al. (2020) on television, and Boxell et al. (2018); Gavazza et al. (2019) on the Internet). There are very few papers examining how social media affects voting (for a review of the literature see Zhuravskaya et al., 2020), and these mainly concentrate on the role played by fake news (Allcott and Gentzkow, 2017). So far, the focus of this literature has mostly been on news consumption, and little is known about the empirical impact social media have on news production by mainstream media. One exception is a work-in-progress article by Hatte et al. (2020) who study the effect of Twitter on the US TV coverage of the Israeli-Palestinian conflict. Compared to this work, our contribution is threefold. First, we focus on the overall activity on Twitter and collect a large representative sample of about 70% of all tweets (about 1.8 billion tweets) rather than the tweets associated with a small number of keywords. Second, we develop an instrument for measuring popularity shocks on Twitter based on the structure of the network that could be of use in different contexts. Finally, we investigate whether there are heterogeneous effects depending on the media characteristics, in particular their business model and their reliance on advertising revenues.

An expanding theoretical literature studies the effects of social media on news. De Cornière and Sarvary (2019) develop a model where consumers allocate their attention between a newspaper and a social platform (see also Alaoui and Germano, 2020, for a theory of news coverage in environments of information abundance). They document a negative impact on the media's incentives to invest in quality. This literature mainly concentrates on competition for attention between newspapers and social media, and documents a trade-off between the business-stealing and the readership-expansion effect of platforms (Jeon and Nasr, 2016).[8] In this article, we highlight the fact that not only are mainstream and social media competing for attention, but also that social media can be used by mainstream media both as a source of news and as a signal to draw inferences on consumers' preferences. We investigate empirically how a story's popularity on Twitter impacts the information produced by traditional media, and in particular the intensity of the coverage they devote to that story.

Our results also contribute to the growing literature in the fields of Economics and Political Science using social media data, and in particular the structure of the social networks – usually Twitter – as a source of information on the ideological positions of actors (Barberá, 2015; Cardon et al., 2019), the importance of ideological segregation and the extent of political polarization (Halberstam and Knight, 2016; Giavazzi et al., 2020), and political language dissemination (Longhi et al., 2019).[9] Gorodnichenko et al. (2018) study information diffusion on Twitter, and Allcott et al. (2019) the spread of false content. While this literature mostly

[8]See Jeon (2018) for a survey of articles on news aggregators.

[9]See also Barberá et al. (2019) who use Twitter data to analyze the extent to which politicians allocate attention to different issues before or after shifts in issue attention by the public.

focuses on relatively small corpuses of tweets and on corpuses that are not representative of the overall activity on Twitter (e.g. Gorodnichenko et al., 2018, make requests to collect tweets using Brexit-related keywords), in this paper, we build a representative corpus of tweets and impose no restriction on the data collection. Furthermore, we contribute to this literature by considering the propagation of information on social media as well as by studying whether and how information propagates from social media to mainstream media. While Cagé et al. (2020) only consider news propagation on mainstream media, we investigate the extent to which the popularity of a story on social media affects the coverage devoted to this story by traditional media outlets.

The impact of "popularity" on editorial decisions has been studied by Sen and Yildirim (2015) who use data from an Indian English daily newspaper to investigate whether editors expand online coverage of stories which receive more clicks initially.[10] Compared to this previous work, our contribution is threefold. First, we use the entire universe of French general information media online (around 200 media outlets), rather than one single newspaper. Second, we not only identify the role played by popularity, but also investigate whether there is heterogeneity depending on the characteristics of the media outlets, as well as the topic of the story. Third, we consider both the extensive and the intensive margin[11], rather than focusing on the subset of stories that receive at least some coverage in the media. Finally, we also contribute to the empirical literature on media by using a split-sample approach; while this approach is increasingly used in economics with the pre-registration of Randomized Controlled Trials, we believe we are the first to use it with "real-world data" on such a large scale.

In addition to this, we contribute to the broader literature on social media that documents its impact on racism (Müller and Schwarz, 2019), political protests (Enikolopov et al., 2020), the fight against corruption (Enikolopov et al., 2018), and the size of campaign donations (Petrova et al., 2017). Overall, social media is a technology that has both positive and negative effects (Allcott et al., 2020). This also holds true for its impact on traditional media: we contribute to this literature by documenting the complex effects social media has on news production, and consequently on news consumption.

Finally, our instrumentation strategy is related on the one hand to the literature that looks at the quantity of newsworthy material at a given moment of time (e.g. Eisensee and Strömberg, 2007; Djourelova and Durante, 2019), and on the other hand to the literature on network interactions (see Bramoullé et al., 2020, for a recent survey). The main issue faced by researchers willing to identify the causal effects of peers is that the structure of the

[10] See also Claussen et al. (2019) who use data from a German newspaper to investigate whether automated personalized recommendation outperforms human curation in terms of user engagement.

[11] The intensive margin here corresponds to whether a story is covered, while on the extensive margin we consider both the total number of articles (conditional on covering the story) and the characteristics of these articles, e.g. their length.

network itself may be endogenous. In this paper, we relax the concern of network endogeneity by considering the interaction between the network and news pressure at a given moment of time.

The rest of the paper is organized as follows. In Section 2 below, we describe the Twitter data and the news content data we use in this paper, review the algorithms we develop to study the propagation of information between social and mainstream media, and provide new descriptive evidence on news propagation. In Section 3, we present our empirical specification, and in particular the new instrument we propose to identify the causal impact of a story's popularity on the subsequent news coverage it receives. Section 4 presents the results and analyzes various dimensions of heterogeneity. In Section 5, we discuss the mechanisms at play, and we perform a number of robustness checks in Section 6. Finally, Section 7 concludes.

2 Data, algorithms, and descriptive statistics

The new dataset we built for this study is composed of two main data sources that we have collected and merged together: on the one hand, a representative sample of tweets, and on the other hand, the online content of the general information media outlets. In this section, we describe these two datasets in turn, and then present the algorithms we develop to identify events on social and on traditional media, and interact them.

2.1 Data: Tweets

First, we collect a representative sample of all the tweets in French during an entire year: July 2018 - July 2019. Our dataset, which contains around 1.8 billion tweets, encompasses around 70% of all the tweets in French (including the retweets) during this time period.[12] For each of these tweets, we collect information on their "success" on Twitter (number of likes, of comments, etc.), as well as information on the characteristics of the user at the time of the tweet (e.g. its number of followers).

To construct this unique dataset, we have combined the Sample and the Filter Twitter Application Programming Interfaces (APIs), and selected keywords. Here, we quickly present our data collection strategy; more details are provided in Mazoyer et al. (2018, 2020a) and we summarize our data collection setup in the online Appendix Figure A.1.

2.1.1 Data collection strategy

There are different ways of collecting large volumes of tweets, although collecting the full volume of tweets emitted during a given period is not possible. Indeed, even if Twitter

[12] See below for a discussion of the completeness of our dataset.

is known for providing a larger access to its data than other social media platforms (in particular Facebook), the Twitter streaming APIs are strictly limited in term of volume of returned tweets. The Sample API continuously provides 1% of the tweets posted around the world at a given moment of time (see e.g. Kergl et al., 2014; Morstatter et al., 2014). The Filter API continuously provides the same volume of tweets (1% of the global volume of tweets emitted at a given moment), but corresponding to the input parameters chosen by the user, including keywords, account identifiers, geographical area, as well as the language of the tweets.

To maximize the size of our dataset, we identify the keywords that maximize the number of returned tweets in French language as well as their representativity of the real Twitter activity. The selected terms had to be the most frequently written words on Twitter, and we had to use different terms (and terms that do not frequently co-occur in the same tweets) as parameters for each API connection. To do so, we extract the vocabulary from a set of tweets collected using the Sample API and obtain a subset of the words having the highest document-frequency. From this subset, we build a word co-occurrence matrix and, using spectral clustering, extract clusters of words that are then used as parameters of our different connections to the Filter API. By doing so, we group terms that are frequently used together (and separate terms that are rarely used together) and thus collect sets of tweets with the smallest possible intersection.

Filtering the tweets An important issue on Twitter is the use of bots, i.e. non-human actors and trolls publishing tweets on the social media (see e.g. Gorodnichenko et al., 2018). In recent years, Twitter has been actively cracking down on bots. In our analysis, we perform some filtering designed to limit the share of tweets from bots in our dataset. However we do not remove all automated accounts: many media accounts, for example, post some content automatically, and are not considered to be bots. Moreover, some types of automatic behaviors on Twitter, such as automatic retweets, may contribute to the popularity of stories and therefore should be kept in our dataset.

Our filtering rules are as follows. First, we use the "source" label provided by Twitter for each tweet.[13] Tweets emanating from a "source" such as "Twitter for iPhone" can be considered valid; however, we excluded sources explicitly described as bots, or referring to gaming or pornographic websites. We also excluded apps automatically posting tweets based

[13]Twitter describes this label as follows: "Tweet source labels help you better understand how a Tweet was posted. This additional information provides context about the Tweet and its author. If you dont recognize the source, you may want to learn more to determine how much you trust the content. [...] Authors sometimes use third-party client applications to manage their Tweets, manage marketing campaigns, measure advertising performance, provide customer support, and to target certain groups of people to advertise to. Third-party clients are software tools used by authors and therefore are not affiliated with, nor do they reflect the views of, the Tweet content. Tweets and campaigns can be directly created by humans or, in some circumstances, automated by an application."

on the behaviour of users: for example, many Twitter users (who are human beings and usually publish tweets they have written themselves) post automatic tweets such as "I like a video on Youtube: [url]". The entire list of the excluded sources is presented in Table A.1 of the online Appendix.

Second, we filter the users depending on their activity on the network: we only keep users with fewer than 1,000 tweets a day[14], and the users who have at least 1 follower. Finally, we only keep the users who post at least 3 tweets in French between July and September 2018.[15]

Completeness of the dataset Ultimately, we obtain 1.8 billion tweets in French between July 2018 and July 2019. While the objective of our data collection method was to maximize the number of tweets we collect – and given we do not know the actual number of tweets emitted in French during the same time period –, we need to use other corpora to get a sense of the completeness of our dataset. We rely on three different metrics to estimate the share of the tweets that we collect.

The DLWeb, i.e. the French Internet legal deposit department at the INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French audiovisual archives) collects tweets concerning audiovisual media by using a manually curated list of hashtags and Twitter accounts. We compare our dataset of tweets with the tweets they collected for 25 hashtags in December 2018. We find that on average we recover 74% of the tweets collected by the DLWeb, and 78% if we exclude retweets[16] (see online Appendix Figure A.2 for more details).

Second, we compare our dataset with the corpus of tweets built by Cardon et al. (2019), which consists of tweets containing URLs from a curated list of 420 French media outlets. Cardon et al. (2019) provide us with all the tweets they collected in December 2018, i.e. 8.7 million tweets, out of which 7.3 million tweets in French. Our dataset contains 70% of these tweets in French, 74% if we exclude retweets (see online Appendix Figure A.3).

Finally, our third evaluation method is based on the total number of tweets sent by a user since the account creation, a metadata that is provided by Twitter for every new tweet. With this metric, we can get an estimate of the total number of tweets emitted by a given user between two of her tweets. We can then compare this value with the number of tweets we actually collect for that user. In practice, we select the tweets of all the users located in France.[17] We find that our dataset contains 61% of the real number of emitted tweets for

[14] As a matter of comparison, the Twitter account of *Le Monde* publishes on average 88 tweets per day, and that of *Le Figaro* 216.

[15] I.e. users who tweet on average at least once a month.

[16] Original tweets are better captured than retweets by our collection method, because each retweet allows us to archive the original tweet to which it refers. Therefore, we only need to capture one of the retweets of a tweet to get the original tweet. Retweets, on the other hand, are not retweeted, so we lose any chance of catching them if they were not obtained at the time they were sent.

[17] We focus on the users located in France given they have a higher probability to only publish tweets in

these users. This evaluation is a lower bound estimation of the percentage of collected tweets however, since some users located in France may write tweets in other languages than French.

All three comparison methods have their flaws, but reassuringly they produce close results. We can therefore conclude that we have collected between 60% and 75% of all the tweets in French during our time period. To the extent of our knowledge, there is no equivalent in the literature of such a dataset of tweets, in terms of size and representativity of the Twitter activity. We hope that both our methodology and data could be of use in the future to other researchers interested in social media.

2.1.2 Descriptive statistics

Split-sample approach As highlighted in the introduction, to address concerns about specification search and publication bias (Leamer, 1978, 1983; Glaeser, 2006), we implement a split-sample approach in this paper (Fafchamps and Labonne, 2016, 2017; Anderson and Magruder, 2017). We split the data into two non-overlapping datasets: July 2018 - September 2018 and October 2018 - July 2019. We used the three-month dataset covering July 2018 - September 2018 to narrow down the list of hypotheses we wish to test and prepare this version of the paper.

The final version of the paper will only use data from October 2018 to July 2019. The idea is to avoid multiple hypothesis testing which has been shown to be an issue in experimental economics (List et al., 2019) and could also be of concern here. Hence, for the remainder of the article, we will rely solely on the first three months of our dataset. This sample includes $417,153,648$ tweets; Table 1 presents summary statistics for these tweets.[18]

Summary statistics For each of the tweets, we have information of its length (102 characters on average or 6.2 words), and know whether it is a retweet of an existing tweet or an original tweet. 63% of the tweets in our dataset are retweets; some of these retweets are "quotes", i.e. comments on the retweeted tweet.[19] Of the original tweets, some are replies to other tweets (17% of the tweets in our sample). Finally, 13% of the tweets contain a URL, most often a link to a news media article or to a video.

We also gather information on the popularity of each of the tweets in our sample. On average, the tweets are retweeted 2.3 times, liked 3.7 times, and receive 0.2 replies (these numbers are only computed on the original tweets, given that retweets, likes and quotes are not attributed to the retweets but to the original tweets).

French, and we only capture here by construction the tweets in French language. In Section 6 below, we discuss in length the issue of user location.

[18]Online Appendix Table C.1 shows statistics on the sample of tweets we collect before applying the filters to exclude the bots as described above. This sample includes over 428 million tweets.

[19]Quote tweets are much like retweets except that they include a new tweet message.

Table 1: Summary statistics: Tweets (split-sample, July 2018-September 2018)

	Mean	St.Dev	P25	Median	P75	Max	Obs
Characteristics of the tweet							
Length of the tweet (nb of characters)	102	52	61	98	140	1,121	417,153,648
Number of words	6.2	4.0	3.0	6.0	9.0	269	417,153,648
=1 if tweet contains an URL	0.13	0.33	0.00	0.00	0.00	1	417,153,648
=1 if the tweet is a retweet	0.63	0.48	0.00	1.00	1.00	1	417,153,648
=1 if the tweet is a reply	0.17	0.38	0.00	0.00	0.00	1	417,153,648
=1 if the tweet is a quote	0.19	0.39	0.00	0.00	0.00	1	417,153,648
Popularity of the tweet							
Number of retweets	2.3	111.5	0.000	0.000	0.000	117,389	154,273,618
Number of replies	0.2	6.6	0.000	0.000	0.000	47,892	154,273,618
Number of likes	3.7	172.2	0.000	0.000	0.000	449,881	154,273,619

Notes: The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for all the tweets included in our dataset. Variables for the "popularity of the tweet" are only for the original tweets, given that the retweets/replies/likes are always attributed to the original tweets (hence the lower number of observations). The maximum number of characters (or length of the tweet) is above the 280 Twitter character limit. This is due to the fact that URLs and mentions (e.g. @*BeatriceMazoyer*) contained in the tweets are not included by Twitter in the character limit. We remove the stop-words before computing the "number of words" statistics. The list of stop-words is provided in the online Appendix Section A.1. Variables are described in more detail in the text.

Furthermore, we compute summary statistics on the Twitter users in our sample. Our dataset includes 4, 222, 734 unique users between July 2018 and September 2018. Table 2 provides these statistics the first time a user is observed in our data.[20] On average, the users tweeted 14, 100 times, liked 7, 463 tweets, and were following 642 other Twitter accounts. The average year of the account creation is 2014 (Twitter was created in 2006; see online Appendix Figure D.1 for the distribution of the users depending on the date on which they created their account). On average, users have 2, 166 followers; however, we observe significant variations: the vast majority of the users have just a few followers, but some of them act as central nodes in the network: the top 1% of the users in terms of followers account for more than 70% of the total number of followers (see online Appendix Figure D.2 for the distribution of the number of followers).

0.5% of the users in our sample have a verified account[21], 0.12% are the accounts of journalists, and 0.011% are media outlets' accounts. We have manually identified the Twitter accounts of the media outlets. For the Twitter accounts of journalists, we proceed to a semi-manual detection with the following method: first, we use the Twitter API to collect the name and description of all accounts that are followed by at least one media account. Second, we only keep the accounts that have some keywords related to the profession of journalist in their description, such as "journalist", "columnist", "news", etc. Third, we hire research assistants to manually select journalists from the remaining accounts by reading their names and description.

2.2 Data: News articles

We combine the Twitter data with the online content of traditional media outlets (alternatively called mainstream media) over the same time period, including newspapers, radio channels, TV stations, online-only news media, and the content produced by the "Agence France Presse" news agency (AFP).[22] The goal here is to gather all the content produced online by the "universe" of French general information news media, regardless of their offline format. The data is collected as part of the OTMedia research project, a unique data collection program conducted by the French National Audiovisual Institute (Cagé et al., 2020). Furthermore, we also gather the content produced online by 10 French-speaking (non-French) media outlets such as the daily newspaper *Le Temps Suisse* from Switzerland. This subset of French-speaking media was selected based on the fact that the tweets included in our sample include at least one URL linked to an article published online by these media.

[20] Alternatively, we compute the users' characteristics the last time we observe them. The results are presented in the online Appendix Table C.2.

[21] According to Twitter, an account may be verified if it is determined to be an account of public interest. Typically this includes accounts maintained by users in music, acting, fashion, government, politics, religion, journalism, media, sports, business, and other key interest areas.

[22] The AFP is the third largest news agency in the world (after the Associated Press and Reuters).

Table 2: Summary statistics: Twitter users

	Mean	St.Dev	P25	Median	P75	Max
User activity						
Total number of tweets	14,100	39,127	192	1,754	11,228	6,020,029
Nb of tweets user has liked	7,463	21,419	95	914	5,414	2,736,965
Nb of users the account is following	642	4,489	76	193	482	1,681,133
User identity						
Date of creation of the account	2014	3	2012	2015	2017	2018
=1 if verified account	0.005	0.073	0	0	0	1
=1 if user is a journalist	0.001	0.034	0	0	0	1
=1 if user is a media	0.0001	0.010	0	0	0	1
User popularity						
Nb of followers	2,166	86,811	24	129	477	58,484,193
Nb of public lists	19	578	0	1	6	1,028,761
Observations	4,222,734					

Notes: The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for all the Twitter users included in our dataset the first time we observe them. Variables are described in more detail in the text.

Newsroom characteristics Our dataset includes 205 unique media outlets (see online Appendix Section A.2 for the list of these media depending on their offline format), which published 929, 764 online news articles between July 2018 and September 2018. Table 3 shows summary statistics for the mainstream media included in our dataset. On average, between July 2018 and September 2018, the mainstream media in our data published 4, 152 news articles (i.e. around 48 articles per day), of which 1, 406 are classified in events (see below for the event definition). 63.4% of the articles come from the newspaper websites, 13, 1% from pure online media, 11.1% from the news agency, 8.8% from the radio station websites and the remainder from TV channel websites (see online Appendix Figure D.3.)

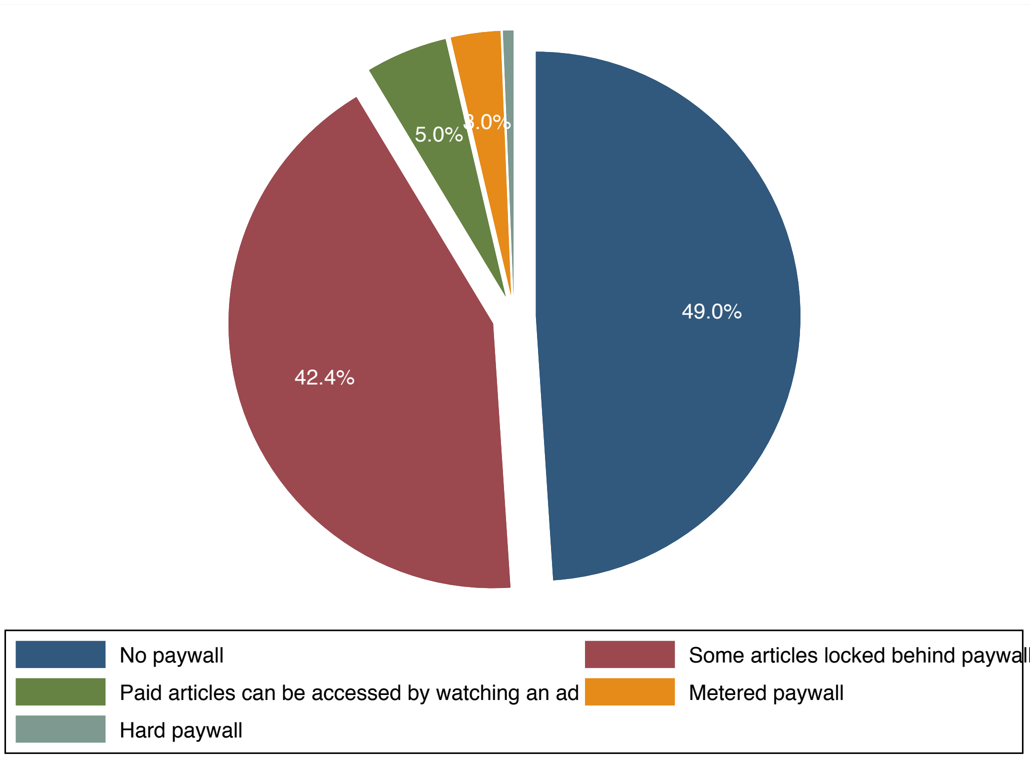
For all the media outlets in our sample, we also collect information on their social media presence. First, we identify their Twitter account(s)[23] and collect information on their popularity (number of followers and public lists, captured the first time we observe them in our sample), as well as the number of tweets posted by these accounts during our period of interest. On average, the media outlets in our sample have 3.1 different Twitter accounts. We compute the date of creation of each of these accounts, and report the oldest one in the table. Additionally, to proxy for the media outlets' social media presence, we compute the share of the articles the media outlet publishes online that it also posts on Twitter. In addition, for each of the media in our sample, we compute the number of journalists with a Twitter account, as well as the characteristics of these accounts.

[23]Some media only have one Twitter account, while others have many; e.g. for the newspaper *Le Monde*: @*lemondefr*, @*lemondelive*, @*lemonde_pol*, etc.

Table 3: Summary statistics: Media outlets

	Mean	St.Dev	P25	Median	P75	Max
Content						
Total content (thsd ch)	10,021	25,055	414	1,965	8,045	222,546
Total number of articles	4,152	10,035	185	837	3,366	85,676
Articles classified in events	1,406	4,524	11	114	864	55,932
Number of breaking news	31.8	135.0	0.0	0.0	13.0	1,656
Online audience (daily)						
Number of unique visitors	210,883	301,686	28,843	90,153	227,008	1,282,498
Number of visits	586,269	852,715	72,165	210,473	714,832	3,283,491
Number of page views	1,510,024	2,544,652	160,117	536,866	1,643,809	15,329,183
Social media presence						
% articles on Twitter	17	17	4	10	24	70
Number of Twitter accounts	3.1	5.7	1.0	1.0	2.0	43
Date of Twitter account creation	2009	1.3	2009	2009	2010	2016
Number of tweets	2,874	4,587	455	1,101	3,043	19,730
Nb journalists with Twitter account	211	354	38	81	220	3,086
Other media characteristics						
Year of media creation	1975	39	1945	1986	2008	2018
Year of website creation	2004	7	1998	2004	2010	2018
Year of paywall introduction	2014	5	2013	2015	2018	2020
Number of journalists	147	183	32	90	208	1121
Observations	205					

Notes: The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for media outlets. The observations are at the media outlet/day level for the online audience statistics, and at the media outlet level for the content data and other media characteristics.



Notes: The Figure reports the share of the media outlets in our sample depending on their online business model. 48.1% of the media in our sample do not have a paywall ("no paywall"), and 5.1% condition the reading of the paid articles on the fact of watching an ad ("paid articles can be accessed by watching an ad"). Of the outlets that do have a paywall, we distinguish between three models: hard paywall, metered paywall, and soft paywall ("some articles locked behind paywall").

Figure 1: News editors' business model

Second, to better understand the mechanisms that may be at play, we collect additional information on the media: (i) their year of creation, (ii) the year of creation of their website (2004 on average), as well as (iii) information on their business model. In particular, for each of the media outlets, we investigate whether it uses a paywall, the characteristics of this paywall (e.g. soft vs. hard), and the date of introduction of the paywall. This information is summarized in Figure 1: while 49% of the media outlets do not have a paywall, 42.4% lock at least some of their articles behind a paywall (soft paywall). Metered paywalls and hard paywalls are much less frequent. The media outlets that use a paywall introduced it on average in 2014. Overall, the large majority of the media outlets in our sample rely at least partly on advertising revenues; however, some of them do not (e.g. the pure online media Mediapart).

Third, given that media outlets may react differently to social media depending on their initial investment in quality (see e.g. de Cornière and Sarvary, 2019), we also compute information on the size of the newsroom and on the average payroll (Cagé, 2016). This information is available for 68 media outlets in our sample. Finally, for the 72 media outlets for which this

Table 4: Summary statistics: Mainstream media articles

	Mean	St.Dev	P25	Median	P75	Max
Length						
Length (number of characters)	2,420	2,224	1,125	1,984	3,184	431,812
Facebook shares						
Number of shares on Facebook	25	1,157	0	0	1	466,114
Number of comments on Facebook	31	540	0	0	0	151,588
Number of reactions on Facebook	121	2,039	0	0	0	643,066
Observations	929,764					

Notes: The table gives summary statistics. Time period is July 2018 - September 2018. Variables are values for the mainstream media articles. The observations are at the article level.

information is available, we collect daily audience information from the ACPM, the French press organization whose aim is to certify circulation and audience data. The average number of daily visits is $586,269$, and the average number of page views $1,510,024$.

Article characteristics Table 4 presents summary statistics for the $929,764$ articles included in our dataset. On average, articles are $2,420$ characters long. For each of these articles, we also compute its originality rate, i.e. the share of the article that is "original" and the share that is copy-pasted (Cagé et al., 2020). Furthermore, to proxy for the audience received by each of these articles, we compute the number of times they are shared on Facebook.[24]

In the remainder of this section, we describe the new algorithms we develop to analyze these two datasets, and in particular to identify the events on Twitter and investigate how they interact with mainstream media events.

2.3 Event detection

An event is defined as a set of documents belonging to the same news story. E.g. all the documents (tweets and media articles) discussing the Hokkaido Easter Iburi earthquake on September 6 will be classified as part of the same event. Events are detected by our algorithms using the fact that the documents share sufficient semantic similarity.

2.3.1 Algorithms

Twitter has been used to detect or predict a large variety of events, from flood prevention (de Bruijn et al., 2017) to stock market movements (Pagolu et al., 2016). However, the

[24]Unfortunately, article-level audience data is not available. The number of shares on Facebook is an imperfect yet relevant proxy for this number, as shown in Cagé et al. (2020).

specificity of social network data (short texts, use of slang, abbreviations, hashtags, images and videos, very high volume of data) makes all "general" detection tasks (without specification of the type of topic) very difficult on tweet datasets. Hence, we develop a new algorithm to detect events on our dataset that we describe briefly here (see Mazoyer et al., 2020b, for more details and a description of the state-of-the-art in the computer science literature). Importantly, we have tested a number of different algorithms[25]; in this paper, we detect the social media events using the algorithm whose performance has been shown to be the best.

In a nutshell, our approach consists in modeling the event detection problem as a dynamic clustering problem, using a "First Story Detection" (FSD) algorithm.[26] The parameters of this algorithm are w, the number of past tweets among which we search for a nearest neighbor, and t, the distance threshold above which a tweet is considered sufficiently distant from past tweets to form a new cluster. The value of w is set to approximately one day of tweets history, based on the average number of tweets per day. We set the value of t so as to maximize the performance of our clustering task.[27]

Regarding the type of tweet representation, we similarly test several embeddings in our companion work, including the representation of both text and images (using TF-IDF, Word2Vec, ELMo, BERT and Universal Sentence Encoder) (Mazoyer et al., 2020b). Online Appendix Table B.1 summarizes the different preprocessing steps we apply to the text before using it as an input in the different models. The best performance is obtained with the TF-IDF model; online Appendix Figure D.4 reports the Best Matching F1 score for the different embeddings used with the FSD algorithm, depending on the threshold t. The relative performance of the different models was evaluated using two datasets: the corpus by McMinn et al. (2013), as well as a corpus of 38 millions original tweets collected from July 15th to August 6th 2018 that we manually annotated.[28] Our algorithm is considered state-of-the-art in the field of event detection in a stream of tweets, and our code and data are published and publicly available on GitHub.

To detect the news events among the stories published online by traditional media outlets, we follow Cagé et al. (2020). For the sake of space here, we do not enter into the details of the

[25]In particular, the "Support Vector Machines", the "First Story Detection", the DBSCAN, and the "Drichlet Multinomial Mixture" algorithms. Their relative performance are described in Mazoyer et al. (2020b).

[26]The superiority of the "First Story Detection" algorithm over topic modeling techniques such as Dirichlet Multinomial Mixture model (Yin and Wang, 2014), or a standard algorithm such as DBSCAN (Ester et al., 1996), most probably comes from the very rapid evolution over time of the vocabulary used to talk about a given event. Dirichlet Multinomial Mixture model and DBSCAN are not designed to take this temporal evolution into account, unlike the FSD algorithm, which allows a gradual evolution of clusters over time.

[27]Generally speaking, lower t values lead to more frequent clustering, and thus better intra-cluster homogeneity (better precision), but may lead to over-clustering (lower recall). Clustering performance is evaluated by using the "best matching F1". This measure is defined by Yang et al. (1998): we evaluate the F1 score of each pair between clusters (detected) and events (annotated). Each event is then matched to the cluster for which the F1 score is the best. Each event can be associated to only one cluster. The best matching F1 thus corresponds to the average of the F1s of the cluster/event pairs, once the matching is done.

[28]More precisely, we hired three graduate political science students to annotate the corpus.

algorithm we use. Roughly speaking, just as for social media, we describe each news article by a semantic vector (using TF-IDF) and use the cosine similarity measure to cluster the articles to form the events based on their semantic similarity.

2.3.2 Interacting social media events and mainstream media events

The aim of this article is to investigate whether mainstream media react to the popularity of a story on social media, and to measure the extent to which it affects their coverage. To generate the intersection between social media events and mainstream media events we proceed as follows: (i) we compute cosine similarity for each (MME,SME) vector pair; (ii) we build the network of events, where two events are connected if $cosine\_sim(MME, SME) > c$ ($c = 0.3$ gives the best results on annotated data) and $time\_difference(MME, SME) \leqslant d$ (d =4 days gives the best results on annotated data); (iii) we use the Louvain community detection algorithm (Blondel et al., 2008) on this network to group together close events. The Louvain algorithm uses a quality metric Q that measures the relative density of edges inside communities with respect to edges outside communities. Each node is moved from community to community until the best configuration (where Q is maximum) is found. We provide more details in the online Appendix Section B.3 and illustrate the different steps in Figure B.2.

We obtain $5,137$ joint events that encompass over 32 million tweets and $273,000$ news articles. Table 5 presents summary statistics on these events that contain on average $6,283$ tweets and 53 media articles published by 17 different media outlets. Of these $5,137$ joint events, $4,392$ break first on Twitter. These articles will be the focus of our analysis in Section 3 below. Their characteristics partly differ from those of the events that appear first on mainstream media, as reported in online Appendix Table C.3 where we perform a t-test on the equality of means. In particular, they tend to last longer and receive slightly more media coverage.

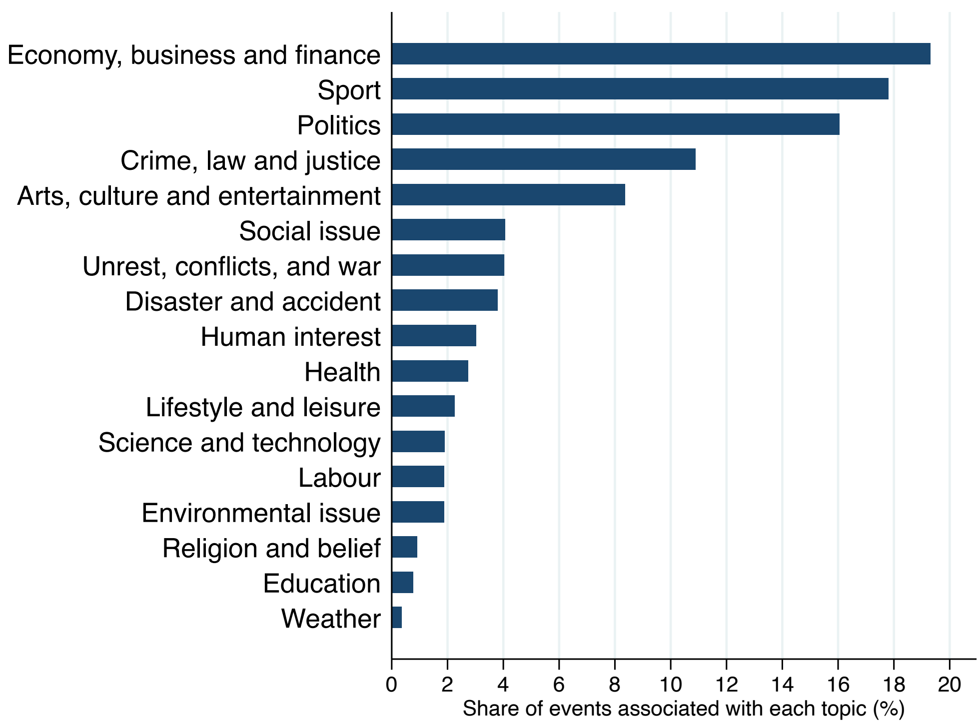
Using the metadata associated with the AFP distaches, we identify the topic of the joint events. The AFP uses 17 IPTC classes to classify its dispatches. These top-level media topics are: (i) Arts, culture and entertainment; (ii) Crime, law and justice; (iii) Disaster and accidents; (iv) Economy, business and finance; (v) Education; (vi) Environment; (vii) Health; (viii) Human interest; (ix) Labour; (x) Lifestyle and leisure; (xi) Politics; (xii) Religion and belief; (xiii) Science and technology; (xiv) Society; (xv) Sport; (xvi) Conflicts, war and peace; and (xvii) Weather.[29] Figure 2 plots the share of events associated with each media topic (given that some events are associated with more than one topic, the sum of the shares is higher than 100%). Nearly one fifth of the events are about "Economy, business and finance", 18% about "Sport", followed by "Politics"."Crime, law and justice" comes fourth.

[29] To define the subject, the AFP uses URI, available as QCodes, designing IPTC media topics (the IPTC is the International Press Telecommunications Council). These topics are defined precisely in the online Appendix

Table 5: Summary statistics: Joint events

	Mean	St.Dev	P25	Median	P75	Max
Length of the event (in hours)	555	507	161	369	819	2,350
Number of documents in event	6,336	94,618	209	761	2,995	6,486,415
Twitter coverage						
Nb of tweets in event	6,283	94,579	192	728	2,928	6,484,204
Number of different Twitter users	3,288	17,902	163	584	2,220	1,077,929
Average number of retweets of tweets in events	3.2	5.6	1.0	1.8	3.4	149
Average number of replies of tweets in events	0.4	0.5	0.1	0.2	0.5	17
Average number of favorites of tweets in events	4.5	8.0	1.0	2.3	5.0	240
Media coverage						
Number of news articles in the event	53	131	13	20	44	4,538
Number of different media outlets	17	12	9	13	22	98
Observations	5,137					

Notes: The table gives summary statistics. Time period is July 2018 - September 2018. The observations are at the event level.



Notes: The figure shows the share of joint events associated with each media topic. The topics correspond to the IPTC media topics described in the text and defined in the online Appendix. Because some events are associated with more than one topic, the sum of the shares is higher than 100%.

Figure 2: Share of the joint events associated with each media topic

There are several reasons why an event may appear on Twitter first, before being covered by mainstream media. First, an event may be described by a media outlet outside of our corpus, such as an English language news media, then be picked by Twitter users before being relayed by the French media. Second, some Twitter users can witness a "real world" event and film it or talk about it on the social network. This was the case with "the battle of Orly airport" in July 2018, when two famous French rappers, Booba and Kaaris, got into a fight inside a duty-free store. Third, some events may originate solely on Twitter, like online demonstrations. For example, in the summer of 2018, far-left MP Jean-Luc Mélenchon called for a boycott of Emmanuel Macron's speech at the Palace of Versailles by spreading the hashtag *#MacronMonarc* with other activists of his party. Sympathizers were invited to tweet massively on the subject using this hashtag. The event was then relayed by mainstream media. Another example in our corpus is the hashtag #BalanceTonYoutubeur: this hashtag spread in August 2018 with accusations that certain French Youtube influencers had raped minors.

2.4 Measures of popularity on Twitter and of media coverage

To proxy the popularity of an event on Twitter, we rely on the activity on Twitter and count, in each event, the total number of tweets and the total number of unique users who tweet about the event. For the tweets, we distinguish between the original tweets and the retweets and replies. Furthermore, we compute the average number of followers of the news breaker or seed of the event.

Importantly, to isolate the specific impact of the popularity on Twitter on mainstream media coverage, we focus on what happens on social media *before* the publication of the first news article; to do so, we compute similar measures but focus only on Twitter activity before the first mainstream media article appears. In the IV strategy described below (Section 3.2), to instrument for this popularity, we also compute the average number of interactions generated by the previous tweets of the seeds' followers.[30]

To study the intensity of mainstream media coverage, we look at different dimensions. First, we use quantitative measures of coverage, in particular the number of articles a given media outlet devotes to the event, as well as the length of these articles. Studying both the intensive and the extensive margin of coverage is of particular importance given that some media outlets can choose to skip some events while others may decide to follow a more systematic approach. This may depend on the characteristics of the media outlet, but also on the topic of the event.

Section A.3. An event can be associated with more than one top-level media topic.

[30] To compute this number of interactions, we also include in our dataset of tweets the period between June 15th 2018 and June 30th 2018. We do so to ensure that we observe at least some tweets before the first event.

Second, we use more "qualitative" measures of coverage, in particular the originality of the articles (following Cagé et al., 2020) and the "reactivity" of the media (i.e. the time it takes the media to cover the event, with respect to both the other media outlets and the beginning of the events).

3 Popularity on social media and news editor decisions: Empirical strategy

In the remainder of the paper, we tackle the following question: does the popularity of a story on social media affect, everything else equal, the coverage that mainstream media devote to this story? While the drivers of news editors' decisions remain essentially a black box, understanding the role played by social media in these decisions is of particular importance. In this Section, we present the empirical strategy we develop to tackle this question; the empirical results are presented in the following Section 4.

3.1 Naive approach

We begin by estimating the correlation between an event popularity on Twitter and its mainstream media coverage. We do so both at the event level and the media level.

Event-level approach At the event-level, we perform the following estimation:

$$\text{news coverage}_e = \alpha + \beta \, \text{number of tweets}_e + \mathbf{Z'_e}\gamma + \lambda_d + \omega_m + \epsilon_{emd} \tag{1}$$

where e index the events, d the day of the week (DoW) of the first tweet in the event, and m the month.

Our dependent variable of interest, news coverage$_e$, is the intensity of the media coverage that we proxy alternatively at the event level by the total number of articles devoted to the event and the number of different media outlets covering the event. Our main explanatory variable, number of tweets$_e$, is the total number of tweets in the event *before* the publication of the first media article, that captures the popularity of the event on Twitter.

$\mathbf{Z'_e}$ is a vector of controls that includes measures of the seed's number of followers (at the time of the event) as well as an indicator variable equal to one if the first tweet in the event is emitted during the night (between 00:00am and 06:59am) and to zero otherwise. We also control for DoW fixed effects (λ_d) and month fixed effects (ω_m).

Media-level approach Bias in editorial decisions may vary depending on the characteristics of the media outlet. Furthermore, some media outlets may decide to cover a news event while others do not. To investigate whether this is the case, we then exploit within media outlet variations (in this case, the standard errors are clustered at the event level). Our specification is as follows:

$$\text{news coverage}_{ec} = \alpha + \beta \text{ number of tweets}_e + \mathbf{Z_e'}\gamma + \delta_c + \lambda_d + \omega_m + \epsilon_{ecmd} \quad (2)$$

where c index the media outlets and the dependent variable, news coverage$_{ec}$, is now alternatively the number of articles published by media c in event e, a binary variable equal to one if the media devotes at least one article to the event and, conditional on covering the event, the average length and originality of the articles and the reactivity of the media. number of tweets$_e$ and $\mathbf{Z_e'}$ are defined as in equation (1). δ_c, λ_d and ω_m are respectively fixed effects for media, DoW and month.

3.2 IV approach

While estimating equations (1) and (2) allows us to analyze the relationship between the popularity of a story on Twitter and its coverage on mainstream media, the estimated relationship may be (at least partly) driven by the unobserved characteristics of the story, e.g. its "newsworthiness". Randomizing story popularity on Twitter is not feasible. To identify the causal effect of a popularity shock on Twitter, we thus need to find and exploit exogenous sources of variation in popularity. In this article, we propose a new IV strategy that relies on the centrality of the users in the Twitter network interacted with the news pressure at the time of the event. Specifically, our instrument is the interaction between the seed's centrality in the network and the news pressure at the time of the tweet, and we control for the direct effects of centrality and news pressure. Our identification assumption is that, once we control for the direct effects of centrality and news pressure, the interaction between the seed's centrality and news pressure should only affect traditional news production through its effect on the tweet's visibility on Twitter. This is conditional on controlling for the seed's number of followers, and we also make sure that our results are robust to dropping the events whose seed is the Twitter account of a media outlet or journalist and/or whose seed is a multiple news breaker, to avoid capturing a celebrity or influencer bias.

In exploiting the structure of the Twitter network, the intention of our identification strategy approach is to mimic a hypothetical experiment that would break the correlation between popularity and unobserved determinants of the story's intrinsic interest. Regarding centrality, our source of variation, in the spirit of an intention-to-treat approach, comes from the number of "impressions" generated by the previous tweets (i.e. before the beginning of the event) of the seed's followers. Figure 3 illustrates the intuition behind this strategy. We consider a simple case with two seeds who have an equal number of followers (sub-Figure 3a), but the followers of one of the seeds (on the left-hand side of sub-Figure 3b) have many more followers than those of the other seed (on the right-hand side of sub-Figure 3b). As a consequence, regardless of the content of the tweet itself, the tweets emitted by the left-hand side seed have a much lower probability of being retweeted than the tweets emitted by the right-hand side seed, everything else equal (sub-Figure 3c). Indeed, everything else equal and regardless of the interest of a given tweet, the higher the number of impressions, the higher the potential number of retweets.

We measure news pressure by the number of interactions generated by all the tweets published in the hour preceding the first tweet in the event. We use the total number of retweets rather than the total number of tweets, because the Twitter API may restrict the number of delivered tweets if it exceeds the threshold of 1% of the global volume. The number of retweets, on the contrary, is known at any time because it is a metadata provided with each tweet. Online Appendix Figure D.5 illustrates our IV strategy: for each minute, we compute the average number of retweets generated by the tweets published during the previous hour.

Formally, given both our dependent variable (the number of articles) and our main (endogenous) explanatory variable (the number of tweets) are count variable, we rely on Wooldridge (2002, 2013)'s control functions approach. Our first stage and reduced-form specifications, respectively, are (for the event-level approach):

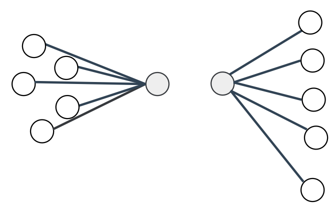
$$\begin{aligned}\text{number of tweets}_e = \beta_1 \text{centrality}_e \times \text{low news pressure}_e + \beta_2 \text{centrality}_e + \beta_3 \text{low news pressure}_e \\ + \mathbf{X_e'}\gamma + \lambda_d + \omega_m + \epsilon_{emd} \quad (3)\end{aligned}$$

$$\begin{aligned}\text{number of articles}_e = \beta_1 \text{centrality}_e \times \text{low news pressure}_e + \beta_2 \text{centrality}_e + \beta_3 \text{low news pressure}_e \\ + \mathbf{X_e'}\gamma + \lambda_d + \omega_m + \epsilon_{emd} \quad (4)\end{aligned}$$

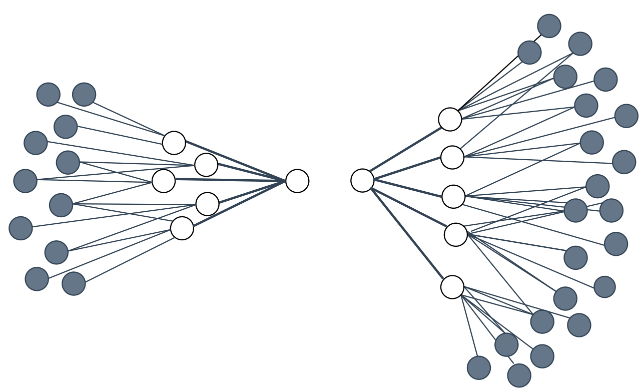
where, in the first stage, number of tweets$_e$ is as before the total number of tweets published in the event e before the publication of the first media article. The instrument centrality$_e$ $\times$ low news pressure$_e$ is relevant if $\beta_1 > 0$.

To obtain the second-stage estimates, we follow a two-step procedure. First, we estimate equation (3) using a negative binomial and obtain the reduced form residuals $\widehat{\epsilon_{emd}}$. Then, we estimate the following model:

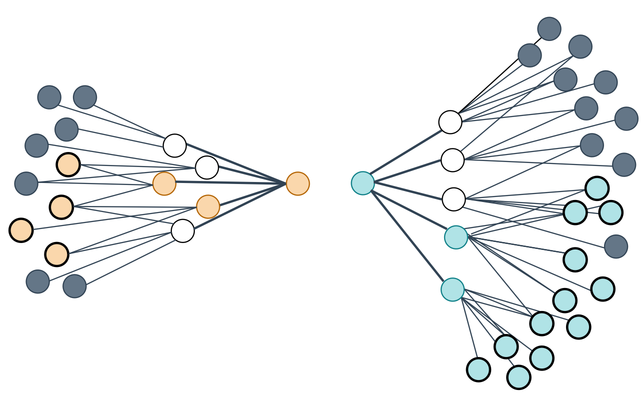
$$\begin{aligned}\text{number of articles}_e = \zeta_1 \text{number of tweets}_e + \zeta_2 \text{centrality}_e + \zeta_3 \text{low news pressure}_e \\ + \zeta_4 \widehat{\epsilon_{emd}} + \mathbf{X_e'}\eta + \lambda_d + \omega_m + \varepsilon_{emd} \quad (5)\end{aligned}$$



(a) **2 seeds with an equal number of followers...**



(b) **... but different numbers of followers' followers**



(c) **, leading to different numbers of tweet impressions**

Notes: The Figure illustrates the measure of centrality we develop in this article, in the spirit of an intention-to-treat approach. The thought experiment is described in details in the text.

Figure 3: Measuring seed's centrality: Illustration

with bootstrapped standard errors. Besides, for robustness, we show that our results are robust to rather using a Generalized Method of Moments estimator of Poisson regression and instrumenting our endogenous variable number of tweets_e by the excluded instrument $\text{centrality}_e \times$ low news pressure_e (Nichols, 2007).

4 Popularity on social media and news editor decisions: Results

In this section, we first present the results of the naive estimations (without instrumenting for event popularity). We then turn to the IV analysis before discussing the heterogeneity of the effects. Each time, we start with the estimates we obtain when performing the analysis at the event level, before turning to the media-level estimations.

4.1 Naive estimates

Event-level analysis Table 6 reports the results of the estimation of equation (1) (event-level analysis). In Columns (1) to (4) the outcome of interest is the total number of articles published in the event, and in Columns (5) to (8) the number of unique media outlets which cover the event. Given the dependent variables are count variables, we first use a negative binomial. We find a positive correlation between the number of tweets published in an event before the first news article ("number of tweets") and the total number of articles published in the event: an increase of $1,000$ in the number of tweets published in the event before the first media article is associated with 3.1 additional articles (Column (1)); this finding is robust to dropping the events whose seed is the Twitter account of a media outlet or journalist and the events whose seed is the Twitter account of a multiple news breaker (as defined above) (Column (4)).

The positive relationship between the popularity of the event on Twitter and the media coverage is driven both by the intensity of the coverage (number of articles) and by the extensive margin: we also find a positive relationship with the number of unique media outlets covering the event (Columns (5) and (8)). An increase of 1,000 in the number of original tweets published in the event before the first media article is associated with an increase of 0.4 in the number of media outlets dealing with the event. These findings are robust to the use of different sets of controls.

Table 7 shows that our results are robust to using a linear-log model. Regarding the magnitude of the effect, we find that a one-percent increase in the number of tweets before the first media article is associated with a 8.89% increase in the number articles (Column (1)). This finding is no affected by changes in the set of controls. The magnitude of the effect slightly decreases however if we drop the events that are broken by media outlets or journalists

Table 6: Naive estimates: Event-level approach, Negative binomial model

	Number of articles				Number of media			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Number of tweets	0.055***	0.055***	0.055***	0.055***	0.021***	0.021***	0.021***	0.021***
	(0.014)	(0.014)	(0.014)	(0.015)	(0.005)	(0.005)	(0.005)	(0.005)
Log # seed's followers	-0.011			-0.011	-0.011***			-0.011***
	(0.011)			(0.011)	(0.004)			(0.004)
# seed's followers		-0.004	-0.004			-0.008	-0.008	
		(0.021)	(0.021)			(0.010)	(0.010)	
# seed's followers-squared		-0.000	-0.000			0.000	0.000	
		(0.001)	(0.001)			(0.000)	(0.000)	
=1 if first tweet during night			0.048				0.039	
			(0.090)				(0.030)	
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Drop media				✓	✓	✓	✓	✓
Drop multiple				✓	✓	✓	✓	✓
Observations	4,411	4,411	4,411	4,332	4,411	4,411	4,411	4,332
Marginal Effect (tweets)	3.1	3.1	3.1	3.1	0.4	0.4	0.4	0.4

Notes: * $p<0.10$, ** $p<0.05$, *** $p<0.01$. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation (robust standard errors are reported between parentheses). An observation is a news event. We only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week and month fixed effects. Columns (1) to (3) and (5) to (7) report the estimates for all the events that appear first on Twitter; in Columns (4) and (8) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event and is given in thousands. More details are provided in the text.

(Column (4)). Regarding the number of media outlets covering the event (Columns 5 to 8), we find that a 1% increase in the number of tweets is associated with a 1.36% increase in this number.

In Table 8, we show that the magnitude of this relationship varies with the topic of the event (as defined using the IPTC classes provided by the AFP). In particular, we find that the magnitude of the effect is the strongest for sport and politics, while it is relatively less important for events related to economy as well as culture.

Media-level analysis Table 9 shows the estimates when we perform the media-level analysis (estimation of equation (2)). The unit of observation is a media-event, and all the media outlets are included in the estimation (even if they do not cover the event; then the value of the number of articles in the event for them is equal to zero). Consistently with the results of Table 6, we find a positive relationship between the popularity of the event on Twitter and the media coverage it receives (Column (1)), and this relationship is robust to dropping the events whose seed is the Twitter account of a media / journalist / multiple news breaker (Column (2)). An increase of $1,000$ in the number of tweets published in the event before the first media article appears increases the average number of articles that *each media outlet* publishes in the event by 0.013.

In Columns (3) and (4), our dependent variable is an indicator variable equal to one if the media outlet publishes at least one article in the event (we use a Probit model to perform the estimation). We find that the probability that a media covers the event is also positively associated with the number of tweets in the event before the first media article.

Next, we focus on the media outlets that devote at least one article to the event and investigate the magnitude of the coverage, *conditional on covering* the event. Indeed, the popularity of an event on Twitter can play a role both at the intensive and at the extensive margin. Table 10 presents the results. Columns (1) and (2) present the estimations for the number of articles published by the media in the event: we see that the positive impact of popularity appears not only at the extensive margin but also at the intensive margin; there is a positive correlation with the number of articles published (conditional on publishing at least one article in the event). An increase of $1,000$ in the number of tweets is associated with an increase of 0.06 in the number of articles published by each media outlet. However, there is no statistically significant relationship with the length of the article(s) published by the media (Columns (3) and (4)). The length of the articles can be considered a proxy – albeit an imperfect one – for the quality of the articles. An alternative proxy for quality is the originality of the articles (Cagé et al., 2020); results are reported in the online Appendix. We will come back to these points in the mechanisms section below.

Finally, we investigate whether our results vary depending on the offline format of the

Table 7: Naive estimates: Event-level approach, Linear-Log model

	Number of articles				Number of media			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log # tweets	8.89***	8.89***	8.89***	8.16***	1.36***	1.36***	1.37***	1.30***
	(0.93)	(0.93)	(0.93)	(0.90)	(0.10)	(0.10)	(0.10)	(0.10)
Log # seed's followers	-0.23			-0.56	-0.17**			-0.22***
	(0.59)			(0.62)	(0.07)			(0.08)
# seed's followers		0.04	0.02			-0.09	-0.10	
		(1.15)	(1.15)			(0.16)	(0.16)	
# seed's followers-squared		-0.02	-0.02			0.00	0.00	
		(0.04)	(0.04)			(0.01)	(0.01)	
=1 if first tweet during night			2.74				0.78	
			(4.90)				(0.52)	
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Drop media				✓				✓
Drop multiple				✓				✓
Observations	4,411	4,411	4,411	4,168	4,411	4,411	4,411	4,168

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a linear-log model (robust standard errors are reported between parentheses). An observation is a news event. We only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week and month fixed effects. Columns (1) to (3) and (5) to (7) report the estimates for all the events that appear first on Twitter; in Columns (4) and (8) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event and is given in thousands. More details are provided in the text.

Table 8: Naive estimates: Event-level approach, Depending on the topic of the event

	Economy	Sport	Politics	Crime	Culture
	(1)	(2)	(3)	(4)	(5)
Number of tweets	0.055	0.100***	0.116***	0.086**	0.002
	(0.042)	(0.023)	(0.038)	(0.044)	(0.010)
Log # seed's followers	-0.028	-0.040*	0.006	-0.014	-0.068*
	(0.024)	(0.023)	(0.024)	(0.030)	(0.035)
Month & DoW FEs	✓	✓	✓	✓	✓
Drop media	✓	✓	✓	✓	✓
Drop multiple	✓	✓	✓	✓	✓
Observations	684	662	574	334	283
Marginal Effect (tweets)	2.8	6.7	8.6	4.7	0.1

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation (robust standard errors are reported between parentheses). An observation is a news event. We only consider the subset of news events that appear first on Twitter, and we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). All specifications include day-of-the-week and month fixed effects. The number of tweets is computed *before* the first news article in the event and is given in thousands. The topics correspond to the IPTC topics described in the text and defined in the online Appendix. More details are provided in the text.

Table 9: Naive estimates: Media-level approach

	Number of articles		=1 if at least one article	
	(1)	(2)	(3)	(4)
Number of tweets	0.047***	0.047***	0.015***	0.015***
	(0.011)	(0.011)	(0.003)	(0.003)
Log # seed's followers	-0.016	-0.017	-0.009***	-0.009***
	(0.011)	(0.011)	(0.003)	(0.003)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Drop media		✓		✓
Drop multiple		✓		✓
Model	Neg bin	Neg bin	Probit	Probit
Observations	839,063	823,974	839,063	823,974
Clusters (events)	4,393	4,314	4,393	4,314
Marginal Effect	0.013	0.013	0.002	0.002

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation in Columns (1) and (2) and a Probit in Columns (3) and (4). Standard errors are clustered at the event level. An observation is a media-news event. We only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, month, and media fixed effects. Columns (1) and (3) report the estimates for all the events that appear first on Twitter; in Columns (2) and(4) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). In Columns (1) and (2), the dependent variable is the number of articles published by the media in the event. In Columns (3) and (4) the dependent variable is an indicator variable equal to one if the media outlet publishes at least one article in the event and to zero otherwise. The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

Table 10: Naive estimates: Media-level approach, Conditional on covering the event

	Number of articles		Articles length	
	(1)	(2)	(3)	(4)
Number of tweets	0.02***	0.02***	1.17	1.52
	(0.01)	(0.01)	(2.55)	(2.61)
Log # seed's followers	-0.00	-0.00	3.51	2.84
	(0.01)	(0.01)	(4.12)	(4.17)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Drop media		✓		✓
Drop multiple		✓		✓
Model	Neg bin	Neg bin	Linear	Linear
Observations	76,703	75,256	76,694	75,246
Clusters (events)	4,393	4,314	4,393	4,314
Marginal Effect	0.06	0.05		

Notes: * $p<0.10$, ** $p<0.05$, *** $p<0.01$. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation in Columns (1) and (2), and OLS in Columns (3) to (4). Standard errors are clustered at the event level. An observation is a media-news event, and only the media outlets that devote at least one article to the event are included. Further, we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, month, and media fixed effects. Columns (1) and (3) report the estimates for all the events that appear first on Twitter; in Columns (2) and (4) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). In Columns (1) and (2), the dependent variable is the number of articles published by the media outlet; in Columns (3) and (4) the average length of these articles. The number of tweets is computed *before* the first news article in the event apears and is given in thousands. More details are provided in the text.

Table 11: Naive estimates: Media-level approach, Depending on the media offline format

	Nat. dail.	Local dail.	Weeklies	Pure online	TV	Radio
	(1)	(2)	(3)	(4)	(5)	(6)
Number of tweets	0.043***	0.042***	0.048***	0.058***	0.094***	0.041***
	(0.013)	(0.013)	(0.011)	(0.013)	(0.024)	(0.012)
Log # seed's followers	-0.007	-0.019	-0.016	-0.030**	-0.012	-0.007
	(0.011)	(0.015)	(0.011)	(0.013)	(0.013)	(0.012)
Media FEs	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media	✓	✓	✓	✓	✓	✓
Drop multiple	✓	✓	✓	✓	✓	✓
Observations	47,454	120,792	155,304	280,410	30,198	47,454
Clusters (events)	4,314	4,314	4,314	4,314	4,314	4,314
Marginal Effect	0.028	0.016	0.012	0.005	0.107	0.017

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation. Standard errors are clustered at the event level. An observation is a media-news event. We only consider the subset of news events that appear first on Twitter, and drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). All specifications include day-of-the-week, month, and media fixed effects. In Column (1), we only consider the national daily newspapers, in Column (2) the local daily newspapers, in Column (3) the weekly newspapers, in Column (4) the pure online media, in Column (5) the television channel websites, and in Column (6) the radio station websites. The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

media outlets we consider. Table 11 reports the estimates separately for the national daily newspapers, the local daily newspapers, the weeklies, the pure online media, the websites of the television channels and the websites of the radio stations. First, it is interesting to note that the estimates we obtain are positive and statistically significant for all the offline formats, despite the lower number of observations.

Second, if we turn to the marginal effects, we find that the relationship between the popularity of an event on Twitter and the media coverage it receives is the strongest for the websites of the television stations, followed by the national daily newspapers. Perhaps surprisingly, the magnitude of the correlation is not higher for the pure online media than for the other media outlets. Yet, we will see below that the intensity of the *social media presence* is an important factor explaining why some media outlets react more to what is happening on Twitter than others.

4.2 IV estimates

The previous estimates may suffer from the fact that the positive relationship between a story's popularity on Twitter and its media coverage can both be driven by the intrinsic interest of the story, regardless of what happens on social media. Hence, in this section, we report the results of the IV estimates following the strategy described in Section 3.2 above. As before, we first report the results of the estimations we obtain at the event level, before turning to the media-level approach.

Event-level approach In Table 12, we begin by reporting the results of the reduced-form estimation (equation (4)) in Columns (1) and (2). The coefficient we obtain for our instrument is positive and statistically significant: the number of previous interactions of the tweets of the seed's followers is positively associated with the number of articles published in the event when the news pressure is low. In Columns (3) to (6), we turn to our two-step procedure. In the first stage, we obtain as expected a positive relationship between the number of tweets in the event (before the first media article appears) and the previous interactions of the tweets of the seed's followers when news pressure is low (Columns (3) and (4)). Columns (5) and (6) show that the findings of Table 6 are robust to controlling for the endogeneity of the number of tweets.

Given that the control function method suffers from some drawbacks, we then show that our results are robust to estimating a Poisson regression model with endogenous regressors using generalized method of moments (GMM) (Windmeijer and Silva, 1997).[31] Table 13 reports the results. In Columns (1) and (2), our endogenous explanatory variable, instrumented as before by $\text{interactions}_e \times \text{news pressure}_e$, is the total number of tweets in the event before the first media article, and our dependent variable the number of articles published in the event. An increase of $1,000$ in this number is associated with an increase of 2 in the number of news articles published in the event. As expected given the direction of the omitted variable bias, the magnitude of the IV effects is smaller than the one of the naive estimates (increase of 3.1 reported in Table 6).

We also find that the increase in the number of media outlets covering the event is robust to instrumenting the number of tweets, but likewise of smaller magnitude (Columns (3) and (4)). According to our estimates, an increase of $1,000$ in the number of tweets before the publication of the first media article is associated with an increase of 0.37 - 0.38 in the number of outlets covering the event.

Media-level approach We next turn to the media-level estimates. Table 14 reports the results of the estimation when we use the Control Function method. As before, in Columns (1) and (2) we report the results of the reduced-form estimation: consistently with the results we obtain when performing the event-level analysis, we find that our instrument is positively associated with the number of articles published by each of the media outlets in the event (controlling for media fixed effects, month fixed effects, and day-of-the-week fixed effects, with

[31] We use the Stata's ivpoisson command.

Table 12: IV estimates: Event-level approach, Control Function method

	Reduced form		First stage		Second stage	
	(1)	(2)	(3)	(4)	(5)	(6)
	Nb articles	Nb articles	Nb tweets	Nb tweets	Nb articles	Nb articles
Instrument						
Low pressure * Centrality	0.03*	0.04**	0.14***	0.14***		
	(0.02)	(0.02)	(0.05)	(0.05)		
Controls						
Low pressure	0.18**	0.17**	0.17	0.17	0.17**	0.17**
	(0.09)	(0.09)	(0.21)	(0.21)	(0.08)	(0.08)
Centrality	-0.03***	-0.03***	-0.02	-0.03	-0.02	-0.02
	(0.01)	(0.01)	(0.03)	(0.03)	(0.01)	(0.01)
Log # seed's followers	-0.01	-0.00	-0.04	-0.04	-0.00	-0.00
	(0.01)	(0.01)	(0.04)	(0.04)	(0.02)	(0.01)
Residuals (1st stage)					0.02	0.02
					(0.10)	(0.10)
Second stage						
Number of tweets					0.06***	0.06***
					(0.02)	(0.02)
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media		✓		✓		✓
Drop multiple		✓		✓		✓
Standard errors	Robust	Robust	Robust	Robust	Bootstrap	Bootstrap
Observations	3,435	3,374	3,435	3,374	3,435	3,374
Marginal effect (tweets)					3.3	3.3

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. An observation is a news event. Robust standard errors are reported between parentheses in Columns (1) to (4). In Columns (5) and (6) we report bootstrapped standard errors. All specifications include day-of-the-week and month fixed effects. Columns (1) and (2) report the results of the reduced form estimation (the dependent variable is the number of articles), Columns (3) and (4) of the first stage (the dependent variable is the number of tweets), and Columns (5) and (6) of the second stage (the dependent variable is the number of articles). In Columns (2), (4) and (6) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

Table 13: IV estimates: Event-level approach, IV Poisson GMM

	Number of articles		Number of media	
	(1)	(2)	(3)	(4)
Number of tweets	0.040***	0.041***	0.022**	0.022**
	(0.010)	(0.010)	(0.009)	(0.009)
Low pressure	0.171*	0.164*	0.087***	0.083***
	(0.091)	(0.092)	(0.030)	(0.031)
Centrality	-0.037**	-0.038**	0.003	0.003
	(0.017)	(0.017)	(0.005)	(0.005)
Log # seed's followers	0.003	0.005	-0.009*	-0.009*
	(0.017)	(0.017)	(0.005)	(0.005)
Month & DoW FEs	✓	✓	✓	✓
Drop media		✓		✓
Drop multiple		✓		✓
Observations	3,435	3,374	3,435	3,374
Marginal Effect (tweets)	2.03	2.06	0.37	0.38

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a a generalized method of moments (GMM) Poisson regression model with endogenous regressors (Stata's ivpoisson command). An observation is a news event. Robust standard errors are reported between parentheses. All specifications include day-of-the-week and month fixed effects. In Columns (1) and (2), the dependent variable is the number of articles published in the event. In Columns (3) and (4), the dependent variable in the number of different media outlets covering the event. In Columns (2) and (4) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

standard errors clustered at the level of the events). The estimated coefficients are statistically significant at the five-percent level.

Columns (3) and (4) report the first-stage estimates of our two-step procedures, and Columns (5) and (6) the second-stage results. In the first stage, we show that the number of tweets in the event increases with the number of interactions of the previous tweets of the seed's followers when the news pressure is low. In the second stage, we find a positive and statistically significant effect of the popularity of the event on Twitter on the media coverage; the effect is of the same order of magnitude than the one we obtain with the naive estimates, and is robust to dropping the events whose seed is the Twitter account of a media outlet / journalist / multiple news breaker (as previously defined).

We then show that our results are robust to rather estimating a Poisson regression model with endogenous regressors. Table 15 presents the results. We estimate the effect separately for each offline media category (as in Table 11). We find that the positive relationship between popularity on Twitter and media coverage holds for all the offline types, expect radio stations, but this might be due to the relatively lower number of observations. The magnitude of the effect is the strongest for national daily newspapers and for the websites of the television stations.

Finally, in Table 16, we focus on the media outlets that devote at least one article to the event and investigate the magnitude of the coverage, conditional on covering the event. Consistently with the previous results, we show that our effects are robust to instrumenting for the popularity of events on Twitter, and that the magnitude of the IV estimates is slightly lower than the one of the naive estimates. If we consider the national daily newspapers, we find that an increase of $1,000$ in the number of tweets is associated with an increase of 0.055 in the number of articles published in the event by the newspapers that devote at least one article to the event. Furthermore, we find no effect on the length of the articles published by the media outlets.

5 Mechanisms

In the previous section, we have documented a positive relationship between the popularity of an event on Twitter and the coverage it receives on mainstream media. In this section, we discuss the different mechanisms at play behind this relationship.

5.1 Journalists monitor Twitter

First, the fact that a number of stories appear first on Twitter and that we observe high reactivity of the mainstream media might be due to the fact that journalists are closely monitoring Twitter. A growing literature in journalism studies indeed highlights the fact

Table 14: IV estimates: Media-level approach, Control Function method

	Reduced form		First stage		Second stage	
	(1)	(2)	(3)	(4)	(5)	(6)
	Nb articles	Nb articles	Nb tweets	Nb tweets	Nb articles	Nb articles
Instrument						
Low pressure * Centrality	0.044**	0.045**	0.140***	0.137***		
	(0.021)	(0.021)	(0.046)	(0.047)		
Controls						
Low pressure	0.148*	0.141*	0.177	0.172	0.139*	0.132
	(0.081)	(0.082)	(0.207)	(0.208)	(0.082)	(0.082)
Centrality	-0.025**	-0.026**	-0.023	-0.023	-0.017	-0.017
	(0.010)	(0.010)	(0.030)	(0.030)	(0.012)	(0.012)
Log # seeds followers	-0.009	-0.008	-0.041	-0.040	-0.005	-0.004
	(0.014)	(0.014)	(0.039)	(0.039)	(0.014)	(0.015)
Residuals (1st stage)					0.052	0.055
					(0.111)	(0.121)
Second stage						
Number of tweets					0.051***	0.050***
					(0.012)	(0.012)
Media FEs	✓	✓			✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media		✓		✓		✓
Drop multiple		✓		✓		✓
Observations	657,995	646,344	657,995	646,344	657,995	646,344
Clusters (events)	3,445	3,384	3,445	3,384	3,445	3,384
Marginal effect (tweets)					0.014	0.014

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. An observation is a media-news event. Standard errors are clustered at the event level. All specifications include the seed's number of followers as a control, and day-of-the-week, month, and media fixed effects. Columns (1) and (2) report the results of the reduced form estimation (the dependent variable is the number of articles), Columns (3) and (4) of the first stage (the dependent variable is the number of tweets), and Columns (5) and (6) of the second stage (the dependent variable is the number of articles). In Columns (2), (4) and (6) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

Table 15: IV estimates: Media-level approach, IV Poisson GMM, Depending on the offline format

	Nat. dail.	Local dail.	Weeklies	Pure online	TV	Radio
	(1)	(2)	(3)	(4)	(5)	(6)
Number of tweets	0.04***	0.03***	0.05***	0.06***	0.04***	-0.52
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(6.06)
Low pressure	0.13	0.21**	0.14	-0.06	0.09	0.30**
	(0.10)	(0.09)	(0.10)	(0.12)	(0.11)	(0.13)
Centrality	-0.03*	-0.02	-0.05**	-0.10**	-0.02	-0.00
	(0.02)	(0.02)	(0.02)	(0.05)	(0.01)	(0.01)
Log # seed's followers	0.01	0.00	0.00	0.00	-0.00	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Media FEs	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media	✓	✓	✓	✓	✓	✓
Drop multiple	✓	✓	✓	✓	✓	✓
Observations	37,224	94,752	121,824	219,960	23,688	37,224
Clusters (events)	3,384	3,384	3,384	3,384	3,384	3,384
Marginal effect (tweets)	0.022	0.011	0.011	0.005	0.016	-0.216

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a a generalized method of moments (GMM) Poisson regression model with endogenous regressors (Stata's ivpoisson command). An observation is a media-news event. We drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). Standard errors are clustered at the event level. All specifications include day-of-the-week, month, and media fixed effects. In Column (1), we only consider the national daily newspapers, in Column (2) the local daily newspapers, in Column (3) the weekly newspapers, in Column (4) the pure online media, in Column (5) the television channel websites, and in Column (6) the radio station websites. The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

Table 16: IV estimates: Media-level approach, IV Poisson GMM, Conditional on covering the event

	National daily newspapers				Television			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Nb articles	Nb articles	Length	Length	Nb articles	Nb articles	Length	Length
Number of tweets	0.02***	0.02***	-0.01	-0.01	0.03***	0.03***	0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Low pressure	0.07	0.07	0.02	0.02	-0.00	-0.01	0.02	0.02
	(0.08)	(0.08)	(0.02)	(0.02)	(0.09)	(0.09)	(0.02)	(0.02)
Centrality	-0.03***	-0.03***	0.00	0.00	-0.04**	-0.05***	-0.00	-0.00
	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)	(0.02)	(0.01)	(0.01)
Log # seed's followers	0.01	0.01	-0.00	-0.00	0.01	0.02	0.00	0.00
	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)	(0.02)	(0.00)	(0.00)
Media FEs	✓	✓	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Drop media		✓		✓		✓		✓
Drop multiple		✓		✓		✓		✓
Observations	8,398	8,250	8,398	8,250	3,902	3,825	3,902	3,825
Clusters (events)	3,188	3,128	3,188	3,128	2,208	2,164	2,208	2,164
Marginal effect (tweets)	0.055	0.057			0.071	0.073		

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a a generalized method of moments (GMM) Poisson regression model with endogenous regressors (Stata's ivpoisson command). An observation is a media-news event, and only the media outlets that devote at least one article to the event are included. Furthermore, we only consider the subset of news events that appear first on Twitter. Standard errors are clustered at the event level. All specifications include day-of-the-week, month, and media fixed effects. Columns (1) to (4) report the results for the national daily newspapers, and Columns (5) to (8) for television channel websites. In even columns, we drop the events whose seed is the Twitter account of a media or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

that social media play an important role as a news source. For example, von Nordheim et al. (2018) examine the use of Facebook and Twitter as journalistic sources in newspapers of three countries; they show that Twitter is more commonly used as a news source than Facebook.[32] Furthermore, McGregor and Molyneux (2018), who have conducted an online survey experiment on working U.S. journalists, show that journalists using Twitter as part of their daily work consider tweets to be as newsworthy as headlines from the Associated Press.

The use of Twitter crosses many dimensions of sourcing, information-gathering, and production of stories (Wihbey et al., 2019). Most media organizations actively encourage journalistic activity on social media. Of the $4,222,734$ Twitter accounts for which we have data, 0.12% are the accounts of journalists (see Table 2 above); while this might seem low, it is actually rather high compared to the share of the total adult population that journalists represent.

To investigate the role played by monitoring, for all the media organizations included in our sample we compute the list of their journalists present on Twitter and investigate the heterogeneity of the effects depending on this variable. Table 17 reports the results of the estimation of the Poisson regression model (online Appendix Table C.4 reports the associated naive estimates). We find that the marginal effect is higher for the media that have a high number of journalists with a Twitter account (Columns (4) and (5)) than for those with only a few numbers (Columns (1) and (2)).

However, this finding may be partly driven by the fact that some media simply have *more journalists* than others (independently of the social media presence of these journalists). As highlighted in the data section 2.2, for the 68 media outlets for which this information is available, we compute data on the size of their newsroom. In Columns (3) and (6), for this subsample of media outlets, we control for their number of journalists.[33] Doing so does not affect our findings; on the contrary, the marginal effect of the popularity on Twitter becomes even stronger for the media outlets whose journalists are relatively more present on the social network.

Hence, while it cannot entirely explain our findings, the monitoring of Twitter by journalists seems to play a role here. Consistently with this finding, we also show in the online Appendix that the magnitude of the effect is stronger for the media that are more present on social media. The intensity of the social media presence is defined with respect to the share of the media outlet's articles that are published on Twitter (Table C.5).

[32] For additional evidence of Twitter as a reporting tool, see Vis (2013).

[33] Note that when we do so, we cannot include media fixed effects given there is no variation at the media outlet level in the number of journalists.

Table 17: IV estimates: Media-level approach, IV Poisson GMM, Depending on the number of journalists with a Twitter account

	Low nb journalists on Twitter			High nb journalists on Twitter		
	(1)	(2)	(3)	(4)	(5)	(6)
Number of tweets	0.048***	0.048***	0.048***	0.039***	0.039***	0.038***
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.009)
Low pressure	0.153	0.145	0.145	0.196**	0.189**	0.191**
	(0.101)	(0.102)	(0.102)	(0.092)	(0.093)	(0.094)
Centrality	-0.042**	-0.044**	-0.044**	-0.034**	-0.035**	-0.034**
	(0.020)	(0.020)	(0.020)	(0.016)	(0.016)	(0.016)
Log # seed's followers	0.003	0.005	0.005	0.005	0.006	0.006
	(0.018)	(0.019)	(0.019)	(0.017)	(0.017)	(0.018)
Number of journalists			0.025***			0.002***
			(0.000)			(0.000)
Media FEs	✓	✓		✓	✓	
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media		✓	✓		✓	✓
Drop multiple		✓	✓		✓	✓
Observations	103,350	101,520	101,520	310,050	304,560	162,432
Clusters (events)	3,445	3,384	3,384	3,445	3,384	3,384
Marginal Effect	0.012	0.012	0.012	0.018	0.018	0.032

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a a generalized method of moments (GMM) Poisson regression model with endogenous regressors (Stata's ivpoisson command). Standard errors are clustered at the event level. An observation is a media-news event. Columns (1) and (4) report the estimates for all the events that appear first on Twitter; in Columns (2), (3), (5) and (6), we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). All specifications include day-of-the-week, month, and media fixed effects. In Columns (3) and (6), we also control for the number of journalists working for the media. In Columns (1) to (3) (respectively (4) to (6), we consider the media with a relatively low (respectively relatively high) number of journalists with a Twitter account. The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

Table 18: Naive estimates: Media-level approach, Depending on the reliance on advertising revenues

	No advertising		Advertising	
	(1)	(2)	(3)	(4)
Number of tweets	0.041***	0.042***	0.049***	0.048***
	(0.009)	(0.009)	(0.011)	(0.011)
Log # seed's followers	-0.002	-0.006	-0.018*	-0.020*
	(0.013)	(0.013)	(0.011)	(0.011)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Drop media		✓		✓
Drop multiple		✓		✓
Observations	145,596	142,989	630,916	619,619
Clusters (events)	4,412	4,333	4,412	4,333
Marginal Effect	0.017	0.017	0.013	0.013

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a negative binomial estimation. Standard errors are clustered at the event level. An observation is a media-news event. Columns (1) and (3) report the estimates for all the events that appear first on Twitter; in Columns (2) and (4) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). All specifications include day-of-the-week, month, and media fixed effects. In Columns (1) and (2) (respectively (3) and (4), we consider the media without online advertising (respectively with advertising revenues). The number of tweets is computed *before* the first news article in the event appears and is given in thousands. More details are provided in the text.

5.2 Editorial decisions and popularity

The causal relationship between the popularity of a story on Twitter and its mainstream media coverage can be due to the existence of a clicks bias. This explanation is consistent with the results of Sen and Yildirim (2015) which show, using data from a leading English-language Indian national daily newspaper, that editors' coverage decisions regarding online news stories are influenced by the observed popularity of the story, as measured by the number of clicks received.

In Sen and Yildirim (2015)'s framework (which builds on Latham (2015)), the newspaper cares about the revenue generated by covering a story, which is assumed to be proportional to the number of readers. To test for this hypothesis, we use the fact that our sample of media outlets includes a lot of different media outlets, some of which rely on advertising revenues while others do not. Table 18 presents our estimates depending on whether the media rely on advertising revenues (around 80% of the media outlets in our sample do so). The order of magnitude of the estimated effects of popularity on Twitter on media coverage is more or less similar in both cases; if anything, the marginal effects are slightly higher for the media that do not rely on advertising online.

Hence our results do not seem to be driven by short-term considerations generated by advertising revenue-bearing clicks. However, even in the absence of such a consideration, publishers may be willing to cover the stories that resonate the most. In other words, news editors may aim to produce news that consumers are interested in. However, news editors do not know consumers' preferences; hence they can use the popularity of an event on Twitter as a signal that allows them to draw inferences about consumers' preferences. We discuss below the extent to which such a signal might be biased, and the welfare implications of our findings.

6 Additional robustness checks and discussion

6.1 Robustness

We perform several robustness checks. This section briefly describes them; the detailed results for these tests are available in the online Appendix.

French media In this article, we compare the popularity of tweets in French with the coverage that French mainstream media devote to a number of events.[34] However, French is a language not only used in France, but also in parts of Belgium, Switzerland and Canada, as well as in a number of North Africa and sub-Sahara African countries. Hence, how can we be sure that our tweets in French are twitted by people living in France or consuming French media online?

On Twitter, there are two different ways to identify location: through the location of the user and through the location of the tweet. First, users can indicate their location in their profile ("user-defined location"): nearly two third of the users in our sample do so. While this location is most often a real location (e.g. "Paris, France" or "Val-d'Oise, Ile-de-France"), this is not necessarily the case (e.g. some users indicate "Gotham City" or "Everywhere and nowhere"). We parse this field and, using OpenStreetMap, we obtain coordinates (latitude and longitude) that we attribute to a country. Out of the $2,693,307$ users for which the location field is filled, the information provided allows us to recover the country where the user is located in 72% of the cases. 47% of these users indicate that they are located in France.

Second, users can share their location with Twitter at the time when they tweet. On the one hand, when they decide to assign a location ("place") to their tweet, they are presented with a list of candidate Twitter places. However, out of our $4,222,734$ unique users, only $62,037$ do so for the first of their tweets that we observe. On the other hand, the "place" field is always present when a Tweet is "geo-tagged". However, even fewer users provide their exact location: only $13,382$ (i.e. 0.32%) do so the first time we observe them, and $13,529$ the last time.

[34] We discuss below the issue of the external validity of our results.

While the issue of the location of the users may seem an important one, it is essential to note that this information is also not available to the media; while they may consider the popularity of an event of Twitter at the time of choosing their coverage, and while they can observe whether the users tweet in French or in a different language, journalists cannot (no more than researchers) know where the users tweeting are located. Further, if people do read French – and we can assume here that they do given that they tweet in French –, they can consume French mainstream media even if they are located outside of France. Hence, if what media outlets care about is the number of clicks, they should ignore the location of the users (even if such information may matter for advertisers, depending on where they are located).

With this in mind and as a first robustness check, we re-run our main analysis but only for the media outlets that are located in France (i.e. we drop the content produced by the French-language not-located-in-France media outlets in our sample). Online Appendix Tables E.1, E.2 and E.3 present the results. Doing so does not affect our main results; if anything, the magnitude of the estimated effects is slightly higher.

Controls In our preferred specification, we only use a reduced set of controls at the event level, including the seed of the event's number of followers. In the online Appendix, we show that our results are robust to adding controls; Table E.4 presents the event-level results for the naive estimates and Table E.5 for the Poisson model with endogenous regressors. In both tables, we report in the first columns our preferred specification for the sake of comparison. In Column (2), we control for the time of the day of the first tweet (using an indicator variable in hourly format). Doing so does not affect our main results.

In Column (3), we also control for the reaction time of the first media (i.e. the time interval between the first tweet in the event and the first news article). We do not do so in our preferred specification to avoid a bad controls problem. However, it is interesting to see that even when doing so, we find a positive and statistically significant relationship between the popularity on Twitter and the media coverage. The magnitude of the estimates coefficient is halved however in the naive approach, and also decreases, but to a lower extent, when we estimate the IV Poisson model.

Besides, in Columns (4) and (5), we show that the magnitude of our IV estimates is not affected by the introduction of additional user-level controls: indeed, we find no change in the magnitude of the estimated marginal effects when we introduce in our control set the number of tweets the seed of the event has liked, the number of Twitter accounts she is following, the number of public lists and her total number of tweets (all these characteristics are computed the first time we observe the user in our data).

Sample Next, we show that our results are robust to changes in our sample of analysis. First, we show that our results are robust to dropping the small events in our sample, i.e. below the 10th percentile in terms of the total number of tweets in the event (online Appendix Tables E.6 and E.7). Second, a number of media outlets in our sample produce on average much less articles than others. We verify that our results are robust to dropping the media outlets who produce on average less than one article a day during our period of interest. We find that doing so does not affect our main results, either qualitatively or quantitatively (online Appendix Table E.8).

Measures of centrality Finally, we show that our results are robust to considering alternative measures of users' centrality. A number of different measures of the centrality of a node in a graph have indeed been proposed in the literature: one of the most well-known is Bonacich centrality (Bonacich, 1987), that relies on the number of walks (existing paths through the graph) emanating from that node. Another metric is the so-called "diffusion centrality", proposed by Banerjee et al. (2012) who models centrality as a probability of contacting other agents in the graph based on the degree of the nodes. To test the robustness of our model to the use of these different measures, given we do not have access to the complete followers-followees graph, we apply these metrics on the graph of interactions (who retweeted/quoted whom).

6.2 External validity of our results

The results presented in this paper are based on French data and the use of Twitter. Hence, one final question is whether we should expect the patterns we have uncovered in the case of France to be repeated in other contexts and the influence of Twitter to be similar to the one of other social media.

First, it is important to highlight that the choice of France was driven by data considerations. As previously described, the fact of relying on tweets in French language allows us to recover around 70% of all the tweets in French during our period of consideration. This wouldn't have been possible with data from the United States for example, relying on English language. Indeed, given that there is much more activity on Twitter in English than in French, the 1% limitation of the Twitter API makes it impossible to recover a very large (and representative) share of the activity in English.

Should the patterns we obtain with the French data hold in other countries? There are good reasons to think this could be the case. First, while the French media market certainly presents specific features, it is by and large very similar to other Western media markets, whether we consider Internet penetration (87%, like Italy and Spain and only slightly below Belgium – 88% – and Germany – 90%), the use of social media for news (36%, compared to 31% for Germany and 39% for the U.K.), or the proportion of the population who paid for online news (11%, like in Spain, slightly above Germany or Canada – 8% – but below Italy – 12%) (Reuters Institute, 2018). In France, like in other Western media markets, many publishers offer online news for free and largely rely on advertising. Moreover France, like the U.S., has an international news agency, the AFP, which is the third leading agency in the world after Reuters and Associated Press. From this point of view, the French market is more similar to the U.S. market than the Spanish, Italian, or German markets. Therefore, overall, we believe that the results presented in this paper have implications for other Western countries.

A second concern may come from the fact that we are using Twitter data, while more citizens use Facebook than Twitter. Beforehand, it is important to highlight that Facebook data to a scale similar to what we are using with Twitter is not available to researchers. Besides and more importantly, Twitter is largely used – in particular by journalists – for news. In fact, Twitter is more commonly used as a news source than Facebook (von Nordheim et al., 2018). Hence, given the focus of this article is on the impact of social media on news production and consequently consumption, it makes more sense to consider Twitter than other social media.

6.3 Welfare implications

What are the welfare effects of social media? In particular, would citizens be better informed absent Twitter? In this article, we have investigated the extent to which traditional media outlets react to stories popularity on Twitter. Given the instrumental variable strategy we use in the paper to isolate the causal impact of a story popularity on Twitter, our estimates capture the effects of a variation in popularity that is uncorrelated with a story underlying newsworthiness. In other words, our findings suggest that social media may provide a biased signal of what readers want.

Twitter representativeness Further, even absent such a bias, it is important to highlight that Twitter users are not representative of the general news-reading population. In Table 19, using data from the Reuters Institute's 2018 Digital News Report (Reuters Institute, 2018) for France, we compute the average characteristics of the news-consuming surveyed individuals depending on whether they use Twitter.[35] We see that Twitter users are younger on average, more educated, much more interested in news, and more often on the Left of the political spectrum than non-users, and that there are relatively more women among them. Further,

[35]The survey data includes 2,006 individuals for France in 2018. Out of them, 88.8% report that they have used either "television news bulletins or programmes", "24 hour news television channels", "radio news programmes or bulletins", "printed newspapers", "printed magazines", "websites/apps of newspapers", "websites/apps of news magazines", "websites/apps of TV and radio companies", "websites/apps of other news outlets" as a source of news in the last week. We focus on these users in the Table.

Table 19: Unrepresentativeness of the Twitter users: Characteristics of the news-consuming citizens depending on whether they use Twitter

	Does not use Twitter	Uses Twitter	Diff/se
Age	51	44	7.4***
			(1.0)
=1 if male (%)	45	59	-13.1***
			(3.2)
=1 if Bachelor degree or more (%)	32	43	-11.1***
			(3.1)
=1 if Annual income $\geq$ €30 000 (%)	45	47	-1.4
			(3.3)
=1 if leaves in Paris region (%)	18	23	-5.8*
			(2.5)
=1 if highly interested in news (%)	49	71	-22.2***
			(3.1)
=1 if places itself on the Left (%)	32	43	-11.0***
			(3.0)
Observations	1,785		

Notes: The table gives summary statistics. Year is 2018. The observations are at the individual level. Data are from the *2018 Digital News Report* (Reuters Institute, 2018). The sample includes 2,006 surveyed individuals for France, out of which 1,785 claim to consume news. Column (1) presents the results for the individuals who do not use Twitter. Column (2) presents the results for the individuals who use Twitter. In Column (3) we perform a t-test on the equality of means (robust standard errors are in parentheses).

the difference is even more striking if rather than considering the citizens who use Twitter for any purpose (16.18% of the surveyed individuals), we only consider those who share news on Twitter (8.54%; see online Appendix Table C.6).

Hence journalists – because they tend to be on Twitter – rely on this social media to obtain information on consumers' preferences, but by doing so they rely on a signal generated by citizens who are not representative of the overall news-consuming population. What's more, the variations we identify in the paper are driven by users who are highly-central in the network and so probably even less representative of the overall population.

While a growing literature documents the fact that journalists are now exposed to copious quantitative data about the preferences of online readers thanks to the adoption of web analytics software programs (Anderson, 2011; Christin, 2020; Christin and Petre, 2020), our findings highlight the bias that may come from journalists reliance on Twitter when they decide on whether to cover an event. Does such a bias affect news quality or citizen satisfaction with the news they consume?

How do consumers react? We know from survey data that a significant share of the population does not find the news produced by the media of interest. In 2018 in France, only 62% of the individual surveyed said they were very or somewhat interested in the news produced by the media (TNS et al., 2019). Can the behaviors documented in this paper explain this lack of interest?

To answer this question, we investigate whether the popularity of a news event on Twitter is associated with an higher demand for the news articles published by the media on this event. Unfortunately, article-level audience data are not available to the researchers. Hence, to compute such a measure, we proceed as follows: for each website, we compute daily-level information on the number of unique visitors and, for each article, we compute the number of times it has been shared on Facebook. Following Cagé et al. (2020), we combine these statistics to proxy the number of views at the article level.[36] Table 4 provides summary statistics on the average number of shares, comments, and reactions on Facebook received by the articles in our sample.

We study whether the news articles covering events that are more popular on Twitter get relatively more views. Our empirical specification is similar to the one presented in equation (1), but the dependent variable is now the average number of times the articles published in the event are shared on Facebook. Table 20 presents the results: we show that on average an increase by 1% in the number of tweets is associated with a 2 to 3% increase in the number of times the articles publish by the media outlets in the event are shared on Facebook.[37] Interestingly, the magnitude of this effect is lower than the magnitude of the observed increase in the news coverage media outlets devote to the event.

Publishers' incentives to invest in quality Finally, as highlighted in the introduction, social media may affect publishers' incentives to invest in quality. de Cornière and Sarvary (2019) show for example theoretically that, following the introduction of platforms, high-quality newspapers invest more in quality but low-quality ones invest less. We are now working on investigating empirically whether the introduction of Twitter led to opposite investment decisions by media outlets depending on the initial size of their newsroom. The overall welfare effects of social media indeed depends on both the media outlets' production decisions and editorial choices and on consumers' reaction.

7 Conclusion

Social media is a complex phenomenon that has both positive and negative effects on the welfare of people (Allcott et al., 2020). This also holds true for its impact on traditional

[36] A number of articles in the literature assume that exposure is proportional to Facebook shares (see e.g. Allcott and Gentzkow, 2017). Cagé et al. (2020) show that the relationship between the number of views of an article and the number of shares on Facebook is almost perfectly linear.

[37] The number of observations is lower in Table 20 than in the previous tables; this is due to the fact that we are still in the process of computing the number of shares on Facebook for all the articles in our dataset.

Table 20: Popularity on Twitter and demand for news

	Average number of shares on Facebook					
	(1)	(2)	(3)	(4)	(5)	(6)
Log # tweets	2.90***	3.03***	2.05**	3.07***	3.04***	2.96***
	(1.06)	(1.08)	(1.04)	(1.09)	(1.08)	(1.14)
# articles		-0.02		-0.02	-0.02	-0.01
		(0.01)		(0.01)	(0.01)	(0.01)
# media outlets			0.77***			
			(0.23)			
Log # seed's followers	-0.11	-0.11	0.05			
	(0.74)	(0.74)	(0.75)			
# seed's followers				-0.67	-0.99	-2.50
				(1.97)	(2.14)	(4.48)
# seed's followers-squared				-0.02	-0.01	0.06
				(0.07)	(0.08)	(0.22)
=1 if first tweet during night					17.52	18.58
					(16.98)	(18.07)
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Drop media						✓
Drop multiple						✓
Observations	1,552	1,552	1,552	1,552	1,552	1,464

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is July 2018 - September 2018. Models are estimated using a linear-log model (robust standard errors are reported between parentheses). An observation is a news event. We only consider the subset of news events that appear first on Twitter. All specifications include the seed's number of followers as a control, and day-of-the-week and month fixed effects. Columns (1) to (3) and (5) to (7) report the estimates for all the events that appear first on Twitter; in Columns (4) and (8) we drop the events whose seed is the Twitter account of a media outlet or journalist ("media") as well as the events whose seed broke more than one event during our time period ("multiple"). The number of tweets is computed *before* the first news article in the event and is given in thousands. More details are provided in the text.

media. According to Cision (2019)'s Global State of the Media Report, the bypassing of traditional media by social media is considered as the biggest challenges facing journalism for the future. At the same time, journalists increasingly rely on social media to stay connected to sources and real-time news.

In this paper, we focus on an important dimension that has been overlooked in the discussions on the implications of the changes brought by social media: namely how it affects publishers' production and editorial decisions. To do so, we built a new dataset encompassing nearly 70% of all the tweets in French over a long time period and the content produced by the general information media outlets during the same time period. We develop new algorithms that allow us to study the propagation of news stories between social and mainstream media. Focusing on the stories that emerge first on Twitter, we show that their popularity on Twitter affects the coverage mainstream media devote to these stories.

These findings shed a new light on our understanding of how editors decide on the coverage for stories, and have to be taken into account when discussing policy implications of the recent changes in media technologies. In particular, while social media compete with mainstream media for audience attention, they can also be used as a signal to draw inferences about consumers' preferences. In future research, we will investigate whether this happens at the expense of the quality of the information produced.

References

Alaoui, Larbi and Fabrizio Germano, "Time scarcity and the market for news," *Journal of Economic Behavior & Organization*, 2020, *174*, 173–195.

Allcott, Hunt and Matthew Gentzkow, "Social Media and Fake News in the 2016 Election," *Journal of Economic Perspectives*, 2017, *31* (2), 211–236.

__ , Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow, "The Welfare Effects of Social Media," *American Economic Review*, 2020, *110* (3), 629–676.

__ , Matthew Gentzkow, and Chuan Yu, "Trends in the Diffusion of Misinformation on Social Media," *Research & Politics*, 2019.

Anderson, C W, "Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms," *Journalism*, 2011, *12* (5), 550–566.

Anderson, Michael L and Jeremy Magruder, "Split-Sample Strategies for Avoiding False Discoveries," Working Paper 23544, National Bureau of Economic Research jun 2017.

Angelucci, Charles and Julia Cagé, "Newspapers in Times of Low Advertising Revenues," *American Economic Journal: Microeconomics*, 2019, *11* (3), 319–364.

__ , __ , and Michael Sinkinson, "Media Competition and News Diets," Working Paper 26782, National Bureau of Economic Research 2020.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson, "The Diffusion of Microfinance," NBER Working Papers 17743, National Bureau of Economic Research, Inc 2012.

Barberá, Pablo, "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data," *Political Analysis*, 2015, *23* (1), 76–91.

__ , Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker, "Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data," *American Political Science Review*, 2019, *113* (4), 883–901.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, oct 2008, *2008* (10), P10008.

Bonacich, Phillip, "Power and Centrality: A Family of Measures," *American Journal of Sociology*, 1987, *92* (5), 1170–1182.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro, "A note on internet use and the 2016 U.S. presidential election outcome," *PLOS ONE*, 2018, *13* (7), 1–7.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Peer Effects in Networks: A Survey," IZA Discussion Papers 12947, Institute of Labor Economics (IZA) 2020.

Cagé, Julia, "Payroll and Inequality within the Newsroom: Evidence from France, 1936-2016," Working Paper, SciencesPo Paris 2016.

__ , "Media Competition, Information Provision and Political Participation: Evidence from French Local Newspapers and Elections, 1944-2014," *Journal of Public Economics*, 2020, *185.*

__ , Nicolas Hervé, and Marie-Luce Viaud, "The Production of Information in an Online World," *The Review of Economic Studies*, 2020, *87* (5), 2126–2164.

Cardon, Dominique, Jean-Philippe Cointet, Benjamin Ooghe-Tabanou, and Guillaume Plique, "Unfolding the Multi-layered Structure of the French Mediascape," Sciences Po médialabal. Working paper 2019-1 1 2019.

Christin, Angèle, *Metrics at Work: Journalism and the Contested Meaning of Algorithms*, Princeton University Press, 2020.

— **and Caitlin Petre**, "Making Peace with Metrics: Relational Work in Online News Production," *Sociologica*, 2020, *14* (2), 133–156.

Cision, "Cision's 2019 Global State of the Media Report," report 2019.

Claussen, Jörg, Christian Peukert, and Ananya Sen, "The Editor vs. the Algorithm: Targeting, Data and Externalities in Online News," Working Paper 2019.

de Bruijn, Jens, Hans de Moel, Brenden Jongman, and Jeroen Aerts, "Towards a global flood detection system using social media," in "EGU General Assembly Conference Abstracts," Vol. 19 2017, p. 1102.

de Cornière, Alexandre and Miklos Sarvary, "Social Media and News: Attention Capture via Content Bundling," Working Paper 2019.

Djourelova, Milena and Ruben Durante, "Media Attention and Strategic Timing in Politics: Evidence from U.S. Presidential Executive Orders," CEPR Discussion Papers 13961, C.E.P.R. Discussion Papers 2019.

Eisensee, Thomas and David Strömberg, "News Droughts, News Floods, and U. S. Disaster Relief," *The Quarterly Journal of Economics*, 2007, *122* (2), 693–728.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova, "Social Media and Protest Participation: Evidence from Russia," *Econometrica*, 2020, *Forthcomin.*

— **, Maria Petrova, and Konstantin Sonin**, "Social Media and Corruption," *American Economic Journal: Applied Economics*, 2018, *10* (1), 150–174.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, and Others, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in "KDD," Vol. 96 1996, pp. 226–231.

Fafchamps, Marcel and Julien Labonne, "Using Split Samples to Improve Inference about Causal Effects," Working Paper 21842, National Bureau of Economic Research jan 2016.

— **and** — , "Do Politicians' Relatives Get Better Jobs? Evidence from Municipal Elections," *The Journal of Law, Economics, and Organization*, 2017, *33* (2), 268–300.

Gavazza, Alessandro, Mattia Nardotto, and Tommaso Valletti, "Internet and Politics: Evidence from U.K. Local Elections and Local Government Policies," *The Review of Economic Studies*, 2019, *86* (5), 2092–2135.

Gentzkow, Matthew, "Television and Voter Turnout," *Quarterly Journal of Economics*, 2006, *121* (3), 931–972.

— **, Jesse M Shapiro, and Michael Sinkinson**, "The Effect of Newspaper Entry and Exit on Electoral Politics," *American Economic Review*, 2011, *101* (7), 2980–3018.

Giavazzi, Francesco, Felix Iglhaut, Giacomo Lemoli, and Gaia Rubera, "Terrorist Attacks, Cultural Incidents and the Vote for Radical Parties: Analyzing Text from Twitter," Working Paper 26825, National Bureau of Economic Research mar 2020.

Glaeser, Edward L, "Researcher Incentives and Empirical Methods," Working Paper 329, National

Bureau of Economic Research oct 2006.

Gorodnichenko, Yuriy, Tho Pham, and Oleksandr Talavera, "Social Media, Sentiment and Public Opinions: Evidence from #Brexit and #USElection," Working Paper 24631, National Bureau of Economic Research 2018.

Halberstam, Yosh and Brian Knight, "Homophily, group size, and the diffusion of political information in social networks: Evidence from Twitter," *Journal of Public Economics*, 2016, *143*, 73–88.

Hatte, Sophie, Etienne Madinier, and Ekaterina Zhuravskaya, "Reading Twitter in the Newsroom: How Social Media Affects Traditional-Media Reporting?," Mimeo 2020.

Jeon, Doh-Shin, "Economics of News Aggregators," Toulouse School of Economics Working Papers 18-912 2018.

_ and Nikrooz Nasr, "News Aggregators and Competition among Newspapers on the Internet," *American Economic Journal: Microeconomics*, 2016, *8* (4), 91–114.

Kergl, Dennis, Robert Roedler, and Sebastian Seeber, "On the endogenesis of Twitter's Spritzer and Gardenhose sample streams," in "2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2014, Beijing, China, August 17-20, 2014" 2014, pp. 357–364.

Latham, Oliver, "Lame Ducks and the Media," *The Economic Journal*, 2015, *125* (589), 1918–1951.

Leamer, Edward E, *Specification Searches: Ad Hoc Inference with Nonexperimental Data* A Wiley-Interscience publication, Wiley, 1978.

_ , "Let's Take the Con Out of Econometrics," *The American Economic Review*, 1983, *73* (1), 31–43.

List, John A, Azeem M Shaikh, and Yang Xu, "Multiple hypothesis testing in experimental economics," *Experimental Economics*, dec 2019, *22* (4), 773–793.

Longhi, Julien, Marinica Claudia, and Després Zakarya, "Political language patterns' dissemination between a political leader and his campaign community: a CMC corpora analysis," *European Journal of Applied Linguistics*, 2019, *0.*

Mazoyer, Béatrice, Julia Cagé, Céline Hudelot, and Marie-Luce Viaud, "Real-time collection of reliable and representative tweets datasets related to news events," in "CEUR Workshop Proceedings," Vol. 2078 2018.

_ , _ , Nicolas Hervé, and Céline Hudelot, "A French Corpus for Event Detection on Twitter," *International Conference on Language Resources and Evaluation*, 2020.

_ , Nicolas Hervé, Céline Hudelot, and Julia Cagé, "Représentations lexicales pour la détection non supervisée d'événements dans un flux de tweets : étude sur des corpus français et anglais," 2020.

McGregor, Shannon C and Logan Molyneux, "Twitter's influence on news judgment: An experiment among journalists," *Journalism*, 2018.

McMinn, Andrew James, Yashar Moshfeghi, and Joemon M Jose, "Building a large-scale corpus for evaluating event detection on twitter," in "22nd {ACM} International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013" 2013, pp. 409–418.

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu, "When is it biased?: assessing the representativeness of twitter's streaming API," in "23rd International World Wide Web Conference, WWW'14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume" ACM Press 2014, pp. 555–556.

Müller, Karsten and Carlo Schwarz, "From Hashtag to Hate Crime: Twitter and Anti-Minority Sentiment," Working Paper 2019.

Nichols, Austin, "IVPOIS: Stata module to estimate an instrumental variables Poisson regression via GMM," Statistical Software Components, Boston College Department of Economics nov 2007.

Pagolu, Venkata Sasank, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi, "Sentiment analysis of Twitter data for predicting stock market movements," in "2016 international conference on signal processing, communication, power and embedded system (SCOPES)" IEEE 2016, pp. 1345–1350.

Petrova, Maria, Ananya Sen, and Pinar Yildirim, "Social Media and Political Donations: New Technology and Incumbency Advantage in the United States," CEPR Discussion Papers 11808, C.E.P.R. Discussion Papers 2017.

Pew Research Center, "Social Media Fact Sheet," Technical Report, techreport 2018.

_ , "Mobile Technololy and Home Broadband 2019," Report 2019.

Reuters Institute, "Digital News Report 2018," Annual Report 2018.

Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo, "Earthquake Shakes Twitter Users: Real-Time Event Detection by Social Sensors," in "Proceedings of the 19th International Conference on World Wide Web" WWW '10 Association for Computing Machinery New York, NY, USA 2010, pp. 851–860.

Sen, Ananya and Pinar Yildirim, "Clicks and Editorial Decisions: How Does Popularity Shape Online News Coverage?," Working Paper 2015.

Snyder, James M and David Stromberg, "Press Coverage and Political Accountability," *Journal of Political Economy*, 2010, *118* (2), 355–408.

Strömberg, David, "Radio's Impact on Public Spending," *The Quarterly Journal of Economics*, feb 2004, *119* (1), 189–221.

TNS, La Croix, and Kantar Media France, "La confiance des Français dans les media," Annual survey January 2019.

Vis, Farida, "Twitter as a Reporting Tool for Breaking News," *Digital Journalism*, 2013, *1* (1), 27–47.

von Nordheim, Gerret, Karin Boczek, and Lars Koppers, "Sourcing the Sources," *Digital Journalism*, 2018, *6* (7), 807–828.

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

_ , Mostafa 'Neo' Mohsenvand, and Deb Roy, "Rumor Gauge: Predicting the Veracity of Rumors on Twitter," *ACM Trans. Knowl. Discov. Data*, 2017, *11* (4), 50:1—-50:36.

Wihbey, John, Kenneth Joseph, and David Lazer, "The social silos of journalism? Twitter, news media and partisan segregation," *New Media & Society*, 2019, *21* (4), 815–835.

Windmeijer, F A G and J M C Santos Silva, "Endogeneity in Count Data Models: An Application to Demand for Health Care," *Journal of Applied Econometrics*, 1997, *12* (3), 281–294.

Wooldridge, Jeffrey M, *Econometric Analysis of Cross Section and Panel Data* Econometric Analysis of Cross Section and Panel Data, MIT Press, 2002.

_ , *Introductory Econometrics: A Modern Approach*, Cengage Learning, 2013.

Yang, Yiming, Thomas Pierce, and Jaime G Carbonell, "A Study of Retrospective and On-Line Event Detection," in "Proc. of ACM-SIGIR" 1998, pp. 28–36.

Yin, Jianhua and Jianyong Wang, “A dirichlet multinomial mixture model-based approach for short text clustering,” in “in” ACM Press 2014, pp. 233–242.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov, “Political Effects of the Internet and Social Media,” *Annual Review of Economics*, 2020, *Forthcomin.*