Social Media Influence Mainstream Media: Evidence from Two Billion Tweets*

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

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

First version: June 2019. This version: June 28th 2022

Abstract

Social media are increasingly influencing society and politics, despite the fact that legacy media remain the most consumed source of news. In this paper, we study the propagation of information from social media to mainstream media, and investigate whether news editors' editorial decisions are influenced by the popularity of news stories on social media. To do so, we build a novel dataset including around 70% of all the tweets produced in French between August 2018 and July 2019 and the content published online by 200 mainstream media outlets. We then 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 "social media news pressure" at the time of the event. We show that the social media popularity of a story increases the coverage of the same story by mainstream media. This effect varies depending on the media outlets' characteristics, in particular on whether they use a paywall. Finally, we investigate consumers' reaction to a surge in social media popularity. Our findings shed new light on news production decisions in the digital age and the welfare effects of social media.

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

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

*An earlier version of this paper circulated under the title "Social Media and Newsroom Production Decisions." We gratefully acknowledge the many helpful comments and suggestions from Dominique Cardon, Edgard Dewitte, Mirko Draca, Michele Fioretti, Matthew Gentzkow, Malka Guillot, Emeric Henry, Céline Hudelot, Brian Knight, Gregory Martin, Benjamin Marx, Elisa Mougin, Maria Petrova, Zeynep Pehlivan, Andrea Prat, Jesse Shapiro, Clémence Tricaud, Philine Widmer, and Katia Zhuravskaya, as well as Denis Teyssou for his help with the AFP data, the *Alliance pour les chiffres de la presse et des médias* for providing us audience data, and Guillaume Plique for his advice on algorithms. We are grateful to seminar participants at Brown University, the CEPR Virtual IO Seminar, the Columbia Institute for Tele-Information, the DC Political Economy Webinar, ETH Zurich, Harvard University, Jonkoping International Business School, the London School of Economics, the Medialab, the Oz Virtual Econ Research Seminar, Sciences Po Paris, the Tinbergen Institute, and the Virtual Digital Economy Seminar, and to conference participants at the Digital Economics Research Network Workshop, the 14th Digital Economics Conference, the SIOE, the RIDGE Forum LACEA-PEG Political Economy workshop, and the NBER Political Economy Meeting. Martin Bretschneider, Michel Gutmann and Auguste Naim provided outstanding research assistance. We gratefully acknowledge financial help from the NET Institute (*www.netinst.org*) and from the French National Research Agency (ANR-17-CE26-0004). Responsibility for the results presented lies entirely with the authors.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

"It's worth keeping in mind that just because a story is generating interest on social media, or a handful of people have tweeted about it, that does not necessarily mean it has news value and needs to be reported or circulated further on social media."

The Guardian's social media guidelines (released in May 2022).

1 Introduction

Many recent papers have shown that social media have changed society and matter for democracy (Fujiwara et al., 2021; Levy, 2021); yet, television remains by far the most popular source of news.[1] So how can we explain the outsized influence of social media? In this paper, relying on nearly two billion tweets and an innovative empirical approach, we provide new evidence in favor of the long-standing hypothesis that social media affect publishers' production and editorial decisions.

Besides, we add a new set of perspectives to the literature studying the welfare effects of social media (Allcott et al., 2020). While there are widespread fears that new technologies are worsening editorial quality, we investigate whether the contagion from social to mainstream media varies depending on the outlets' characteristics, and in particular on whether they offer digital news for free. The influence of social media may indeed increase information inequality (Kennedy and Prat, 2019), which would in turn affect voting outcomes.

To do so, we built a novel dataset including around 70% of all the tweets in French emitted during an entire year (August 1st 2018-July 31st 2019) and the content produced online by French-speaking general information media outlets during the same time period. 200 media outlets are included, encompassing newspapers, television channels, radio stations, pure online media, and news agencies' dispatches. Our dataset contains around 1.8 billion tweets as well as 4 million news articles. Producing these data is our first contribution. To the best of our knowledge, it is the most exhaustive dataset on social and mainstream media available to researchers, and the algorithms we develop to analyze these data could be of future use to other researchers studying online information.

We investigate whether news editors' editorial decisions are influenced by the popularity of news stories on Twitter. The major empirical challenge lies in disentangling the causal effect of the popularity of a story on social media; we leverage the large size of our dataset to propose a novel instrument that relies on the interaction between the centrality of the users and the news pressure on the social media at the time of the event.

More precisely, for each tweet as well as each news article, we determine their precise "publication" time. We merge this data with information we collect on the Twitter users,

[1] In 2018, 49% of US adults said they often get news on television, compared to 20% on social media (https://www.pewresearch.org/fact-tank/2019/09/11/key-findings-about-the-online-news-landscape-in-america/). In the UK, a Reuters Institute study shows that voters mainly got their news from BBC News during the 2019 general election (Fletcher et al., 2020). Similarly in France, according to the 2022 Viavoice survey, television was used by 75% of the citizens to get information on the Presidential campaign (compared to 33% for the Internet and 20% for the social media accounts of the mainstream media).

Electronic copy available at: https://ssrn.com/abstract=3663899

in particular their number of followers and the number of interactions (i.e. the number of retweets, favorites and replies) generated by their tweets. We develop a "community detection" algorithm to identify the joint news events that happen both on Twitter and on mainstream media. This allows us to study the propagation of information between Twitter and traditional media. We determine 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 and, for the subset of events that originate first on social media, we study whether their popularity on Twitter impacts the coverage that mainstream media devote to them.

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).[2] Following Fafchamps and Labonne (2016, 2017) and Anderson and Magruder (2017), we first perform our analysis on the August 2018-November 2018 time period (four months of data). The results of this draft rely on this sub-sample, which 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 (December 2018-July 2019).

The sample we use in this version of the article includes 531 million tweets and $1,239,552$ news articles. We identify $3,992$ joint events, i.e. events that are covered both on Twitter and on traditional media. Focusing on the subset of news stories that originate first on Twitter ($3,904$ events), we investigate how their popularity affects the coverage that traditional media devote to them. 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.[3]

The main empirical challenge 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. We propose a novel instrument that relies on the interaction between users' centrality in the Twitter network and the "news pressure" on Twitter at the time of the event.[4] We consider the interaction between centrality and pressure rather than centrality and/or pressure taken independently, given that both centrality and social media news pressure likely directly correlate with the news editors' decision (in other words, they both assumably violate the exclusion restriction). On the contrary, the interaction between centrality and news pressure is a valid instrument under an arguably effective identification assumption: once we control for the direct effects of centrality and news pressure, the interaction between users' centrality and news pressure should only affect traditional news production through its effect on the tweet's visibility on

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

[3] In the robustness Section 6.1, we show that our results are unchanged if we instead consider only the number of original tweets. We call "original tweets" the tweets that are not retweets.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

Twitter, i.e. its number of "impressions".[5] (We also control for day-of-the-week fixed effects, calendar-month fixed effects, as well as an indicator variable equal to one if the first tweet in the event is tweeted during the night.) While we cannot rigorously test this assumption, we show that our instrument is not correlated with a number of observable event characteristics, such as its topic (economy, sport, environment, etc.) or the number of named entities (e.g. the mentions of places such as Paris or of individuals such as Boris Johnson).

To measure the centrality of a user on Twitter, we use the PageRank algorithm (Page et al., 1999). This algorithm relies on the idea that the importance of a node is related to the importance of the nodes linking to it, in a recursive approach. Hence, for a given user's importance (that we can directly proxy by her number of followers), a user will be more "central" on Twitter if her tweets tend to be retweeted by users whose tweets are also more retweeted on average. This implies in particular that her tweets – independently of their intrinsic interest – will have a higher probability of generating interactions.

To measure news pressure on Twitter, we first compute the number of interactions generated by all the tweets published in the hour preceding the first tweet in the event. To ensure the validity of the exclusion restriction, we then consider an alternative measure of pressure, where we isolate a non-news dimension: we estimate the number of interactions generated by all the tweets *except the tweets generated by the Twitter accounts of journalists and media outlets*.[6] Doing so does not affect our main findings.

We show that a fifty-percent increase in the number of tweets published before the first media article appears is correlated with an increase of 1.35 in the number of media articles published in the event, corresponding to 3.1% of the mean; this increase is partly driven by a higher number of media outlets covering the event. This result is robust to controlling for the endogeneity of the event popularity on Twitter. First, we provide evidence of the relevance of our instrument. We show that the interaction between centrality and social media news pressure is a strong and statistically significant predictor of the popularity of a Twitter event (first stage). Reduced-form evidence also shows that this interaction is correlated with the subsequent coverage of the event by traditional media. Besides, while we cannot rigorously test the exclusion restriction, we show that we cannot predict the event characteristics with our instrument. Ultimately, our instrumental-variable analysis yields results consistent with the OLS approach: we show that, at the media level, a fifty-percent increase in the instrumented number of tweets (before the first media article) leads to an increase in the number of articles corresponding to 17% of the mean. In other words, Twitter activity seems to affect publishers' editorial decisions in a quantitatively important way. Reassuringly, our IV results are robust to controlling for additional characteristics of the seed of the event[7], and doing so does not affect the magnitude of the estimates.

To understand the mechanisms behind our findings, we then investigate the heterogeneity

[5] The number of impressions is a total tally of all the times the tweet has been seen. Unfortunately, this statistic is not directly available to researchers.

[6] So that we are sure that it affects Twitter activity but not traditional media activity.

[7] We call "seed of the event" the first Twitterer who tweets in an event.

Electronic copy available at: https://ssrn.com/abstract=3663899

of our results depending on the characteristics of the media outlets. 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. This result is consistent with the fact that journalists monitor Twitter – according to Muck Rack's "State of Journalism 2019" report, nearly 60% of reporters turn to digital newspapers or magazines as their first source of news and 22% check Twitter first.

Next, 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. We show that the magnitude of the effect is larger for the media outlets with no paywall or a soft paywall than for the media with a metered or a hard paywall, pointing to the existence of a clicks bias (Sen and Yildirim, 2015). These findings also imply that the influence of social media mostly affects people who cannot afford or are unwilling to pay for news, i.e. people who are less informed to begin with (Angelucci and Prat, 2021). In other words, the contagion from social to mainstream media may increase existing information inequality.

Finally, we discuss additional 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. Furthermore, Twitter users are not representative of the general news-reading population. This points to a negative effect of social media driven by the production side, consistent with recent changes in both *The Guardian* and *The New York Times* social media guidelines, which highlight the fact that journalists tend to rely too much on Twitter as both a reporting and feedback tool, and that it may distort their view of who their audience is. Using survey data, we show nonetheless that the magnitude of our effects is higher for the media outlets whose consumers' use of Twitter is relatively high.

Turning to the demand for news, we examine whether 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. Unfortunately, we do not have audience data at the article level, but we follow Allcott and Gentzkow (2017) and Cagé et al. (2020) and assume that audience is proportional to Facebook shares. Once we control for the endogeneity of the popularity of the events on Twitter, we find no relationship between popularity and the audience of the articles then published by the media in the event. Hence, media outlets seem to overreact to what is happening on Twitter, and this might lead journalists to distort the information they produce compared to what consumers want.

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 whose

Electronic copy available at: https://ssrn.com/abstract=3663899

Twitter account is verified, 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 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 (Levy, 2021; Braghieri et al., 2022)[8], and little is known about the empirical impact social media have on news production by mainstream media. One exception is Hatte et al. (2021) 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).[9] 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

[8] On the impact of social media on online news consumption, see also Aral and Zhao (2019). Di Tella et al. (2021) study the role of Twitter in fostering political polarization.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

dissemination (Longhi et al., 2019).[10] Gorodnichenko et al. (2018) study information diffusion on Twitter, and Allcott et al. (2019) the spread of false content.[11] While this literature mostly 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 was first 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.[12] Compared to this previous work, our contribution is threefold. First, we use the entire universe of French general information media online (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. Third, we consider both the extensive and the intensive margin[13], 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

[10] 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.

[11] See also Ershov and Morales (2021) for a study of the efficiency of different interventions to reduce the sharing of certain content on social media.

[12] 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.

[13] 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.

Electronic copy available at: https://ssrn.com/abstract=3663899

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 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 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 novel 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 in Section 6 the welfare implications of our findings. 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 the joint events on social and traditional media.

2.1 Data: Tweets

First, we collect a representative sample of all the tweets in French during an entire year: August 1st 2018 - July 31st 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.[14] 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. Readers uninterested in these technical details can jump directly to Section 2.1.2 below.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

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.[15] Indeed, even if Twitter is known for providing broader access to its data than other social media platforms (in particular Facebook), the Twitter streaming APIs are strictly limited in terms 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).[16] 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 and geographical area. One can also combine the said parameters with a filter on 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 as well as their representativity of 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 (we use three different API keys[17]). 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 frequency. From this subset, we build a word co-occurrence matrix and 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 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 (see online Appendix Section A.1 for details). 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.

Completeness of the dataset Ultimately, we obtain 1.8 billion tweets in French between August 1st 2018 and July 31st 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

[15] The only way to collect the full volume of tweets would be to purchase it directly from Twitter's static archival data. However, the price charged for a volume of tweets such as the one we use in this paper is completely prohibitive to researchers.

[16] The limits in terms of volume of tweets that we mention here were in effect at the time of our data collection but will most probably cease to be valid at the end of 2022, since Twitter announced that it will then only provide access to 10 million tweets per month through the Filtered stream.

[17] One for each author of the paper.

Electronic copy available at: https://ssrn.com/abstract=3663899

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[18] (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 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 were 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.[19] We find that our dataset contains 61% of the real number of emitted tweets for 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, and our data collection method does not allow us to capture the tweets that do not contain any text (e.g. the tweets that only include an animated gif).

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 Twitter activity. We hope that both our methodology and data could be of use in the future to other researchers interested in social media.

[18] 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.

[19] We focus on the users located in France given they have a higher probability of only publishing tweets in French, and we only capture here by construction the tweets in French. In Section 6 below, we discuss in length the issue of user location.

Electronic copy available at: https://ssrn.com/abstract=3663899

2.1.2 Descriptive statistics

Online Appendix Figure D.1 plots the daily number of tweets included in our dataset. We see that, on average, we capture around 5 million tweets a day (including retweets), of which 2 million were original tweets. As highlighted above, the focus of this article will be on the subperiod August 1st 2018-November 30th, 2018. However, we implemented the tweets collection beginning on July 20th, 2018 (rather than on August 1st) so as to be able (i) to compute the centrality of the Twitterers before the very first event, and (ii) to identify and exclude the events that began before August 1st and would otherwise either not be identified or be censored in our dataset.[20]

Table 1 provides summary statistics on the tweets we collected. For each of the tweets, we have information of their length (100 characters on average or 6.1 words), and know whether the tweet is a retweet of an existing tweet or an original tweet. 61% of the tweets in our dataset are retweets and 22% are "quotes", i.e. comments on the retweeted tweet.[21] Among the original tweets, 19% are in reply to other tweets.

We also gather information on the popularity of each of the tweets in our sample. On average, the tweets are retweeted 2.1 times, liked 4.2 times, and receive 0.22 replies.[22]

2.2 Data: Twitter users

Summary statistics Furthermore, we compute summary statistics on the Twitter users. Our dataset includes $4,395,129$ unique users. Table 2 provides these statistics the first time a user is observed in our data.[23] On average, the users tweeted $11,582$ times, liked $6,187$ tweets, and were following 520 other Twitter accounts the first time we observe them. The average year of the account creation is 2015 (Twitter was created in 2006; see online Appendix Figure D.2 for the distribution of the users depending on the date on which they created their account). On average, users have $1,645$ 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 (online Appendix Figure D.3).

0.5% of the users in our sample have a verified account[24], 0.2% are the accounts of journalists, and 0.02% are media outlets' accounts. We have manually identified the Twitter accounts of the media outlets. For the Twitter accounts of journalists, we carry out a semi-

[20] Hence, we only keep the events that began on August 1st, or after. We proceed similarly at the end of our sample.

[21] Quote Tweets share another tweet with an additional comment, unlike Retweets that repost the original tweet without any modification.

[22] These numbers are only computed on the original tweets, given retweets, likes and quotes are not attributed to the retweets but to the original tweets.

[23] Alternatively, we compute the users' characteristics the last time we observe them. The results are presented in the online Appendix Table C.1.

[24] 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.

Electronic copy available at: https://ssrn.com/abstract=3663899

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 1: Summary statistics: Tweets (full sample)

	Mean	St.Dev	P25	Median	P75	Max	Obs
Characteristics of the tweet							
Length of the tweet (nb of characters)	100	53.93	60	97	140	1400	1,705,453,269
Number of words	6.1	4.08	3	6	8	260	1,705,453,269
=1 if the tweet is a retweet	0.61	0.49	0	1	1	1	1,705,453,269
=1 if the tweet is a reply	0.19	0.4	0	0	0	1	1,705,453,269
=1 if the tweet is a quote	0.22	0.41	0	0	0	1	1,705,453,269
Popularity of the tweet							
Number of retweets	2.1	112.76	0	0	0	510000	659,639,905
Number of likes	4.2	241.25	0	0	0	1600000	659,639,905
Number of replies	0.22	10.12	0	0	0	84000	659,639,905

Notes: The table gives summary statistics. Time period is June 18 2018 - August 10 2019. Variables are values for all the tweets included in our dataset (i.e. after filtering the tweets following the methodology described in Section 2.1 and online Appendix Section A.1). 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 details in the text.

Table 2: Summary statistics: Twitter users (full sample)

	Mean	St.Dev	P25	Median	P75	Max
User activity						
Total number of tweets	11,582	36,925	44	745	6,598	4,669,894
Nb of tweets user has liked	6,187	20,387	25	457	3,449	2,783,878
Nb of users the account is following	520	3,883	45	140	375	1,672,681
User identity						
Date of creation of the account	2015	3	2013	2016	2018	2019
=1 if user located in France	0.54	0.50	0	1	1	1
=1 if verified account	0.005	0.067	0	0	0	1
=1 if user is a journalist	0.002	0.043	0	0	0	1
=1 if user is a media	0.0002	0.014	0	0	0	1
User popularity						
Nb of followers	1,645	64,904	9	69	322	30,648,885
Nb of public lists	16	549	0	0	4	1,028,761
PagerRank centrality	0.023	0.38	0.014	0.014	0.014	299
Observations	4,395,129					

Notes: The table gives summary statistics. Time period is August 2018 - July 2019. Variables are values for all the Twitter users included in our dataset the first time we observe them. The variable "= 1 if user located in France" is defined with respect to the users whose information on the location is available. Variables are described in more details in the text.

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.

Finally, we collect information on the **location of the users**. On Twitter, the main way to identify this location is through the fact that users can indicate their location in their profile ("user-defined location"): nearly two thirds of the users in our sample do so.[25] 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 $2,708,411$ users for which the location field is filled (i.e. around 62% of the users in our dataset), the information provided allows us to recover the country where the user is located in 82% of the cases. 55% of these users indicate that they are located in France.

[25] Alternatively, one can rely on the location of the tweet. Users can indeed share their location with Twitter at the time when they tweet. First, when they decide to assign a location ("place") to their tweet, users are presented with a list of candidate Twitter places. However, out of our $4,395,135$ unique users, only $49,133$ do so for the first of their tweets that we observe. Next, the "place" field is always present when a Tweet is "geo-tagged". However, even fewer users provide their exact location (0.22%).

Electronic copy available at: https://ssrn.com/abstract=3663899

Users' centrality As highlighted by Azzimonti and Fernandes (2018), there are several statistics in the literature that can proxy for the degree of influence or centrality of an agent. In this article, we use Google's *PageRank* centrality (a variant of eigenvector centrality commonly used in network analysis), a measure that has the advantage *"to account not only by quantity (e.g. a node with more incoming links is more influential) but also by quality (a node with a link from a node which is known to be very influential is also influential)"* (Azzimonti and Fernandes, 2018). In other words, this algorithm relies on the idea that the importance of a node is related to the importance of the nodes linking to it, in a recursive approach.

Specifically, to compute the centrality of a given user i on a given day t in the network of retweets, we proceed as follows. First, for each day, we build a directed network of users[26], where an edge from i to j corresponds to the fact that user i has retweeted user j during the past 40 days, and we measure the total number of times she has done so. Second, we use the PageRank algorithm (Page et al., 1999) to compute the daily centrality of each user in that network.[27]

Compared to other measures of centrality such as Bonacich centrality and "diffusion centrality" (Banerjee et al., 2013, 2014), the main advantage of the PageRank algorithm is that it is designed to handle directed networks. Is is also optimized for very large networks with millions of nodes. Further, PageRank is a simple and robust algorithm adopted by other papers such as Xu and Corbett (2019) and Devereux et al. (2020) in different contexts.

2.3 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: August 2018-November 2018 and December 2018-July 2019. We used the four-month dataset covering August 2018-November 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 December 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 four months of our dataset. This sample includes $531,050,553$ tweets; online Appendix Table C.2 presents summary statistics for these tweets.

2.4 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 chan-

[26] A directed network is a network in which the links between nodes have a direction. In the present case, the direction is important, since "user i retweeted user j" is not equivalent to "user j retweeted user i".

[27] We rely on the Python networkx implementation to do so.

Electronic copy available at: https://ssrn.com/abstract=3663899

nels, TV stations, online-only news media, and the content produced by the "Agence France Presse" (AFP) news agency as well as the French dispatches of the news agency Reuters. 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 14 French-speaking (non-French) media outlets such as the daily newspaper *Le Temps* 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.

Newsroom characteristics Our dataset includes 200 unique media outlets (see online Appendix Section A.2.1 for the list of these media depending on their offline format), which published 1,239,552 online news articles between August 1st 2018 and November 30th 2018. Table C.3 shows summary statistics for these mainstream media.[28]

First, we gather information on their social media presence. We identify their Twitter account(s)[29] and collect information on their popularity (number of followers and public lists, captured the first time we observe them), as well as on the number of tweets they posted. On average, the media outlets in our sample have 2.9 different Twitter accounts. We compute the date of creation of each of these accounts, and report the oldest one in the table. In addition, for each of the media, we compute its number of journalists with a Twitter account, as well as the characteristics of these accounts.

Second, to better understand the mechanisms at play, we collect additional information on the media: their year of creation; the year of creation of their website; as well as information on their business model. In particular, we investigate whether the media uses a paywall (at the time of the data collection), the characteristics of this paywall (e.g. soft vs. hard), and the date of introduction of the paywall. This information is summarized in the online Appendix Figure D.5: while 52% of the media outlets do not have a paywall, 39.5% lock at least some of their articles behind a paywall (soft paywall). Metered paywalls and "watching an ad" paywalls are much less frequent (3.4 and 4.5% respectively), as well as hard paywalls. Overall, even if some of them do not (e.g. the pure online media Mediapart), the large majority of the media outlets rely at least partly on advertising revenues.

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 compute information on the size of their newsroom (Cagé, 2016). This information is available for 68 media outlets in our sample.

[28] 59.6% of the articles come from the newspaper websites, 14.7% from pure online media, 8.4% from the news agencies, 11.1% from the radio station websites and the remainder from TV channel websites (see online Appendix Figure D.4).

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

Electronic copy available at: https://ssrn.com/abstract=3663899

Fourth, for the 79 media outlets for which this information is available, we collect daily audience information from the ACPM, the French press organization whose aim is to certify circulation and audience data. Besides, to improve our understanding of the welfare effects of social media, we recover information on the characteristics of the audience. First, we rely on the ACPM's OneNext study that provides information on the age of the readers of the newspapers and magazines sites. Online Appendix Figure D.6 provides summary statistics for the 50 sites for which this information is available. Second, we use data from the Reuters Institute's 2018 Digital News Report (Reuters Institute, 2018), which allows us to recover information on the characteristics of the online consumers of 21 media outlets as well as of private radio stations (pulled together) (online Appendix Tables C.5 and C.6). In particular, this includes information on whether consumers use Twitter.[30]

Finally, we rely on "Decodex", the *Le Monde*'s fact-checking product to estimate the "reliability" of the media outlets included in our dataset. The outlets are classified into three categories: (i) the reliable sites, (ii) the unreliable sites, which occasionally publish fake news and do not cite their sources, and (iii) the false sites, which regularly disseminate false information.[31] Out of the 200 media outlets in our data, 102 are classified by *Le Monde*; 90 of them are reliable sites, 5 unreliable sites, and 7 false sites (online Appendix Figure D.7).

Article characteristics Online Appendix Table C.4 presents summary statistics for the $1,239,552$ articles included in our dataset. On average, articles are $2,343$ characters long. We compute their originality rate, i.e. the share of the article that is "original" and the share that is copy-pasted (Cagé et al., 2020). We also collect information on the number of times they are shared on Facebook to proxy their audience.[32]

Finally, in order to detect the topic of the events, we start by determining the topic of each of the news articles. To do so, we train a classifier to tag each news article with the 17 top-level IPTC categories (see online Appendix Section A.3 for details). Our training dataset is composed of the AFP dispatches for the years 2012 to 2017, i.e. the 6 years preceding our corpus (around 1.5 million dispatches). We use a neural network approach with TF-IDF encoding of the documents; Section A.4 describes the details of the procedure. Once trained, our network allows us to predict for any news article a score between 0 and 1 for each of the IPTC categories.

In the remainder of this section, we describe the new event detection 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.

[30] The survey data includes $2,006$ individuals for France in 2018.

[31] A fourth category is defined by *Le Monde*, but does not cover any of the websites included in our dataset: the satirical sites, which are deemed irresponsible with their labeling or in the kind of content and sentiment they are projecting.

[32] 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).

Electronic copy available at: https://ssrn.com/abstract=3663899

2.5 Event detection algorithms

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 Eastern Iburi earthquake on September 6th 2018 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.

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 specificity of social network data (short texts, use of slang, abbreviations, hashtags, very high volume of data, etc.) makes all "general" detection tasks (without specification of the type of topic) very difficult on tweet datasets. Hence, for this research project, we develop a novel algorithm to detect social media events on a very large unspecified dataset of tweets.[33]

In a nutshell, our approach consists in modeling the event detection problem as a dynamic clustering problem, using a "First Story Detection" (FSD) algorithm. 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 12 hours 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. Regarding the type of tweet representation, we use the TF-IDF model (we tested several models in our companion work – including Word2Vec, ELMo, BERT and Universal Sentence Encoder – and show that TF-IDF provides the best performances (Mazoyer et al., 2020b); similarly, we used different representations of tweets, counting the representation of both text and images, and obtain the best performances with only text). 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 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.

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 $time_difference(MME, SME) \leqslant d$

[33] See online Appendix Section B and 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 – in particular, the "Support Vector Machines", the "First Story Detection", the DBSCAN, and the "Dirichlet Multinomial Mixture" algorithms – and detect in this paper the social media events using the algorithm whose performance has been shown to be the best.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 3: Summary statistics: Joint events that appear first on Twitter

	Mean	St.Dev	P25	Median	P75	Max
Length of the event (in hours)	755	666	203	531	1,145	2,849
Number of documents in event	2,296	5,275	289	823	2,343	177,280
Twitter coverage						
Nb of tweets in event	2,252	5,238	274	798	2,285	177,268
Number of different Twitter users	1,612	2,684	234	662	1,827	46,262
Average number of retweets of tweets in events	2.5	3.9	0.9	1.5	2.7	89
Average number of replys of tweets in events	0.3	0.4	0.1	0.2	0.4	6
Average number of favorites of tweets in events	4.2	9.3	1.1	2.2	4.5	279
Media coverage						
Number of news articles in the event	44	81	13	20	42	2,070
Number of different media outlets	17	11	9	13	21	84
Observations	3,904					

Notes: The table gives summary statistics. Time period is August 1st 2018 - November 30th 2018. The observations are at the event level. Variables are described in details in the text.

($d = 1$ day 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.3. We obtain $3,992$ joint events, of which $3,904$ break first on Twitter. Table 3 presents summary statistics on these events that contain on average $2,252$ tweets and 44 media articles published by 17 different media outlets.

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.

Electronic copy available at: https://ssrn.com/abstract=3663899

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 is of particular importance. In this section, we present the empirical strategy we develop to tackle this question; the empirical results are displayed in Section 4 below.

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:

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

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

Our dependent variable of interest, news coverage$_e$, is the intensity of the media coverage that we proxy alternatively 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 (logarithm of the) total number of tweets in the event *before* the publication of the first media article, and captures the popularity of the event on Twitter.

$\mathbf{Z}'_{\mathbf{e}}$ 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 the topic(s) of the event, and for calendar-month (ω_m) and day-of-the-week fixed effects (λ_d).

Media-level approach Bias in editorial decisions may vary depending on the characteristics of the media outlet. To investigate whether this is the case, we exploit within media outlet variations (in this case, the standard errors are clustered at the event level). Our specification is as follows:

$$\text{news coverage}_{ec} = \alpha + \beta \,\text{number of tweets}_e + \mathbf{Z}'_{\mathbf{e}}\gamma + \delta_c + \omega_m + \lambda_d + \epsilon_{ecmd} \tag{2}$$

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,

Electronic copy available at: https://ssrn.com/abstract=3663899

the average length and originality of the articles. number of tweets$_e$ and $\mathbf{Z'_e}$ are defined as in equation (1). δ_c, ω_m and λ_d are respectively fixed effects for media, calendar month and DoW.

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 on Twitter at the time of the event.

Specifically, our instrument is the interaction between the first Twitter users' centrality in the network (measured just before the event) and the news pressure in the social media at the time of the first tweets on the event, 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 users' centrality and news pressure should only affect traditional news production through its effect on the tweets' visibility on Twitter. We make sure that our results are robust to dropping the events whose seed is the Twitter account of a media outlet or journalist.

In exploiting the structure of the Twitter network, the intention of our identification strategy is to mimic a hypothetical experiment that would break the correlation between popularity and unobserved determinants of the story's intrinsic interest. While both centrality and news pressure are likely correlated with the popularity of an event on Twitter, none of them can be used separately as an instrument given that they presumably violate the exclusion restriction. E.g. news pressure on Twitter at a given moment of time likely correlates with news pressure on traditional media, for instance during an election night or because a terrorist attack is under way. Similarly, even if the fact of using PageRank centrality – i.e. a measure of centrality computed at the second rank rather than the number of followers itself – partly alleviates this concern, the identity of the person(s) tweeting might be both correlated with the popularity generated by her tweets and with the traditional media interest for these tweets (e.g. because she is a celebrity or a prominent politician).

Regarding centrality, our source of variation, in the spirit of an intention-to-treat approach, comes from the number of "impressions" generated by the Twitterers' previous tweets.[34] The intuition here is that, regardless of the content of the tweet itself, the tweets emitted by central users (as defined using PageRank) have a much higher probability of being retweeted than the tweets emitted by non-central users, everything else equal (and in particular controlling for the

[34] As described in Section 2.2 above, centrality is computed for each of the user in the dataset on a daily basis; this allows us to measure the centrality of the users just before the event in which they tweet.

Electronic copy available at: https://ssrn.com/abstract=3663899

Twitterers' number of followers). 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. In our preferred specification, we use the maximum value of the centrality of the first twenty Twitterers in the event; in the Robustness Section 6.1, we show that our results are robust to the use of alternative measures of centrality.

We measure news pressure on Twitter by the number of interactions (retweets, favorites and replies) generated by all the tweets published in the hour preceding the first tweet in the event. We use the number of interactions rather than the 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, favorites and replies, on the contrary, is known at any given time because it is a metadata provided with each tweet. The intuition is that, for a given intrinsic interest, a tweet has a higher probability of being seen by Twitter users – and so of generating more traffic on Twitter – if it is posted at a time of low rather than high pressure. Everything else equal, this probability will be even higher if the first Twitterers are central in the network. In our preferred specification, we rely on a binary approach of pressure and define "low news pressure" on Twitter as an indicator variable equal to one if the number of interactions generated by the tweets posted in the hour preceding the event is in the bottom ninety percent of the distribution[35] Importantly, in the Robustness Section 6.1, we show that our findings are robust to considering an alternative measure of pressure, where we isolate a non-news dimension: we estimate the number of interactions generated by all the tweets except the tweets generated by the Twitter accounts of journalists and media.

Formally, we estimate the following model:

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

$$\text{news coverage}_e = \zeta_1 \widehat{\text{number of tweets}}_e + \zeta_2 \text{centrality}_e + \zeta_3 \text{low news pressure}_e + \mathbf{X}'_{\mathbf{e}}\eta + \omega_m + \lambda_d + \varepsilon_{emd} \quad (4)$$

where, in the first stage (equation (3)), number of tweets$_e$ is as before the (logarithm of the) total number of tweets published in the event e before the publication of the first media article, and, in equation (4), the dependent variable is the intensity of the media coverage. The instrument, centrality$_e$ $\times$ low news pressure$_e$, is relevant if $\beta_1 > 0$; we show below that it is a strong and significant predictor of the popularity of an event on Twitter.

[35] Online Appendix Figure D.8 plots the distribution of the number of interactions generated by these tweets.

Electronic copy available at: https://ssrn.com/abstract=3663899

3.3 Validity of the exclusion restriction

The interaction between users' centrality and low news pressure on Twitter can serve as a valid instrument for an event popularity on Twitter if the assumptions of exogeneity and the exclusion restriction are satisfied. Namely, the interaction between centrality and pressure is not correlated with factors that affect news coverage, other than through its effect on the activity on Twitter. In particular, our instrument should not be related to the underlying newsworthiness of an event. While we cannot rigorously test this assumption, we can show that we cannot predict a number of observable event characteristics with our instrument.

First, in the online Appendix Table C.7, we show that our instrument is not correlated with the topic of the event. The topic of the event is defined using the IPTC categories described above (Section 2.4); for each of the topics, we use the average score (defined between 0 and 1) over all the news articles published in the event. We associate to each event the topic whose probability has the highest score and, using a logistic regression, show that our instrument does not predict any of the topics.

As a second observable characteristic of the events, we compute the number of named entities. To do so, we extract the named entities from all the media articles using the Spacy library[36], and investigate in the online Appendix Table C.8 the correlation between the total number of entities in the event (controlling for the total number of words in the event, given that this number varies with the overall media coverage) and our instrument. We find that our instrument does not predict the number of named entities, whether we consider the mentions of places (e.g. Paris, Columns (1) to (3)), organizations (e.g. the United Nation, Columns (4) to (6)), or individuals (e.g. Emmanuel Macron, Columns (7) to (9)).

Finally, as a third observable characteristic of the events, we calculate the share of the media articles in the event published by media outlets depending on their reliability (as measured by Decodex and described above). Online Appendix Table C.9 presents the results; reassuringly, we show that, just like for the number of named entities and the topic of the events, our instrument is not correlated with the share of articles published by reliable, unreliable or false sites in the event, nor with the share of articles published by media outlets that have not been classified by Decodex. Overall, these findings bolster our confidence that our instrumental variable can be viewed as being "as good as randomly" distributed across the news stories.

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

[36] https://spacy.io/

Electronic copy available at: https://ssrn.com/abstract=3663899

at the event level, before turning to the media-level estimations.

4.1 Naive estimates

Event-level analysis Table 4 reports the results of the estimation of equation (1). In Columns (1) to (5), the outcome of interest is the total number of articles published in the event, and in Columns (6) to (10) the number of unique media outlets which cover the event. We find a positive correlation between the number of tweets published in an event before the first news article ("Log # tweets") and the total number of articles published in the event: according to our estimates, a fifty-percent increase in the number of tweets is equivalent to an increase of 1.46 in the number of articles, corresponding to 3.42% of the mean (Column (1)). This finding is robust to adding a number of event-level controls (Columns (2) to (4)), and to dropping the events whose seed is the Twitter account of a media outlet or journalist (Column (5)).

The positive relationship between the popularity of the event on Twitter and the media coverage is driven both by the overall number of articles published and by the number of unique media outlets covering the event (Columns (6) to (10)). According to our estimates, a fifty-percent increase in the number of tweets is associated with an increase of 0.44 in the number of media outlets covering the event (2.66% of the mean).

Media-level analysis Table 5 shows the estimates when we perform the media-level analysis (estimation of equation (2)); the unit of observation is a media-event. All the media outlets are included in Columns (1) and (2) (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 4, we find a positive relationship between the popularity of the event on Twitter and the media coverage it receives. A fifty-percent increase in the number of tweets published in the event before the first media article is correlated with an increase in the average number of articles that *each media outlet* publishes in the event corresponding to 2.57% of the mean.

In Columns (3) to (8), we restrict our sample to the media outlets that publish at least one article in the event, and consider the number of articles they publish (Columns (3)-(4)), the length of these articles (Columns (5)-(6)), and their originality (Columns (7)-(8)). We see that the positive impact of popularity is driven by the extensive margin; the correlation with the number of articles published (conditional on publishing at least one article in the event) is indeed not statistically significant. Further, we find a negative correlation with the length of these articles (Columns (4) and (5)). This length can be considered a proxy – albeit an imperfect one – for the quality of the articles. An alternative proxy is the originality of the articles (Cagé et al., 2020); we show that a one-percent increase in the number of tweets is associated with an increase in the originality rate corresponding to 1.3% of the mean (Columns (7) and (8)).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 4: Naive estimates: Event-level approach

	Number of articles					Number of media				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Log # tweets	3.02***	3.17***	3.26***	3.33***	3.23***	1.01***	1.09***	1.10***	1.11***	1.10***
	(0.99)	(1.04)	(1.03)	(1.06)	(1.10)	(0.11)	(0.11)	(0.11)	(0.11)	(0.11)
Log # seed's followers		1.32**					0.11			
		(0.57)					(0.07)			
# seed's followers			-0.60	-0.51	-1.13			-0.14	-0.14	-0.36
			(0.59)	(0.59)	(1.10)			(0.12)	(0.12)	(0.25)
# seed's followers-squared			0.01	0.01	0.01			0.00	0.00	0.00
			(0.01)	(0.01)	(0.02)			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				2.96	3.11				0.27	0.29
				(2.65)	(2.69)				(0.35)	(0.36)
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓		✓	✓	✓	✓
Drop media & journalist					✓					✓
Observations	3,904	3,904	3,904	3,904	3,773	3,904	3,904	3,904	3,904	3,773
Mean DepVar	43.5	43.5	43.5	43.5	43.5	16.6	16.6	16.6	16.6	16.6
Sd DepVar	81.1	81.1	81.1	81.1	81.8	11.1	11.1	11.1	11.1	11.1

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 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 calendar-month fixed effects, and in Columns (2) to (5) and (7) to (10) we also control for the topic of the event ("Topic of the event"). Columns (1) to (4) and (6) to (9) report the estimates for all the events that appear first on Twitter; in Columns (5) and (10) we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). The number of tweets is computed *before* the first news article in the event. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 5: Naive estimates: Media-level approach

	All media outlets		Conditional on covering the event					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	# articles	# articles	# articles	# articles	Length	Length	Originality	Originality
Log # tweets	0.020***	0.019***	0.077	0.074	-21.404***	-23.132***	0.012***	0.013***
	(0.006)	(0.006)	(0.059)	(0.062)	(7.467)	(7.585)	(0.001)	(0.001)
# seed's followers	-0.003	-0.007	-0.015	-0.035	8.610	9.492	0.000	0.001
	(0.003)	(0.006)	(0.026)	(0.077)	(8.653)	(21.661)	(0.002)	(0.006)
# seed's followers-squared	0.000	0.000	0.000	0.000	-0.375**	-0.391	0.000	0.000
	(0.000)	(0.000)	(0.001)	(0.001)	(0.175)	(0.389)	(0.000)	(0.000)
=1 if first tweet during night	0.018	0.019	0.145	0.148	-21.607	-16.995	-0.008*	-0.008*
	(0.016)	(0.016)	(0.130)	(0.132)	(22.478)	(22.817)	(0.004)	(0.004)
Media FEs	✓	✓	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓	✓	✓
Drop media & journalist		✓		✓		✓		✓
Observations	658,917	636,844	64,806	62,572	64,806	62,572	64,806	62,572
Clusters (events)	3,904	3,773	3,904	3,773	3,904	3,773	3,904	3,773
Mean DepVar	0.3	0.3	2.6	2.6	2,570.3	2,563.9	0.4	0.4
Sd DepVar	2.0	2.0	5.9	5.9	1,424.4	1,423.5	0.4	0.4

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using a linear-log model 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, calendar month, and media fixed effects, and we also control for the topic of the event ("Topic of the event"). Odd columns report the estimates for all the events that appear first on Twitter; in even Columns, we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). In Columns (1) and (2), all the media outlets are included and the dependent variable is the number of articles published by the media in the event. In Columns (3) to (8), we only include the media outlets that publish at least one article in the event. In Columns (3)-(4), the dependent variable is the number of articles they publish, in Columns (5)-(6) the length of these articles, and in Columns (7)-(8) their originality. The number of tweets is computed *before* the first news article in the event appears. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

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 6, we begin by reporting the results of the reduced-form estimation in Columns (1) to (4). In Column (1), we look at the relationship between the centrality of the first Twitterers and the number of articles published in the event; as expected, we find that it is positive and statistically significant. In Column (2), our explanatory variable of interest is the social media news pressure at the time of the event; while we show that media coverage is higher at times of low pressure on Twitter, the estimated coefficient is not statistically significant.

Finally, in Columns (3) and (4), we investigate the relationship between our instrument – the interaction between centrality and news pressure – and media coverage (controlling for the direct effects of centrality and pressure). The coefficient we obtain for our instrument is positive and statistically significant: the media coverage of an event is higher when central Twitterers tweet at times of low news pressure on Twitter (compared to events involving non-central Twitterers or Twitterers tweeting at times of high pressure on Twitter, when their tweets get less impressions). This finding is robust to controlling for a number of event characteristics, as well as to dropping the news events whose seed is a media or a journalist (Column (4)). These reduced-form findings are reassuring as to the quality of our IV strategy.

In Columns (5) and (6), we report the first stage of the estimation (corresponding to equation (3)). The dependent variable is the logarithm of the number of tweets (published in the event before the first media article). We find that our instrument is positively and consistently associated with an increase in the number of tweets, regardless of the set of controls included and the sample used.

Finally, Table 7 reports the estimates of the second stage (equation (4)). For each regression, we report the F-statistic from the first stage; F-statistics are sufficiently high not to worry about the weak instrument problem. We find that a fifty-percent increase in the instrumented number of tweets is associated with an increase of 7 in the number of articles, corresponding to 16.4% of the mean (Column (4)). This finding is robust to the use of different specifications and control variables, and in particular to dropping the events whose seed is the Twitter account of a media or journalist (Column (5)).[37]

[37]Note that the magnitude of our IV estimates is larger than that of the OLS estimates. While this could initially seem surprising – given that our inability to control for the underlying newsworthiness of the events should in fact result in an upward bias in the OLS estimates – the IV estimates may in fact be larger due to measurement errors. A measurement error from the misclassification of tweets in joint events would indeed bias OLS estimates toward zero.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 6: Event-level approach, Reduced form and First stage estimates

	Reduced form				First stage	
	(1)	(2)	(3)	(4)	(5)	(6)
	# articles	# articles	# articles	# articles	Log # tweets	Log # tweets
Instrument						
Low pressure * Centrality			3.03**	3.35**	0.17***	0.14***
			(1.46)	(1.47)	(0.05)	(0.05)
Controls						
Centrality	2.21**		-0.31	-0.56	-0.19***	-0.16***
	(0.94)		(0.97)	(0.96)	(0.05)	(0.05)
Low pressure		2.57	35.71**	39.84**	2.04***	1.65***
		(2.42)	(16.30)	(16.40)	(0.57)	(0.59)
Log # seed's followers						
# seed's followers	-0.81	-0.33	-0.77	-1.47	0.08***	0.16***
	(0.67)	(0.60)	(0.67)	(1.06)	(0.02)	(0.05)
# seed's followers-squared	0.01	0.00	0.01	0.02	-0.00***	-0.00***
	(0.01)	(0.01)	(0.01)	(0.02)	(0.00)	(0.00)
=1 if first tweet during night			2.39	2.47	-0.26***	-0.25***
			(2.76)	(2.81)	(0.05)	(0.05)
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist				✓		✓
Observations	3,904	3,904	3,904	3,773	3,904	3,773
Mean DepVar	43.5	43.5	43.5	43.5	4.1	4.1
Sd DepVar	81.1	81.1	81.1	81.8	1.6	1.6

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. An observation is a news event. We only consider the subset of news events that appear first on Twitter. The dependent variable is the number of media articles published in the event in Columns (1) to (4) (reduced form estimates) and the logarithm of the number of tweets in Columns (5) and (6) (first stage estimates). Robust standard errors are reported between parentheses. All specifications include day-of-the-week and calendar-month fixed effects, and we also control for the topic of the event ("Topic of the event"). Columns (1) to (3) and (5) report the estimates for all the events that appear first on Twitter; in Columns (4) and (6), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 7: Event-level approach, IV estimates (Second stage)

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	21.1*	17.5*	17.2*	17.7*	24.1*
	(11.4)	(9.6)	(9.7)	(9.9)	(13.8)
Low pressure	3.533	3.136	2.696	-0.311	0.121
	(3.057)	(2.869)	(2.829)	(3.478)	(3.826)
Centrality	3.217**	2.612**	2.996**	3.130**	3.352**
	(1.254)	(1.192)	(1.268)	(1.299)	(1.423)
Log # seed's followers		0.674			
		(0.554)			
# seed's followers			-2.364*	-2.207*	-5.317*
			(1.237)	(1.194)	(2.865)
# seed's followers-squared			0.044*	0.041	0.094*
			(0.026)	(0.025)	(0.056)
=1 if first tweet during night				7.051*	8.503*
				(3.817)	(4.506)
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	10.6	13.2	12.5	12.2	7.5
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	43.5	43.5	43.5	43.5	43.5
Sd DepVar	81.1	81.1	81.1	81.1	81.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. An observation is a news event. Robust standard errors are reported between parentheses. All specifications include day-of-the-week and calendar-month fixed effects, and in Columns (2) to (4) we also control for the topic of the event ("Topic of the event"). Columns (1) to (3) report the estimates for all the events that appear first on Twitter; in Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Media-level approach We next turn to the media-level estimates. Table 8 reports the results of the second stage of the estimations.[38] We find a positive and statistically significant effect of the popularity of the event on Twitter on the media coverage: a fifty-percent increase in the instrumented number of tweets is associated with an increase of 0.044 in the number of articles written by each media outlet in the event, corresponding to 17% of the mean (Column (4)). As before, this effect is robust to dropping the events whose seed is the Twitter account of a media outlet or journalist (Column (5)); if anything, the magnitude of the estimated coefficients is then higher.

Finally, in the online Appendix Table C.11, 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 naive estimates (Table 5), conditionally on covering the event, we find no relationship between the popularity of an event on Twitter and the number of articles the media outlets devote to this event. Further, once we take into account the endogeneity of the number of tweets, we find no statistically significant relationship with the length of these articles, nor with their originality.

5 Mechanisms and welfare implications

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 that may be at play behind this relationship, in particular journalists' monitoring of Twitter and the existence of a click bias, and then discuss the welfare implications of our findings.

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 journalists' monitoring of Twitter. A growing literature in journalism studies indeed highlights the fact 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 for newspapers of three countries; they find that Twitter is more commonly used as a news source than Facebook.[39] Furthermore, McGregor and Molyneux (2018) show that US journalists using Twitter as part of their daily work consider tweets to be as newsworthy as headlines from the Associated Press. Consistently, Weaver et al. (2019) document that, as early as 2013, 60% of US journalists were using social

[38] See online Appendix Table C.10 for the corresponding reduced-form and first-stage estimations. 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, and that the first stage is positive and statistically significant.

[39] The same has been shown regarding local television newsrooms: in a nationwide survey of US news directors at network-affiliated television stations, Adornato (2016) show that social media are an important factor in choosing stories to cover. For additional evidence of Twitter as a reporting tool, see Vis (2013).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 8: Media-level approach, IV estimates (Second stage)

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.13*	0.11*	0.10*	0.11*	0.15*
	(0.07)	(0.06)	(0.06)	(0.06)	(0.09)
Low pressure	0.02	0.02	0.02	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Centrality	0.02**	0.02**	0.02**	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.01*	-0.01*	-0.03*
			(0.01)	(0.01)	(0.02)
# seed's followers-squared			0.00*	0.00	0.00*
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.04*	0.05*
				(0.02)	(0.03)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	658,917	658,917	658,917	658,917	636,844
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	9.75	12.32	11.69	11.33	7.02
Underidentification (p-value)	0.00	0.00	0.00	0.00	0.01
Mean DepVar	0.26	0.26	0.26	0.26	0.26
Sd DepVar	2.00	2.00	2.00	2.00	2.02

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects, and in Columns (2) to (5), we also control for the topic of the event ("Topic of the event"). In Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

media to find new ideas for their stories. Hence, 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).

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 9 reports the results of the estimation (online Appendix Table C.12 reports the associated naive estimates). We find that the relationship between the instrumented number of tweets and the number of articles is higher for the media that have a high number of journalists with a Twitter account (Columns (3) to (6)) than for those with only a few (Columns (1) to (3)).[40]

However, this finding may be partly driven by the fact that some media simply have *more journalists* than others (regardless of the social media presence of these journalists). As highlighted in the data section 2.4, 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.[41] Doing so does not affect our findings; on the contrary, the effect of the popularity on Twitter becomes even stronger for the media outlets whose journalists are relatively more present on the social network. Besides, we investigate whether the magnitude of our effects varies depending on the media outlets' strategy on Twitter. In the online Appendix Table C.13, we show that our effects are driven by the media outlets that tweeted a relatively high number of times (above the median) in the event.[42] Hence, while it cannot entirely explain our findings, the monitoring of Twitter by journalists seems to play a role here.

5.2 Editorial decisions and popularity

The causal relationship between the popularity of a story on Twitter and its mainstream media coverage can also be due to the existence of a clicks bias. This explanation would be consistent with the results of Sen and Yildirim (2015) which show, using data from a leading 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

[40] "High" vs. "low" is defined here with respect to the median number of journalists with a Twitter account (25).

[41] 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.

[42] However, this last result should be interpreted with a pinch of salt, given that the media outlets that cover an event are also the ones that produce articles on the event they can tweet about. Alternatively, in Table C.14, we estimate the relationship between the popularity of an event on Twitter and the subsequent media coverage depending on the overall media outlets' propensity to tweet on events (computed over all the joint events). We find that the relationship between popularity on Twitter and media coverage is positive and statistically significant for all the media outlets, but that the magnitude of the effect is higher for the outlets with a higher propensity to tweet on events.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 9: Media-level approach, IV estimates (Second stage), Depending on the number of journalists with a Twitter account

	Low nb journalists Twitter			High nb journalists Twitter		
	(1)	(2)	(3)	(4)	(5)	(6)
Log(Number of tweets)	0.03	0.03	0.05	0.22*	0.30*	0.40*
	(0.02)	(0.03)	(0.05)	(0.12)	(0.17)	(0.23)
Low pressure	0.01	0.00	0.01	-0.01	-0.00	-0.01
	(0.01)	(0.01)	(0.01)	(0.04)	(0.05)	(0.06)
Centrality	0.01*	0.01*	0.00	0.04**	0.04**	0.05**
	(0.00)	(0.00)	(0.00)	(0.02)	(0.02)	(0.02)
# seed's followers	-0.00*	-0.01	-0.01	-0.03*	-0.07**	-0.10**
	(0.00)	(0.01)	(0.01)	(0.01)	(0.04)	(0.05)
# seed's followers-squared	0.00*	0.00	0.00	0.00	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.01*	0.02*	0.03*	0.09*	0.11*	0.15**
	(0.01)	(0.01)	(0.02)	(0.05)	(0.05)	(0.07)
Total number of journalists			0.00***			0.00***
			(0.00)			(0.00)
Media FEs	✓	✓		✓	✓	
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist		✓	✓		✓	✓
Observations	240,394	232,378	52,016	274,942	265,687	182,497
Clusters (events)	3,904	3,773	3,773	3,904	3,773	3,773
Number of media outlets included	71	71	71	73	73	73
F-stat for Weak identification	10.7	6.6	7.1	12.1	7.5	7.7
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.1	0.1	0.1	0.5	0.5	0.7
Sd DepVar	0.8	0.8	0.7	3.0	3.0	3.5

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week and calendar month fixed effects, and we also control for the topic of the event ("Topic of the event"). In Columns (1)-(2) and (4)-(5), we control for media fixed effects, and in Columns (3) and (6), we instead control for the number of journalists working for the media (given this variable is time invariant at the level of the media outlets). 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, defined with respect to the median (25). In Columns (3) and (6), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

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, some of which rely on advertising revenues while others do not. We investigate the specific role played by their business model, and in particular whether they have a paywall and the kind of paywall they use.

Table 10 presents the results. Media outlets are ordered from the left to the right according to their reliance on advertising revenues. We see that the magnitude of the effect is much higher for the media outlets that fully ("No paywall") or strongly ("soft paywall" and "watch an ad") rely on advertising revenues (Columns (1) to (3)) than for media outlets whose online revenues mostly depend on subscriptions (Columns (4) and (5)). For the former, a fifty-percent increase in popularity leads to an increase in news coverage corresponding respectively to 22, 20.3 and 21.1% of the mean, compared to 6.2% of the mean for the outlets using a metered paywall (a coefficient that is furthermore not statistically significant).[43]

Hence our results seem to be at least partly driven by short-term considerations generated by advertising revenue-bearing clicks. Furthermore, 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; interestingly, in the online Appendix Table C.16, we show that the positive relationship between popularity on social media and online news coverage is positive both for private and for public media outlets considered independently (however, due to the low number of public media outlets, it is not statistically significant for the latter). 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 these preferences. We discuss below the extent to which such a signal might be biased, and the welfare implications of our findings.

Depending on the offline format Prior to that, we finally investigate whether our results vary depending on the offline format of the media outlets. Table 11 reports the estimates separately for (i) the national daily newspapers, (ii) the local daily newspapers, (iii) the weeklies and the monthlies, (iv) the pure online media, (v) the websites of the television channels, and (vi) the websites of the radio stations.

We show that the positive relationship between the popularity of an event on Twitter and its subsequent media coverage is first driven by the national daily newspapers (Column (1)) and the television channels (Column (6)), for which we observe respectively that a fifty-percent increase in the number of tweets before the event leads to an increase in the number of articles corresponding to 18.1% and 16.5% of the mean. Interestingly, the local newspapers do not seem to react to what is happening on Twitter (Column (2)), a finding that we link

[43] Alternatively, in the online Appendix Table C.15, for the subset of the media outlets for which we were able to recover the information (see Section A.2.2 for details on the data construction), we investigate the extent to which our results vary depending on the media dependence on advertising revenues. We separate the media outlets in our sample into four quartiles of share of advertising revenues in total revenues, and obtain results consistent with those presented in Table 10, i.e. the lower the reliance on advertising, the lower the magnitude of the relationship between the popularity of an event on Twitter and the subsequent media coverage.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 10: Media-level approach, IV estimates (Second stage), Depending on the paywall

	No paywall	Soft paywall	Watch an ad	Metered	Hard paywall
	(1)	(2)	(3)	(4)	(5)
	# articles	# articles	# articles	# articles	# articles
Log(Number of tweets)	0.13*	0.17*	0.24*	0.02	0.03
	(0.07)	(0.10)	(0.14)	(0.04)	(0.03)
Low pressure	0.00	-0.01	0.04	-0.02	0.00
	(0.02)	(0.03)	(0.05)	(0.02)	(0.01)
Centrality	0.02***	0.03**	0.04**	0.01	0.00
	(0.01)	(0.01)	(0.02)	(0.01)	(0.00)
# seed's followers	-0.01*	-0.02**	-0.03*	-0.01	-0.00
	(0.01)	(0.01)	(0.02)	(0.01)	(0.00)
# seed's followers-squared	0.00	0.00*	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.05*	0.06*	0.09**	0.02	0.02**
	(0.02)	(0.04)	(0.05)	(0.02)	(0.01)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Observations	395,535	175,022	18,297	29,596	19,724
Clusters (events)	3,904	3,904	3,904	3,904	3,904
Number of media outlets included	113	47	6	8	7
F-stat for Weak identification	9.0	9.2	10.2	10.1	7.1
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.24	0.34	0.46	0.13	0.04
Sd DepVar	2.29	1.53	2.07	0.96	0.30

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week, calendar-month and media fixed effects. Column (1) includes the media outlets that do not have a paywall, Column (2) those that use a soft paywall, Column (3) those that require consumers to watch an ad before been allowed to read the articles, Column (4) the media outlets that use a metered paywall, and finally Column (5) the outlets that rely on a hard paywall. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

to the characteristics of their audience (see also below).

5.3 Welfare implications

What are the welfare effects of social media? In particular, would citizens be better informed in the absence of Twitter? In this article, we have investigated the extent to which traditional media outlets react to the popularity of stories on Twitter. We have documented heterogeneity in media outlets' response depending on their characteristics, and in particular on whether they offer digital news for free. Our finding – that media outlets providing free content are more sensitive to what is happening on Twitter than those that are behind a paywall – implies that citizens, depending on their willingness and/or capacity to pay for news, will be offered different kinds of news. This may partly explain the extent of existing information inequality (Kennedy and Prat, 2019). If the contagion from social to mainstream media were to further increase, this would also imply that already less-informed citizens will be even more exposed to click-bait news, which might in turn affect voting outcomes.

Furthermore, given the instrumental variable strategy we use to isolate the causal impact of a story's popularity on Twitter, our estimates capture the effects of a variation in popularity that is uncorrelated with a story's 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 the online Appendix Table C.17, 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.[44] 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, 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%; Table C.18).

Hence journalists – because they tend to be on Twitter – rely on this social media platform 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. As highlighted in April 2022 by Dean Baquet, former executive director of *The New York Times*, Twitter warps how journalists report news *"by changing who they see as their audience."*[45]

[44] As highlighted above, the survey data includes 2,006 individuals. Of these, 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.

[45] The quote is taken from a NiemanLab's article published by Joshua Benton (https://www.niemanlab.org/2022/04/the-new-york-times-would-really-like-its-reporters-to-stop-scrolling-and-get-off-twitter-at-least-once-in-a-while/), which also refers to Dean Baquet's memo to the *Times*

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 11: Media-level approach, IV estimates (Second stage), Depending on the offline format of the mainstream media outlets

	Nat. dail.	Local dail.	Weeklies & Monthlies	Pure online	TV	Radio
	(1)	(2)	(3)	(4)	(5)	(6)
	Nb articles	Nb articles	Nb articles	Nb articles	Nb articles	Nb articles
Log(Number of tweets)	0.21*	0.07	0.05	0.03	0.24*	0.09
	(0.11)	(0.06)	(0.03)	(0.02)	(0.14)	(0.08)
Low pressure	0.00	-0.03	0.01	0.00	0.00	-0.02
	(0.04)	(0.02)	(0.01)	(0.00)	(0.05)	(0.03)
Centrality	0.04**	0.01	0.01**	0.01**	0.05***	0.02**
	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)	(0.01)
# seed's followers	-0.03**	-0.01*	-0.01**	-0.00	-0.04**	-0.02*
	(0.01)	(0.01)	(0.00)	(0.00)	(0.02)	(0.01)
# seed's followers-squared	0.00*	0.00	0.00*	0.00	0.00*	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.06	0.05*	0.02	0.01	0.09	0.06*
	(0.04)	(0.03)	(0.01)	(0.01)	(0.06)	(0.03)
Media FEs	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Observations	74,176	98,222	183,808	226,272	27,328	41,303
Clusters (events)	3,904	3,904	3,904	3,904	3,904	3,904
Number of media outlets included	19	26	53	69	7	11
F-stat for Weak identification	12.28	11.96	11.14	10.57	12.27	11.90
Underidentification (p-value)	0.00	0.00	0.00	0.00	0.00	0.00
Mean DepVar	0.47	0.27	0.16	0.07	0.59	0.37
Sd DepVar	1.78	1.39	1.06	0.70	2.28	1.58

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. The dependent variable is the number of articles. An observation is a media-news event. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. Column (1) includes the national daily newspapers, Column (2) the local daily newspapers, Column (3) the weeklies and the monthlies, Column (4) the pure online media, Column (5) the television channels, and Column (6) the radio stations. All specifications include day-of-the-week, calendar-month and media fixed effects, and we also control for the topic of the event ("Topic of the event"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

What is 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.

Note however that the magnitude of our estimated effects varies depending on the media consumers' characteristics, and in particular on their use of Twitter. In the online Appendix Table C.19, using the previously described Reuters' data, for the subset of the media outlets covered in the survey we separate them into outlets whose consumers use Twitter to a relatively high degree, and those whose consumers use it to a relatively low degree. We find that our effects are of larger magnitude (and only statistically significant) for the former. Further, we show in Table C.20 that the magnitude of the effect tends to be larger for the media outlets whose audience is relatively young.[46]

Hence, while a growing literature documents that journalists are now exposed to copious quantitative data on 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 deciding 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 is not interested in the news produced by the media. In 2018 in France, only 62% of the individuals 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 a 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, we proxy this number by the number of times each article has been shared on Facebook.[47]

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 (4), but the dependent variable is now the average number of times the articles published in the event are shared on Facebook. Table 12 presents the results: we find no statistically significant relationship between the instrumented popularity of an event on Twitter and the number of times the articles published by the media outlets in this event are shared on Facebook, whether we consider the average number of shares of each article on Facebook (Columns

newsroom staff: *"We can rely too much on Twitter as a reporting or feedback tool – which is especially harmful to our journalism when our feeds become echo chambers."*

[46] In the online Appendix Table C.21, we show consistent evidence using the ACPM's OneNext data to measure the age of the audience. However, in this case, the difference is not statistically significant between the media outlets whose audience is relatively young and the ones whose audience is relatively old.

[47] 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. Online Appendix Table C.4 provides summary statistics on the average number of shares, comments, and reactions on Facebook received by the articles in our sample.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table 12: IV estimates (Second stage) – Popularity on Twitter and demand for news

	Average number of shares on Facebook			Sum # shares		
	(1)	(2)	(3)	(4)	(5)	(6)
Low pressure	0.5	1.4	1.4	66.3	84.7	83.8
	(3.4)	(3.6)	(3.6)	(112.7)	(116.5)	(112.0)
Seed centrality	1.615*	1.812*	1.901*	29.393	29.735	4.378
	(0.953)	(1.020)	(1.077)	(36.735)	(38.326)	(39.008)
# seed's followers	1.145	-4.328*	-4.469*	-28.521	-109.483*	-69.258
	(2.657)	(2.373)	(2.478)	(36.795)	(66.291)	(64.108)
# seed's followers-squared	-0.006	0.077*	0.080*	0.671	1.991	1.277
	(0.043)	(0.046)	(0.048)	(0.724)	(1.279)	(1.232)
=1 if first tweet during night	1.955	2.299	2.524	129.283	144.528	80.203
	(3.319)	(3.670)	(3.782)	(115.802)	(123.346)	(117.204)
Nb articles			-0.026			7.565***
			(0.020)			(1.841)
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist		✓	✓		✓	✓
Observations	3,904	3,773	3,773	3,904	3,773	3,773
F-stat for Weak identification	12.2	7.5	7.1	12.2	7.5	7.1
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0	0.0
Mean DepVar	19.8	19.7	19.7	639.8	639.0	639.0
Sd DepVar	87.4	86.4	86.4	2,764.8	2,783.8	2,783.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. The dependent variable is the average number of times articles published in an event have been shared on Facebook in Columns (1) to (3), and the total number of times in Columns (4) to (6). The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week and calendar-month fixed effects, and in Columns (2)-(3) and (5)-(6) we also control for the topic of the event ("Topic of the event"). Columns (1)-(2) and (4)-(5) report the estimates for all the events that appear first on Twitter; in Column (3) and (6) we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

(1) to (3)) or the overall number of shares received by all the articles in the event (Columns (4) to (6)). Hence, this finding further reflects the fact that the journalists' reliance on Twitter might distort the information they produce compared to what citizens actually prefer.

Publishers' incentives to invest in quality Finally, social media may affect publishers' incentives to invest in quality. For example, de Cornière and Sarvary (2019) show theoretically that, following the introduction of social media platforms, high-quality newspapers invest more in quality, while low-quality ones invest less.[48] In this article, we cannot isolate the specific

[48]Further, even if editors know that Twitter is an imperfect signal for consumer preferences, journalists may rely on the social network as a way to get news material cheaply. See e.g. Martin and McCrain (2019) who make a similar argument regarding Sinclair television stations in the US; the stations are fine with paying a rating penalty given that they benefit from clear economies of scale.

Electronic copy available at: https://ssrn.com/abstract=3663899

impact of the introduction of Twitter on the quality choices of the media outlets (given that this introduction affected all the media outlets simultaneously). However, we can examine whether media outlets' reactivity to what is happening on Twitter varies depending on their "quality" that we can proxy by the number of journalists. In the online Appendix Table C.22, focusing on the subset of media outlets for which the information is available, we report the magnitude of the effects separately for the media with "small" vs. "large" newsrooms (using the median number of journalists). While both media outlets with small and large newsrooms increase their coverage of events that are more popular on Twitter (yet not the size nor the originality of the articles they publish in these events), the magnitude of the effect is larger for media outlets with more journalists.

Furthermore, in the online Appendix Table C.23, we study whether there is heterogeneity depending on the "reliability" of the media outlets, which can be seen as an alternative proxy for their quality. Relying on the classification performed by Decodex, we consider separately the media labeled as "reliable" and the media considered as "unreliable" or "false" sites. We see that our results are entirely driven by the reliable sites (Column (1)). The effect is not statistically significant and close to zero for the unreliable (Column (2)) and false sites (Column (3)), even when we pull them together (Column (4)).

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 Section E.

Alternative measures of news pressure on social media In our preferred specification, we instrument the number of tweets by the interaction between the centrality of the users and the news pressure on Twitter just before the event, and 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. To ensure the validity of the exclusion restriction, as a first robustness check, we consider an alternative measure of pressure on Twitter, where we isolate a non-news dimension of pressure. More precisely, we measure pressure by the number of interactions generated by all the tweets *except the tweets generated by the Twitter accounts of journalists and of media outlets*. Online Appendix Tables E.1 and E.2 present the results; the main findings are unaffected, either qualitatively or quantitatively, by this alternative measure of pressure.

Explanatory variable In our preferred specification, we rely on the total number of tweets – including the retweets – to measure the popularity of an event on Twitter. In the online Appendix Tables E.3 and E.4, we show that our results are robust to solely using the number of *original* tweets (if anything, the magnitude of the estimated effects is larger).

Electronic copy available at: https://ssrn.com/abstract=3663899

Controls In our preferred specification, we use a reduced set of controls, including the seed of the event's number of followers and an indicator variable equal to one if the first tweet is tweeted during the night. In the online Appendix, we show that our results are robust to adding further controls. Table E.6 presents the event-level results.[49] We report in the first column our preferred specification for the sake of comparison. In Column (2), 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 she is included in, and her total number of tweets (all these characteristics are computed the first time we observe the user in our data). Further, in Column (3), we show that our results are robust to controlling for an indicator variable equal to one if the user is located in France; in Column (4), to the inclusion of an indicator variable equal to one if the language of the users is French; and in Column (5), to controlling for the date of creation of the user Twitter account. These findings – that our main estimates do not depend on the set of controls included – are of particular importance given that they corroborate the validity of our IV strategy. Indeed, they show that our instrument is efficient enough at capturing the endogeneity that might be linked to user characteristics.

Verified accounts Further, in Column (6), we show that our results are also robust to controlling for whether the Twitter account of the seed of the event is verified. Alternatively, in the online Appendix Table E.8, we show that our results are robust to dropping the events whose seed has a verified Twitter account. This is reassuring as to the fact that our findings are not driven by a celebrity bias or by tweets by influencers.

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.[50] 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 African and sub-Saharan African countries. Hence, how can we be sure that our results are not (at least partly) driven by media outlets and/or users based in these other countries?

First, it is important to note that, while the issue of users' location may seem an important one, this information is not available to the media; when visiting Twitter and reviewing tweets, journalists observe the language of the tweets but not the location of the users. Further, if people do read French – and we can assume 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 users' location (even if such information may matter for advertisers, depending on where they are

[49] The media-level estimates are reported in online Appendix Table E.7 and are similar.
[50] We discuss below the issue of the external validity of our results.

Electronic copy available at: https://ssrn.com/abstract=3663899

located).

As a robustness check, we nonetheless 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 Table E.5 presents the estimates; doing so does not affect our main results. Note also that controlling for the location of the users and/or their language does not affect our findings either (Tables E.6 and E.7).

Sample Next, we show that our results are robust to changes in our sample of analysis. First, we show that they are robust to using a winsorized value of the number of tweets, trimmed at the 99th percentile (Table E.9). Second, we verify that they are robust to dropping the media outlets that produce only a few articles (fewer than 10) on events during our period of interest (Table E.10). Finally, in Table E.11, we show that they are robust to excluding the news agencies AFP and Reuters, whose behavior might be different to that of media outlets aimed directly at consumers.

Measures of centrality Finally, we show that our results are robust to considering alternative measures of users' centrality. In our preferred specification, we take the maximum value of the centrality of the first 20 Twitterers in the event. In the online Appendix Figure E.1, we show that our results are robust to using the first 15 to 25 users.

In the end, note that our innovative – at least regarding the use of "real-world data" – split-sample approach should reassure readers as to the empirical validity of the results presented in this paper, which will be recomputed on a different sample for the final version of the article.

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 allows us to recover around 70% of all the tweets in French during our period of consideration. This would not have been possible with data from the United States for example, relying on English-language tweets. 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,

Electronic copy available at: https://ssrn.com/abstract=3663899

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 UK), or the proportion of the population who paid for online news (11%, like in Spain, but slightly above Germany or Canada – 8% – and 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 USA, 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 US 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. First, 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 that 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.

7 Conclusion

In May 2022, Elon Musk's bid to buy Twitter attracted widespread media coverage, demonstrating the importance of social media. As highlighted by Roger McNamee in *Time* magazine, *"journalists and politicians depend on the platform to share their ideas and build their brands."*[51] In other words, Twitter exerts outside power and influence on the public discourse. Twitter, therefore, like other social media, is a complex phenomenon and may have both positive and negative effects on people's welfare (Allcott et al., 2020). This also holds true for its impact on traditional media. According to Cision (2019)'s Global State of the Media Report, the bypassing of traditional media by social media is considered the biggest challenge 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 – and thus in turn the news available to and consumed by citizens. 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, and leverage the

[51] https://time.com/6172636/twitter-policy-failures-elon-musk/

Electronic copy available at: https://ssrn.com/abstract=3663899

enormity of our data to propose a novel instrument that allows us to isolate the causal impact of social media. Focusing on the stories that emerge first on Twitter, we show that their popularity on Twitter affects the coverage that mainstream media devote to these stories.

These findings shed 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. This contagion calls into question the business model of the legacy media, as well as the welfare effects of the platforms.

References

Adornato, Anthony C, "Forces at the Gate: Social Media's Influence on Editorial and Production Decisions in Local Television Newsrooms," *Electronic News*, 2016, *10* (2), 87–104.

Alaoui, Larbi and Fabrizio Germano, "Time scarcity and the market for news," *Journal of Economic Behavior and 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 Andrea Prat, "Is Journalistic Truth Dead? Measuring How Informed Voters Are about Political News," Working Paper 2021.

_ 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.

Aral, Sinan and Michael Zhao, "Social Media Sharing and Online News Consumption," Working Paper 2019.

Azzimonti, Marina and Marcos Fernandes, "Social Media Networks, Fake News, and Polarization," Working Paper 24462, National Bureau of Economic Research mar 2018.

Banerjee, Abhijit, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson, "The Diffusion of Microfinance," *Science*, 2013, *341* (6144), 1236498.

_ , _ , _ , and _ , "Gossip: Identifying Central Individuals in a Social Network," Working Paper 20422, National Bureau of Economic Research aug 2014.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

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.

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.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin, "Social Media and Mental Health," Technical Report 2022.

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.

Devereux, Michael B, Karine Gente, and Changhua Yu, "Production Networks and International Fiscal Spillovers," Working Paper 28149, National Bureau of Economic Research nov 2020.

Di Tella, Rafael, Ramiro H Gálvez, and Ernesto Schargrodsky, "Does Social Media cause Polarization? Evidence from access to Twitter Echo Chambers during the 2019 Argentine Presidential Debate," Working Paper 29458, National Bureau of Economic Research nov 2021.

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*

Electronic copy available at: https://ssrn.com/abstract=3663899

Journal: Applied Economics, 2018, *10* (1), 150–174.

Ershov, Daniel and Juan S. Morales, "Sharing News Left and Right," techreport 2021.

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.

Fletcher, Richard, Nic Newman, and Anne Schulz, "A Mile Wide, an Inch Deep: Online News and Media Use in the 2019 UK General Election," Digital News Project 2020 2020.

Fujiwara, Thomas, Karsten Muller, and Carlo Schwarz, "The Effect of Social Media on Elections: Evidence from the United States," Technical Report 2021.

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 of Conflicts," CEPR Discussion Papers 16167, C.E.P.R. Discussion Papers may 2021.

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.

Kennedy, Patrick J and Andrea Prat, "Where Do People Get Their News?," *Economic Policy*, 2019, *34* (97), 5–47.

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.

Electronic copy available at: https://ssrn.com/abstract=3663899

Levy, Ro'ee, "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment," *American Economic Review*, mar 2021, *111* (3), 831–870.

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*.

Martin, Gregory J. and Joshua McCrain, "Local News and National Politics," *American Political Science Review*, 2019, *113* (2), 372–384.

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.

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.

Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web.," Technical Report 1999-66, Stanford InfoLab nov 1999.

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.

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

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.

Electronic copy available at: https://ssrn.com/abstract=3663899

Weaver, David H, Lars Willnat, and G Cleveland Wilhoit, "The American Journalist in the Digital Age: Another Look at U.S. News People," *Journalism & Mass Communication Quarterly*, 2019, *96* (1), 101–130.

Xu, Ying and Jennifer Corbett, "Using Network Method to Measure Financial Interconnection," Working Paper 26499, National Bureau of Economic Research nov 2019.

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

Electronic copy available at: https://ssrn.com/abstract=3663899

Online Appendix to the Paper: Social Media Influence Mainstream Media: Evidence from Two Billion Tweets

Julia Cagé*, Nicolas Hervé, and Béatrice Mazoyer

June 21, 2022

Contents

A Data sources **2**
A.1 Tweet collection: Additional details . . . 2
A.2 News media data . . . 4
A.2.1 News media content data . . . 4
A.2.2 Additional characteristics of the news media . . . 10
A.3 IPTC topics . . . 11
A.4 Topic detection algorithm . . . 13

B Algorithms **15**
B.1 Mainstream media event detection algorithm . . . 15
B.2 Social media event detection . . . 16
B.3 Joint events . . . 18
B.3.1 First Story Detection . . . 18
B.3.2 Event-similarity graph . . . 18
B.3.3 Community detection . . . 21
B.3.4 Performance of the joint events detection algorithm . . . 21

C Additional tables **25**

D Additional figures **48**

E Robustness checks **56**

*Corresponding author. julia [dot] cage [at] sciencespo [dot] fr.

Electronic copy available at: https://ssrn.com/abstract=3663899

A Data sources

A.1 Tweet collection: Additional details

Choice of the keywords parameters As described in Section 2.1.1 the choice of "neutral words" to query tweets was made with a view to optimizing two metrics: the number of collected tweets, and their representativity of actual Twitter activity. The selected terms thus had to be the most frequently written words on Twitter, and we had to use different terms (and terms that do not co-occur in the same tweets) as parameters for each connection. In this way, the multiple connections would return sets of tweets with little intersection, and thus a greater total volume. Figure A.1 details the experiments conducted to select the keywords and optimize the distribution of these words on the different connections.

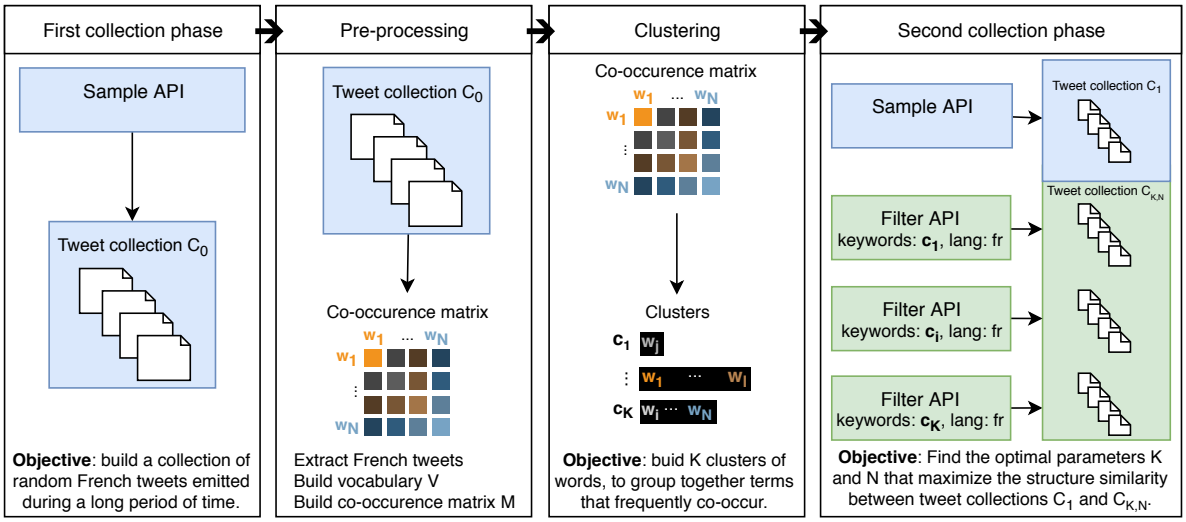
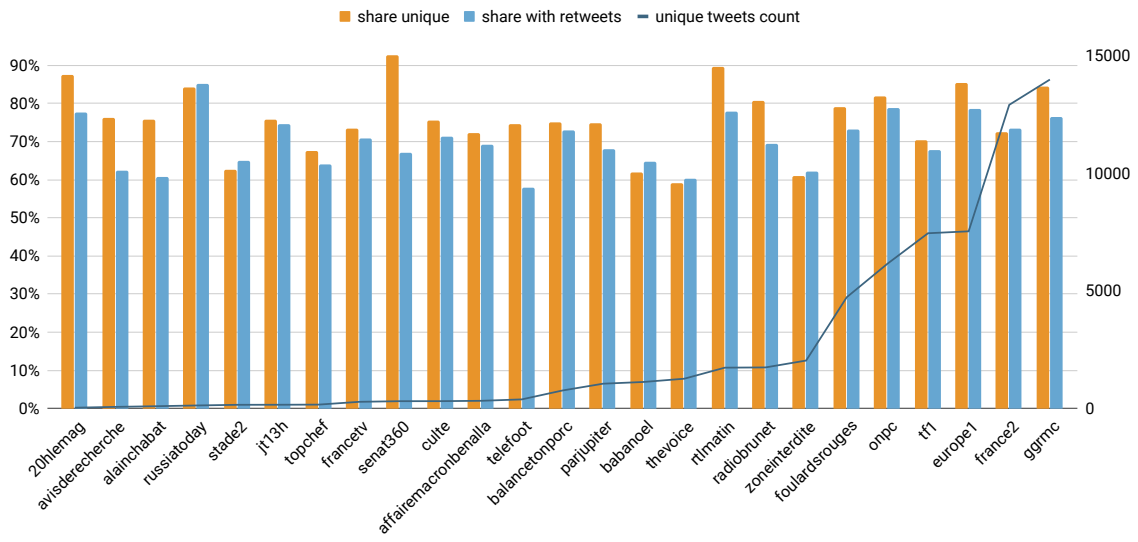


Figure A.1: Diagram of our experimental setup to select the best tweet collection method

Share of collected tweets As described in the core of the article, we use three different methods to evaluate the share of tweets we have collected. These evaluation methods are briefly presented in Section 2.1.1. Here, we provide more details on the methods based on the number of tweets per user, and report the associated Figures A.2 and A.3

In order to select users that write mostly in French (tweets written in other languages are not collected with our method), we used the OpenStreetMap API to locate users depending on what they indicate in the "location" field. We obtained 920,000 users located in France that emitted 241 million tweets, according to the "number of tweets" field. With our collection method, we captured 147 million tweets from these users, *i.e.* 61% of the real number of tweets emitted. We found the same percentage with the sample of users who geolocate their tweets in France (27,000 users). This method gives us a high estimate of the real number of tweets emitted in French, since some of these users probably write in other languages than

Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: This figure plots the share of tweets from the DLWeb that we were also able to capture using our collection method. Blue columns represent the ratio for all tweets, yellow columns represent the ratio for original tweets (*i.e.* retweets excluded). The grey line shows the number of original tweets (*i.e.* retweets excluded) captured by the DLWeb for each hashtag. Tweets were collected from December 1st to December 31st, 2018.

Figure A.2: Share of DLWeb tweets captured using our collection method for a set of 25 hashtags

French, even if they are located in France.

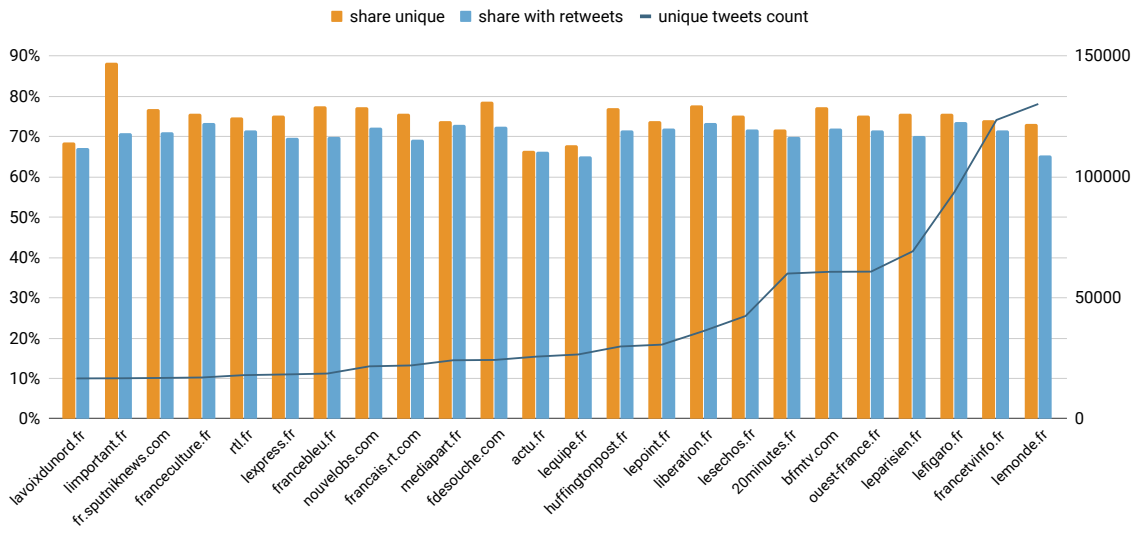
List of stop words To compute the average number of words included in the tweets, we have first removed the stop words listed in Figure A.4.

Sources excluded from our dataset We excluded sources explicitly described as bots, or referring to gaming or pornographic websites, from our tweet collection.

Our filtering rules are as follows. First, we use the "source" label provided by Twitter for each tweet.[1] 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 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

[1]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 don't 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."*

Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: This figure plots the share of tweets from the Médialab that we were also able to capture using our collection method for the first 25 domain names in terms of original tweets in their dataset. Blue columns represent the ratio for all tweets, yellow columns represent the ratio for original tweets (*i.e.* retweets excluded). The grey line shows the number of original tweets (*i.e.* retweets excluded) captured by the Médialab for each domain. Tweets were collected from December 1st to December 31st, 2018.

Figure A.3: Share of tweets from the Médialab also captured using our collection method for 25 domain names

video on Youtube: [url]". The entire list of the excluded sources is presented in Table A.1.

Second, we filter the users depending on their activity on the network: we only keep users with fewer than 1,000 tweets a day[2], and the users who have at least 1 follower. Finally, we only keep the users who post at least 12 tweets in French between August 2018 and July 2019.[3]

A.2 News media data

A.2.1 News media content data

The content data is from the OTMedia research projet. This projet was subsidized by the *Agence Nationale de la Recherche* (ANR – National Agency for Research), a French institution tasked with funding scientific research. The INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French radio and television audiovisual archives) was the project leader. The OTMedia research projet used the RSS feeds of the media outlets to track every piece of content they produced online. For the media outlets whose RSS feeds were not tracked by INA, we completed the OTMedia data by scrapping the Sitemaps of their

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

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

Electronic copy available at: https://ssrn.com/abstract=3663899

```
STOP_WORDS_FR = ['0', '1', '2', '3', 'a', 'ah', 'ai', 'aime', 'aller', 'alors', 'ans', 'apres', 'après', 'as', 'au',
    'aussi', 'autre', 'autres', 'aux', 'avais', 'avait', 'avant', 'avec', 'avez', 'avoir', 'b', 'bah', 'bcp',
    'beaucoup', 'bien', 'bon', 'bonjour', 'bonne', 'bref', 'c', "c'est", "c'était", 'ca', 'ce', 'cela',
    'celle', 'celui', 'ces', 'cest', 'cet', 'cetait', 'cette', 'ceux', 'chaque', 'chez', 'co', 'comme',
    'comment', 'compte', 'contre', 'coup', 'cours', 'crois', 'cétait', 'c'est', 'd', 'dans', 'de', 'deja',
    'depuis', 'des', 'detre', 'deux', 'dire', 'dis', 'dit', 'dm', 'dois', 'doit', 'donc', 'du', 'déjà',
    'dêtre', 'e', 'eh', 'elle', 'elles', 'en', 'encore', 'entre', 'envie', 'es', 'est', 'estce', 'et', 'etais', 'etait',
    'etc', 'ete', 'etes', 'etre', 'eu', 'f', 'faire', 'fais', 'fait', 'faites', 'faut', 'fois', 'font', 'g',
    'genre', 'gens', 'grave', 'gros', 'gt', 'h', 'hein', 'https', 'i', 'il', 'ils', 'j', "j'ai", "j'aime",
    "j'avais", "j'me", "j'suis", "j'vais", 'jai', 'jaime', 'jamais', 'javais', 'je', 'jen', 'jme', 'jour',
    'journee', 'journée', 'jsp', 'jsuis', 'jte', 'juste', 'jvais', 'jveux', 'jetais', 'jétais', 'j'ai', 'k', 'l', 'la',
    'le', 'les', 'leur', 'leurs', 'lol', 'lui', 'là', 'm', 'ma', 'maintenant', 'mais', 'mal', 'mdr', 'mdrr',
    'mdrrr', 'mdrrrr', 'me', 'mec', 'meme', 'merci', 'merde', 'mes', 'met', 'mettre', 'mieux', 'mis', 'mm',
    'moi', 'moins', 'moment', 'mon', 'monde', 'mtn', 'même', 'n', 'na', 'nan', 'ne', 'nest', 'ni', 'nn',
    'non', 'nos', 'notre', 'nous', 'o', 'of', 'oh', 'ok', 'on', 'ont', 'ou', 'ouais', 'oui', 'où', 'p', 'par',
    'parce', 'parle', 'pas', 'passe', 'pcq', 'pense', 'personne', 'peu', 'peut', 'peutetre', 'peutêtre', 'peux',
    'plus', 'pour', 'pourquoi', 'pq', 'pr', 'prend', 'prendre', 'prends', 'pris', 'ptdr', 'ptdrrr', 'ptn',
    'pu', 'putain', 'q', 'qd', 'qu', "qu'il", "qu'on", 'quand', 'que', 'quel', 'quelle', 'quelque', 'quelques',
    'quelquun', 'qui', 'quil', 'quils', 'quoi', 'quon', 'r', 'rien', 'rt', 's', 'sa', 'sais', 'sait', 'sans',
    'se', 'sera', 'ses', 'sest', 'si', 'sil', 'soir', 'soit', 'son', 'sont', 'suis', 'super', 'sur', 't',
    'ta', 'tas', 'te', 'tellement', 'temps', 'tes', 'tete', 'the', 'tjrs', 'tjs', 'toi', 'ton', 'toujours',
    'tous', 'tout', 'toute', 'toutes', 'tres', 'trop', 'trouve', 'trouvé', 'très', 'tt', 'tu', 'tête', 'u',
    'un', 'une', 'v', 'va', 'vais', 'vas', 'veut', 'veux', 'via', 'vie', 'viens', 'voila', 'voilà', 'voir',
    'vois', 'voit', 'vont', 'vos', 'votre', 'vous', 'vrai', 'vraiment', 'vs', 'vu', 'w', 'wsh', 'x', 'xd',
    'y', 'ya', 'z', 'à', 'ça', 'ça', 'étais', 'était', 'été', 'êtes', 'être', '——', '-', '']
```

Notes: The figure reports the list of stop words we use.

Figure A.4: List of stop words

website. Finally, we obtained all the AFP dispatches (respectively all the Reuters dispatches in French) directly from the AFP (from Reuters).

Our dataset includes the following media outlets:

Local daily newspapers:

1. *L'Ardennais*;
2. *Aisne Nouvelle*;
3. *Le Berry Republicain*;
4. *La Charente Libre*;
5. *Corse Matin*;
6. *Le Courrier Picard*;
7. *La Depeche Du Midi*;
8. *Est Eclair*;
9. *L'Eveil De La Haute Loire*;
10. *L'Independant Pyrenees Orientales*;
11. *Le Midi Libre*;
12. *La Montagne*;
13. *Nice Matin*;

Electronic copy available at: https://ssrn.com/abstract=3663899

14. *La Nouvelle Republique Des Pyrenees*;
15. *La Nouvelle Republique Du Centre Ouest*;
16. *Paris Normandie*;
17. *Le Parisien*;
18. *Le Petit Bleu D'Agen*;
19. *La Provence*;
20. *La Republique Des Pyrenees*;
21. *Sud Ouest*;
22. *Le Telegramme*;
23. *L' Union*;
24. *Var Matin*;
25. *La Voix Du Nord*;
26. *Yonne Republicaine.*

National daily newspapers:

1. *La Croix*;
2. *Les Echos*;
3. *L'Equipe*;
4. *Le Figaro*;
5. *France Soir*;
6. *La Gazette Des Communes Des Departements Et Des Regions*;
7. *L'Humanite*;
8. *Liberation*;
9. *Le Monde*;
10. *La Tribune.*

Free (national daily) newspapers:

1. *20 Minutes.*

Weekly (national & local) newspapers:

1. *10 Sport*;
2. *Agefi*;
3. *L'Avenir De Artois* (local);
4. *Capital*;
5. *Challenges*;
6. *Closer*;
7. *Courrier International*;
8. *L'Echo De La Lys* (local);
9. *Echo Le Valentinois Drome Ardeche* (local);
10. *Elle*;
11. *L'Essor Savoyard* (local);
12. *L'Express*;
13. *Femme Actuelle*;
14. *Gala*;
15. *Grazia*;

Electronic copy available at: https://ssrn.com/abstract=3663899

16. *L'Impartial De La Drome* (local);
17. *Les Inrockuptibles*;
18. *Investir*;
19. *Jeune Afrique*;
20. *Le Journal De Millau* (local);
21. *Le Journal Du Dimanche*;
22. *La Lettre De L Expansion*;
23. *L'Hebdo Du Vendredi*;
24. *La Manche Libre* (local);
25. *Marianne*;
26. *Le Monde Diplomatique*;
27. *Le Nouvel Economiste*;
28. *L'Obs*;
29. *L'Observateur De Beauvais* (local);
30. *Paris Match*;
31. *Le Paysan Du Haut Rhin* (local);
32. *Le Point*;
33. *Point De Vue*;
34. *Le Republicain De L'Essonne* (local);
35. *Le Réveil De Berck* (local);
36. *La Semaine Dans Le Boulonnais* (local);
37. *La Semaine Des Pyrénées* (local);
38. *Strategies*;
39. *Tele 7 Jours*;
40. *L'Usine Nouvelle*;
41. *Valeurs Actuelles*;
42. *Version Femina*;
43. *Voici*;
44. *La Volonte Paysanne De L'Aveyron* (local).

Monthly (national) newspapers:

1. *Alternatives Economiques*;
2. *Causeur*;
3. *Cb News*;
4. *Geo*;
5. *GQ Magazine*;
6. *Japon Infos*;
7. *Marie Claire*;
8. *Marie France*;
9. *Mon Viti*;
10. *Premiere*;
11. *La Revue Des Deux Mondes*;
12. *Santé Magazine*;
13. *Science Et Vie*;
14. *Sciences Et Avenir*;
15. *Sciences Humaines*;
16. *Têtu*;
17. *Vanity Fair*;
18. *Vogue*;
19. *Zibeline*.

Electronic copy available at: https://ssrn.com/abstract=3663899

TV:

1. BFM TV;
2. Eurosport;
3. France 24;
4. LCI;
5. Public Senat;
6. TF1;
7. TV5 Monde.

Radio:

1. Europe 1;
2. France Bleu (Radio France);
3. France Culture (Radio France);
4. France Inter (Radio France);
5. France Musique (Radio France);
6. France Info (also TV);
7. Radio Classique;
8. RCF;
9. RFI;
10. RTL;
11. Tendance Ouest.

News agencies:

1. Agence France Presse;
2. Reuters.

Pure online media:

1. 01 Net;
2. Acrimed;
3. Actu;
4. Actu Mag;
5. Aleteia;
6. Basta;
7. Boursier Com;
8. Boursorama;
9. Bref Eco;
10. Buzzfeed;
11. C Net;
12. Contrepoints;
13. Le Courrier Des Balkans;
14. Dreuz Info;
15. Les Echos Du Touquet;
16. Echos Start;
17. Echosdunet;
18. Foot Mercato;
19. Football;

Electronic copy available at: https://ssrn.com/abstract=3663899

20. Ginjfo;
21. Goodplanet Info;
22. Herault Tribune;
23. Huffington Post;
24. Influenth;
25. Jeune Nation;
26. Le Journal Du Net;
27. Le Journal Des Femmes;
28. Konbini;
29. L'ADN;
30. Le Libre Penseur;
31. Le Media;
32. Le Tribunal Du Net;
33. L'Explicite;
34. L'Incorrect;
35. L'Internaute;
36. Là Bas Si J'y Suis;
37. LVSL;
38. Maddyness;
39. Made In Foot;
40. Mag Centre:
41. Marsactu;
42. Mashable;
43. Mediacites;
44. Medialot;
45. Mediapart;
46. Meta Media;
47. Medias Presse Info;
48. Minutenews;
49. Mondafrique;
50. Les Moutons Enrages;
51. Next Inpact;
52. Numerama;
53. Ohmymag;
54. Paris Depeches;
55. Le Petit Journal;
56. Pourquoi Docteur;
57. Pure Medias;
58. Purepeople;
59. Resistance Republicaine;
60. Rue 89 Lyon;
61. Rue89 Bordeaux;
62. Rue89 Strasbourg;
63. Slate;
64. Sputniknews;
65. Toulouse 7;
66. Toute La Culture;
67. Up Magazine;
68. L'Usine Digitale.

Electronic copy available at: https://ssrn.com/abstract=3663899

French-speaking foreign media

1. *20 Minutes Suisse* (Switzerland);
2. 24h Au Benin;
3. Africa Intelligence;
4. Express Mu Ile Maurice;
5. Jeune Indépendant Algérie;
6. Journal De Montreal Canada;
7. Journal De Quebec Canada;
8. Nouvelles Caledoniennes;
9. Nouvelle Tribune Benin;
10. Petit Journal Turquie;
11. Quotidien Canada (Canada);
12. Slate Afrique;
13. *Temps Suisse* (Switzerland);
14. Wort Luxembourg.

A.2.2 Additional characteristics of the news media

Media and website creation date The creation date for websites was either found through a credible source (the media itself, a news article announcing the creation of the website, etc.) or, absent such a source, through a proxy. The first proxy we used is the first appearance of the website on the Internet Archive. The date was only recorded when the page registered on the Archive was usable (i.e. it opened, did not display a blank page) and was a page for an active media outlet (a page that was under construction, even if it was for the outlet, was not counted as an appearance). However, this proxy does not give the date of creation of the website and usually the date of the first appearance on the Archive is later than the actual opening of the website, as the site has to receive some attention from online users before it is registered by a web crawler and then in the Archive.

The second proxy we built is the date of the registration of the domain name for the website, obtained through a WHOIS Lookup. The date inputted is always that of the domain in the first Internet Archive registration. In some cases, the date came after the first Archive appearance, which signifies the domain had been re-registered, in which case the data was marked as missing. In addition, if the domain was registered before the first Archive appearance but it was clear from looking at the archive that the domain was registered for a website other than that of the media outlet, the data was marked as missing. This proxy gives an indication of the earliest possible time at which the media outlet created its website, without accounting for the website's construction time. Using the two proxies when no precise date is available can give a good idea of when the website was created, most certainly in between the two dates, and maybe slightly closer to the Internet Archive date than to the WHOIS date.

Business model The business model was found by consulting the website of each media outlet. If no article can be read without paying a subscription fee then the paywall is coded

Electronic copy available at: https://ssrn.com/abstract=3663899

as "hard". Otherwise, unless there is unlimited free access, the paywall is coded as "soft". Soft paywalls either limit which articles the user can access for free, the number of articles users can access, or the number of articles of a certain type that users can access. Next, a few websites used an advertisement-wall for their content, through which some articles were accessible to users if they watched an ad.

The paywall introduction date was found either through a credible source or by looking at the first appearance of any trace of the paywall on the Internet Archive. Service diversification (non-editorial services offered) was checked by looking for credible sources on non-editorial services a media could have in its revenue streams. The presence of advertising was determined by consulting the outlet's website.

Estimation of non-editorial revenues We tried to reconstruct the value of non-editorial revenue for each media outlet in an attempt to estimate the weight of advertising revenue in the outlet's total revenues. To do so, we collected the outlet's available accounting through Altares' Intuiz Platform, and then collected, for each year, the reported values of service-based revenues and total revenues. Each outlet has been linked with a unique legal identity, identified by its SIREN number. When multiple entities controlled the outlet, or different entities controlled different parts of the outlet, we considered all entities together.

Furthermore, for the **print media**, we used the *"Alliance pour les Chiffres de la Presse et des Medias"* (Alliance for Data on the Press and Media) data for the sales of digital issues and subscriptions for each outlet and year, as well as the Alliance reported price per issue and subscription price per year to compute an estimation of the total sales revenues. Subtracting this value from the total revenues gives us an (admittedly imperfect) estimation of the non-editorial (mostly advertising) revenues.

For **television**, we collected the advertising data reported in the *"Centre National du Cinéma"* (National Center for Cinema) yearly television station guides.

For radio, we relied on press reports of advertising revenue (Les Echos, Nicolas Madelaine) and Radio France's own website to estimate the advertising revenues of Radio France. For private radio stations, no data is publicly available in a systematic way on the share of advertising revenues in total revenues. Similarly, we had to proceed on a case-by-base basis for the **pure online media**.

A.3 IPTC topics

To define the subject of its dispatches, AFP uses URI, available as QCodes, designing 17 IPTC media topics. The IPTC is the International Press Telecommunications Council.

The 17 topics are defined as follows:

Electronic copy available at: https://ssrn.com/abstract=3663899

- **Arts, culture and entertainment**: matters pertaining to the advancement and refinement of the human mind, of interests, skills, tastes and emotions.
- **Crime, law and justice**: establishment and/or statement of the rules of behaviour in society, the enforcement of these rules, breaches of the rules and the punishment of offenders. Organisations and bodies involved in these activities.
- **Disaster and accident**: man-made and natural events resulting in loss of life or injury to living creatures and/or damage to inanimate objects or property.
- **Economy, business and finance**: all matters concerning the planning, production and exchange of wealth.
- **Education**: all aspects of furthering the knowledge of human individuals from birth to death.
- **Environment**: all aspects of the protection, damage, and condition of the ecosystem of the planet earth and its surroundings.
- **Health**: all aspects pertaining to the physical and mental welfare of human beings.
- **Human interest**: items concerning individuals, groups, animals, plants or other objects with a focus on emotional facets.
- **Labour**: social aspects, organisations, rules and conditions affecting the employment of human effort for the generation of wealth or provision of services and the economic support of the unemployed.
- **Lifestyle and leisure**: activities undertaken for pleasure, relaxation or recreation outside paid employment, including eating and travel.
- **Politics**: local, regional, national and international exercise of power, or struggle for power, and the relationships between governing bodies and states.
- **Religion and belief**: all aspects of human existence involving theology, philosophy, ethics and spirituality.
- **Science and technology**: all aspects pertaining to human understanding of nature and the physical world and the development and application of this knowledge.
- **Society**: aspects of the life of humans with regard to their social relationships.
- **Sport**: competitive exercise involving physical effort. Organizations and bodies involved in these activities.

Electronic copy available at: https://ssrn.com/abstract=3663899

- **Conflicts, war and peace**: acts of socially or politically motivated protest and/or violence and actions to end them.
- **Weather**: the study, reporting and prediction of meteorological phenomena.

A.4 Topic detection algorithm

As described in Section 2.4, we train a classifier to tag each news article with the 17 top-level IPTC categories. Our training dataset is composed of the AFP dispatches for the years 2012 to 2017, i.e. the 6 years preceding our corpus (around 1.5 million dispatches). For each of these dispatches, we keep only the 17 top-level IPTC categories, whether they have been directly assigned by the AFP or inferred from the sub IPTC categories reported in the dispatch's metadata. Each article can thus be included in several categories (e.g. talks both about politics and international affairs).

For a good overview of text categorization methods, see Aggarwal (2018). In this article, we are in a multilabel text classification setting. We choose a neural network approach with TF-IDF encoding of the documents. The learning of the classifier consists of two steps. First, a vocabulary is learned from the dataset. This vocabulary will then be used to obtain the vector representation of the documents. We consider single words as well as bi-grams. For this feature selection step, the vocabulary is learned by keeping the $15,000$ words that have the highest Chi-square scores with respect to the 17 categories. We therefore have the most discriminating terms for our labeling task. The vector representation of the documents is based on a standard TF-IDF weighting.

Next, a simple fully-connected non-linear neural network is trained for 10 epochs (input layer of dimension $15,000$, single hidden layer of dimension 128 with tanh activation function and dropout, output layer with 17 dimensions). We use a multi-label loss provided by the Pytorch framework. This approach is standard and is described in a guide provided by Google. Once trained, the network is able to predict for any news article a score between 0 and 1 for each of the IPTC categories.

Electronic copy available at: https://ssrn.com/abstract=3663899

BotDuCul	http://louphole.com
BotGentil	http://www.benjamingibeaux.fr
Botbird tweets	https://slmame.com
Botpoto	https://anthony-dumas.fr
CatenaBot	http://vnatrc.net
CeQuiNeTeBotPas	http://louphole.com/bots/
Cheap Bots, Done Quick!	http://cheapbotsdonequick.com
Comparobot	http://louphole.com/
Curious Cat	https://curiouscat.me
EmergencyPipou0	http://louphole.com
Gamepush2	http://gamepush.fr
Games Trailers	http://gamestrailers.org
Google	https://www.google.com/
Google	http://www.google.com/
H. M. Despladt's BOT	http://vnatrc.net/
JeuDuDicoBot	http://louphole.com
LVRFD Bot	http://twitter.com/LVRFD_Bot
LaCoFD Twitter Bot	https://www.fire.lacounty.gov/
MNBot	http://moringanutrition.fr
ManageTweetBot	https://twitter.com/EasterEd35
MonChatBot	https://mon-chatbot.com
MypornSight autopublish	http://www.mypornsights.com
ONE PIECE TREASURE CRUISE	http://www.bandaigames.channel.or.jp/list/one_...
Paradise Island 2	http://www.game-insight.com/
PsychoAFALISTOBOT	http://www.vnatrc.net/
Radio King LiveTweet	https://www.radioking.com
Random Taxi bot	https://whatever.com
RoboTribz	http://jhroy.ca
Temperature Bot MC901	http://www.notyet.com/
Tweetbot for Mac	https://tapbots.com/software/tweetbot/mac
Tweetbot for iOS	http://tapbots.com/tweetbot
Unfollow.fr	http://www.unfollow.fr/
WizeBot.tv	https://wizebot.tv
bondageartbot_s3	http://121.170.193.209/muse
dtc randposts	http://vnatrc.net/
emploisjob	http://emploisjob.com/
glissantBot	http://www.villaempain.com/en/
gnapblbot	http://vnatrc.net/
lapresse_diff	http://ruebot.net
manuuuuu	https://curiouscat.me/Saphirewall
myfirsttwitbotfabien	http://www.cookngo.paris
porc_bot	https://github.com/clemonster
pyTweetInfoBot	http://www.nilsschaetti.com/index.php/projects...
rakubo2	https://rakubots.kissa.jp/
test essai bot	http://twitter.com/LucieWankel
twitbilbot_	http://bilboeee.fr
twittbot.net	http://twittbot.net/
vnatrcASCIIBOT	http://www.vnatrc.net/

Table A.1: List of the "sources" labels excluded from our dataset

Electronic copy available at: https://ssrn.com/abstract=3663899

B Algorithms

B.1 Mainstream media event detection algorithm

Description of the algorithm The goal of online topic detection is to organize a continuous incoming stream of news articles by the events they discuss. The algorithms place all the documents into appropriate and coherent clusters. Consistency is ensured both at the temporal and the semantic levels. As a result, each cluster provided by the algorithm covers the same topic (event) and only that topic. Following Allan et al. (2005), who have experienced their TDT system in a real world situation, we adopt the following implementation:

1. As in most natural language processing methods, we first pre-process our documents by removing very common words (called stop words) and applying a stemming algorithm so as to keep only the stem of the words.

2. Each document is then described by a semantic vector which takes into account both the headline and the text. We apply a multiplicative factor of five to the words of the title as they are supposed to describe the event well, resulting in an overweight in the global vector describing the document. A semantic vector represents the relative importance of each word of the document compared to the full dataset. A standard scheme is TF-IDF: term frequency-inverse document frequency, a numerical statistic intended to reflect how important a word is to a document in a corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. More precisely, the weight of a word w in a document d is: $TFIDF(w,d) = wf\,(w,d) * log\,(N/df\,(w))$, where wf is the frequency of word w in the considered document, dw is the number of documents in which it appears, and N is the total number of documents. The total vector is $TFIDF(d) = [TFIDF(w_1,d), TFIDF(w_2,d), ..., TFIDF(w_n,d)]$, where $w_1, ..., w_n$ are the words occurring in the whole text stream to segment.

3. The documents are then clustered in a bottom-up fashion to form the events based on their semantic similarity. The similarity between two documents is given by the distance between their two semantic vectors. We use the cosine similarity measure (Salton et al., 1975).

4. This iterative agglomerative clustering algorithm is stopped when the distance between documents reaches a given threshold. We have determined this threshold empirically based on manually created media events.

5. A cluster is finalized if it does not receive any new document for a given period of time.

Electronic copy available at: https://ssrn.com/abstract=3663899

We use a one-day window.[4]

Performance of the algorithm This event detection algorithm can be compared to other detection systems in its ability to put all the stories in a single event together. We test the quality of the algorithm by running it on a standard benchmark dataset: the Topic Detection and Tracking (TDT) Pilot Study Corpus. The TDT dataset contains events that have been created "manually": the goal is to compare the performance of the algorithm with that of humans. The goal of the TDT initiative is to investigate the state of the art in finding and following events in a stream of news stories (see e.g. Allan et al., 1998). To test the performance of our algorithm on the English corpus, we slightly adapt it. There is indeed no similar test corpus in French. First, we use an English stop-word list and an English stemming algorithm. Second, given that the time frame of the test corpus is wider than ours, the one-day window used to close clusters is not adapted. When testing our algorithm we thus follow the literature (Allan et al., 2005) and close a cluster when $2,500$ documents have been treated by the algorithm and none of them has been added to the cluster.

In the TDT Pilot Study Allan et al. (1998), two types of algorithms are evaluated: a "retrospective algorithm" and an "online algorithm". A retrospective algorithm needs to know all the articles in order to detect media events, whereas an online algorithm is fed by the stream of articles, one by one. Given that the OTMedia platform must be able to manage articles in real time, we implemented an online algorithm.

Note also that we find that the main parameter of our implementation, the distance threshold on semantic similarity, is the same for this English test corpus and our corpus of French news articles. While we were expecting these thresholds to be of similar order of magnitude (the TF-IDF representation of text is only based on word appearance frequencies; given that both corpuses include articles that are of the same nature, it is not surprising to obtain relatively close thresholds for French and English), finding a similar threshold is nonetheless very reassuring as to the quality of our algorithm. In particular, it ensures that the news events detected by our algorithm are as close as possible to what a human would be able to do.

B.2 Social media event detection

Description of the algorithm We use the same algorithm (which we call "First Story Detection") for event detection in tweets as the one used for mainstream media event detection, with some minor changes in the preprocessing step (see Table B.1). This algorithm was chosen after we tested its superiority over topic modelling techniques such as Dirichlet

[4]Events can last more than one day. But if during a 24-hour period of time no document is placed within the cluster, then the cluster is closed. Any new document published after this time interval becomes the seed of a new event cluster.

Electronic copy available at: https://ssrn.com/abstract=3663899

Multinomial Mixture model (Yin and Wang, 2014), or a standard clustering algorithm such as DBSCAN (Ester et al., 1996). This superiority 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.

The semantic vectors used to represent tweets also use the TF-IDF weighting scheme, which performs better on our data than more recent embedding models such as Word2Vec, ELMo, Universal Sentence Encoder or Sentence BERT.

Performance of the algorithm The relative performance of the different models is evaluated using two datasets: the corpus by McMinn et al. (2013), as well as a corpus of 38 million original tweets collected from July 15th to August 6th 2018 that we manually annotated.[5] Figure B.1 reports the performance of different embeddings used with the FSD algorithm, depending on the distance threshold parameter t. 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.

Preprocessing Each text embedding model takes different text formats as inputs: for example, models able to deal with sentences, such as BERT, Sentence-BERT, ELMo or Universal Sentence Encoder, take the full text with punctuation as input. For Word2Vec and TF-IDF models, we lowercase characters and remove punctuation. Table B.1 summarizes all preprocessing steps depending on the type of model. Each column corresponds to a preprocessing step:

- Remove mentions: mentions are a Twitter-specific way of referring to another Twitter user in a tweet, so that she is notified that the tweet is talking about her or is addressed to her. Entries take the following form: @name_of_the_user. For TF-IDF models, removing mentions is a way to reduce the size of the vocabulary. For most Word2Vec models, mentions are not part of the vocabulary, except for w2v_twitter_en.
- Unidecode: we use the Python module unidecode to convert Unicode characters to ASCII characters. In French, for example, all accents are removed: "Wikipédia" becomes "Wikipedia".

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

Electronic copy available at: https://ssrn.com/abstract=3663899

- Lower: we set the text in lowercase letters.
- Hashtag split: we split hashtags on capital letters. "#HappyEaster" becomes "Happy Easter". This step is of course applied before lowercasing.
- Remove long numbers: we remove numbers longer than 4 digits.
- Remove repeated characters: we limit the number of repeated characters inside a word to three. "looooool" becomes "loool".

B.3 Joint events

Our approach can be broken down into three steps: first, we perform the detection of Twitter events and media events separately. Then we represent the similarity between detected events in a weighted bi-partite graph. Finally, we apply a community detection algorithm in order to discover common events across the two spheres. Figure B.3 represents the approach used for the last two steps.

B.3.1 First Story Detection

Tweets and news articles are quite different in terms of length and type of vocabulary. Detecting joint events directly from a heterogeneous collection of documents may thus lead to a poor performance. In order to let Twitter-specific and media-specific clusters emerge, we thus perform a first event detection step separately for each type of document.

B.3.2 Event-similarity graph

Once events are detected separately in each sphere, we model the relationships between Twitter events and media events as a weighted bi-partite graph. In the rest of the Section, we denote $E_T = \{e_{T,1}, \ldots, e_{T,f}\}$ the set of all Twitter events, and $E_M = \{e_{M,1}, \ldots, e_{M,g}\}$ the set of all media events. A Twitter event is composed of a set of tweets, and a media event is composed of a set of news articles. We explore three types of links between Twitter events and media events: word-similarity, URLs and hashtags.

Word-similarity. In order to compute a word-similarity metric between Twitter events and media events, we represent each event as the average of the TF-IDF vectors of all documents it contains. The vocabulary used to compute TF-IDF is the union of the two vocabularies (vocabulary of tweets and vocabulary of news). The word-similarity between two events is computed as the cosine similarity between these average vectors and used to weight the edge between the two events:

$$weight_{text}(e_T, e_M) = \frac{\vec{e_T} \cdot \vec{e_M}}{\|\vec{e_T}\| \|\vec{e_M}\|} \quad (1)$$

Electronic copy available at: https://ssrn.com/abstract=3663899

Electronic copy available at: https://ssrn.com/abstract=3663899

model	rm mentions	unidecode	lower	rm punctuation	hashtag split	rm long numbers	rm repeated chars	rm urls
tfidf_all_tweets	X	X	X	X	X	X	X	X
tfidf_dataset	X	X	X	X	X	X	X	X
w2v_afp_fr	X	X	X	X	X	X	X	X
w2v_twitter_fr	X	X	X	X	X	X	X	X
w2v_gnews_en	X			X	X	X	X	X
w2v_twitter_en				X	X	X	X	X
elmo					X	X	X	X
bert					X	X	X	X
bert_tweets					X	X	X	X
sbert_sts					X	X	X	X
sbert_nli_sts					X	X	X	X
sbert_tweets_sts					X	X	X	X
sbert_tweets_sts_long					X	X	X	X
use					X	X	X	X

Table B.1: Preprocessing applied for each model

Where $\vec{e}$ is the average of the TF-IDF vectors of all documents in e. To facilitate the community detection step (see Section B.3.3), we then remove the edges with a too weak cosine similarity. We denote s the similarity threshold.

Hashtags. The graph of hashtag relationships between Twitter events and media events is built as follows: if some of the tweets of a Twitter event and some of the articles of a media event have hashtags in common, we draw an edge between the two events. The hashtag weight is computed as follows:

$$weight_{htag}(e_T, e_M) = \frac{h(e_T, e_M)}{\max\{h(e_T, e_M) : e_T \in E_T, e_M \in E_M\}} \tag{2}$$

where $h(e_T, e_M)$ is the number of times the hashtags common to e_T and e_M appear in the Twitter event e_T. In order to limit the role of one individual hashtag (e.g. #BreakingNews) we remove edges where the number of different hashtags is too low.

URLs. The graph of URLs is constructed in the same way as the graph of hashtags: if tweets within a Twitter event e_T contain a URL pointing to one of the articles in media event e_M, we draw an edge between e_T and e_M. The weight of urls is computed as follows:

$$weight_{url}(e_T, e_M) = \frac{u(e_T, e_M)}{\max\{u(e_T, e_M) : e_T \in E_T, e_M \in E_M\}} \tag{3}$$

where $u(e_T, e_M)$ is the number of times urls linking to articles that are part of media event e_M appear in the tweets of Twitter event e_T.

Multidimensional graph. We combine these three different layers into a single multidimensional graph where the weight of the edges is computed as follows:

$$weight(e_T, e_M) = \sum_{i \in \{text, url, htag\}} \alpha_i weight_i(e_T, e_M) \tag{4}$$

where $0 \leq \alpha_i \leq 1$ and $\sum_{i \in \{text, url, htag\}} \alpha_i = 1$.

Time of events. In addition to including word similarity, hashtags and urls in the construction of the event similarity graph, we also take into account the time dimension of the events. We therefore introduce a final parameter, Δ, which indicates the maximum time difference (in days) between a pair of events. Figure B.2 shows examples of configurations where edges between events are kept, and others where edges are removed because the two events are too distant in time.

Electronic copy available at: https://ssrn.com/abstract=3663899

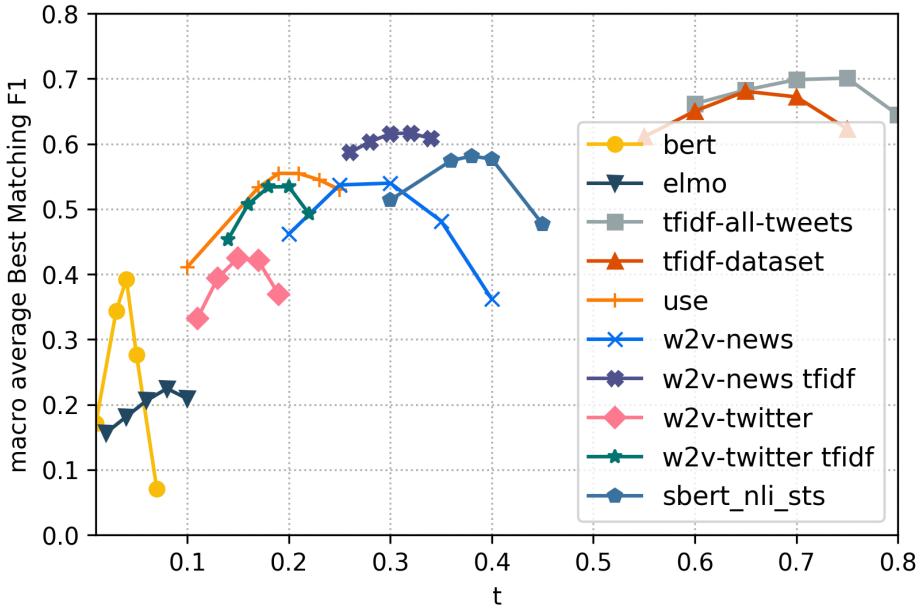
B.3.3 Community detection

Community detection within a network consists in decomposing the network into sub-groups of highly connected nodes. Researchers have proposed many strategies to solve this task, many of them based on the optimization of a given objective function. We use the implementation by Traag et al. (2015), which is a variant of the the Louvain algorithm (Blondel et al., 2008) with a different objective function (*Surprise* (Aldecoa and Marin, 2011) instead of *Modularity* (Newman and Girvan, 2004)) to find a partition of the nodes within the event-similarity graph.

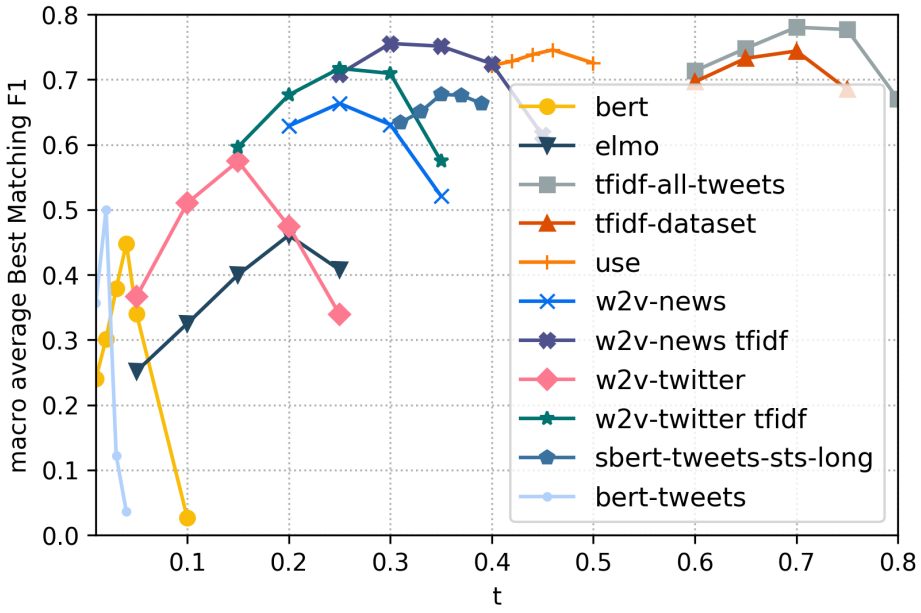
B.3.4 Performance of the joint events detection algorithm

We evaluate the performance of our algorithm using the "best matching" F1 score (Yang et al., 1998) for each event in the manually annotated dataset. The values of α_{text}, α_{url}, α_{htag}, Δ and s are set to maximize this metric. URLs and hashtags seem to have a much lower effect than text. Increasing α_{url} or α_{htag} may slightly improve the performance on some time periods but it also degrades it on other sub-samples of our dataset. Overall, there is no configuration of α_{url} and α_{htag} that performs equally well on each subset. We therefore chose to eliminate these modalities, in order to simplify the graph construction step and to obtain a model that is more stable to the change of dataset. The final choice of parameters is the following: $(\alpha_{text} = 1, \alpha_{url} = 0, \alpha_{htag} = 0, s = 0.3, \Delta = 1)$.

Electronic copy available at: https://ssrn.com/abstract=3663899



(a) **English**



(b) **French**

Figure B.1: Best Matching F1 score for FSD clustering depending on the distance threshold parameter t for each corpus

Electronic copy available at: https://ssrn.com/abstract=3663899

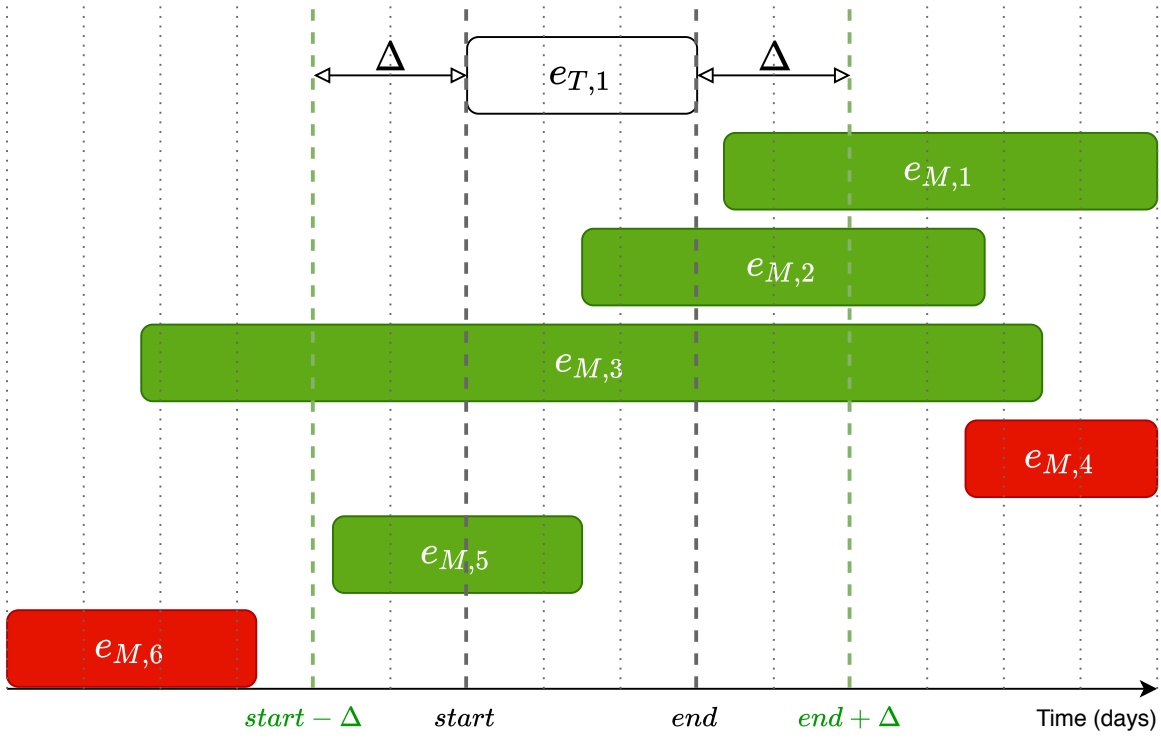
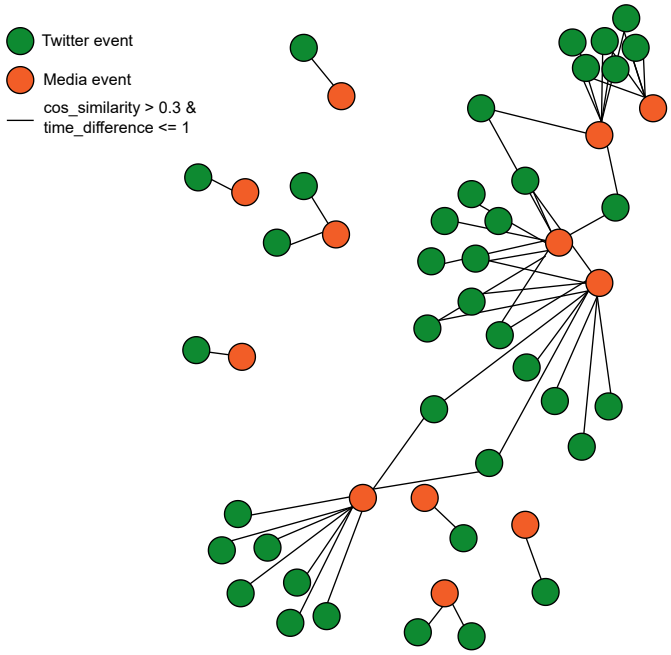
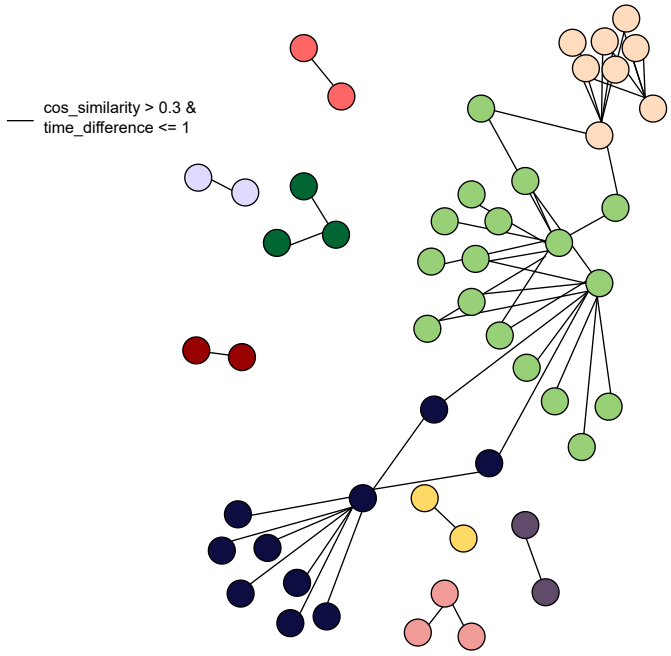


Figure B.2: Example of different time configurations between a given Twitter event and some media events. Media events ending before $start - \Delta$ or beginning after $end + \Delta$ appear in red. The edges between these red events and $e_{T,1}$ are removed in the graph.

Electronic copy available at: https://ssrn.com/abstract=3663899



(a) **Building the similarity network**



(b) **Applying Louvain algorithm**

Figure B.3: Graphical representation: Building the joint events

Electronic copy available at: https://ssrn.com/abstract=3663899

C Additional tables

Table C.1: Summary statistics: Twitter users (full sample; last time the user is observed)

	Mean	St.Dev	P25	Median	P75	Max
User activity						
Total number of tweets	16,067	43,680	241	1,801	12,977	4,783,226
Nb of tweets btw first & last time	388	1,611	19	44	178	202,104
Nb of tweets user has liked	10,957	29,300	193	1,326	8,161	3,613,810
Nb of users the account is following	666	4066	84	200	500	1568851
User identity						
Date of creation of the account	2,015.085	2.892	2,013	2,016	2,018	2,019
=1 if verified account	0.005	0.069	0	0	0	1
=1 if user is a journalist	0.0018	0.043	0	0	0	1
=1 if user is a media	0	0	0	0	0	1
User popularity						
Nb of followers	1,933	68,887	27	121	450	32,642,187
Nb of public lists	17	550	0	1	5	1,023,854
Observations	4,395,135					

Notes: The table gives summary statistics. Time period is August 2018 - July 2019. Variables are values for all the Twitter users included in our dataset the last time we observe them. Variables are described in more details in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.2: Summary statistics: Tweets (split-sample, August 1st 2018-November 30th 2018)

	Mean	St.Dev	P25	Median	P75	Max	Obs
Characteristics of the tweet							
Length of the tweet (nb of characters)	102	53	61	99	140	1,415	531,050,553
Number of words	6.2	4.1	3.0	6.0	9.0	234	531,050,553
=1 if the tweet is a retweet	0.63	0.48	0.00	1.00	1.00	1	531,050,553
=1 if the tweet is a reply	0.18	0.38	0.00	0.00	0.00	1	531,050,553
=1 if the tweet is a quote	0.20	0.40	0.00	0.00	0.00	1	531,050,553
Popularity of the tweet							
Number of retweets	2.2	121.1	0.00	0.00	0.00	436,776	198,166,184
Number of replies	0.2	10.7	0.00	0.00	0.00	64,980	198,166,184
Number of likes	3.7	230.2	0.00	0.00	0.00	1,210,203	198,166,184

Notes: The table gives summary statistics. Time period is August 1st 2018-November 30th 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.

Table C.3: Summary statistics: Media outlets (split-sample, August 1st 2018-November 30th 2018)

	Mean	St.Dev	P25	Median	P75	Max
Content						
Total content (thsd ch)	16,789	34,328	1,295	4,914	15,749	346,310
Total number of articles	7,168	14,808	496	2,150	6,026	127,563
Online audience (daily)						
Number of unique visitors	204,224	289,642	32,094	86,840	249,134	1,374,814
Number of visits	576,439	828,343	74,002	223,857	693,494	3,466,883
Number of pages views	1,429,691	2,315,493	174,359	475,453	1,648,408	14,029,836
Social media presence						
Number of Twitter accounts	2.9	5.0	1.0	1.0	2	37
Date of Twitter account creation	2010	2.2	2009	2009	2,011.0	2018
Number of tweets	1,912	2,732	300	882	2400	20,191
Nb journalists with Twitter account	76	136	7	24	88	1,067
Other media characteristics						
Year of media creation	1976	39	1946	1988	2,008	2021
Year of website creation	2004	6	1998	2004	2010	2021
Year of the paywall introduction	2014	5	2013	2015	2018	2020
Number of journalists	139	174	34	84	201	1121
Observations	200					

Notes: The table gives summary statistics. Time period is August 1st 2018-November 30th 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.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.4: Summary statistics: Mainstream media articles (split-sample, August 1st 2018-November 30th 2018)

	Mean	St.Dev	P25	Median	P75	Max
Length						
Length (number of characters)	2,343	2,141	958	1,837	3,117	126,800
Originality rate (%)	43	42	3	22	100	100
Facebook shares						
Number of shares on Facebook	26	572	0	0	0	173,380
Number of comments on Facebook	32	685	0	0	0	335,842
Number of reactions on Facebook	118	2,245	0	0	0	813,161
Observations	1,239,552					

Notes: The table gives summary statistics. Time period is August 1st 2018-November 30th 2018. Variables are values for the mainstream media articles. The observations are at the article level.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.5: Characteristics of the media consumers: Newspapers

(a) **Daily newspapers**

	20 Minutes	La Croix	Les Echos	Le Figaro	Liberation	Le Monde
	Mean	Mean	Mean	Mean	Mean	Mean
Age	48.5	45.1	48.2	47.6	46.4	45.4
=1 if male (%)	49.8	59.3	57.9	58.0	62.5	57.4
=1 if Bachelor degree or more (%)	39.0	47.9	53.4	46.5	53.5	53.6
=1 if Annual income $\geq$ €30 000 (%)	45.5	55.3	57.9	52.1	51.9	53.2
=1 if leaves in Paris region (%)	22.6	24.1	21.5	25.2	23.2	25.7
=1 if highly interested in news (%)	63.0	72.2	72.0	73.6	78.6	71.8
=1 if places itself on the Left (%)	34.3	48.1	40.2	30.0	57.1	47.5
=1 if uses Twitter (%)	23.2	27.8	29.9	27.6	38.4	30.3

(b) **Other newspapers**

	Courrier International	L'Express	Marianne	L'Obs	Le Point
	Mean	Mean	Mean	Mean	Mean
Age	41.7	49.4	51.1	49.4	51.4
=1 if male (%)	72.9	50.6	65.4	66.9	54.2
=1 if Bachelor degree or more (%)	45.8	40.6	42.9	47.7	43.9
=1 if Annual income $\geq$ €30 000 (%)	54.8	42.1	47.7	55.4	50.0
=1 if leaves in Paris region (%)	32.9	20.1	15.4	19.4	22.6
=1 if highly interested in news (%)	72.9	72.0	76.9	79.0	80.0
=1 if places itself on the Left (%)	55.7	38.4	51.9	56.5	40.0
=1 if uses Twitter (%)	38.6	29.9	30.8	35.5	26.5

Notes: The table gives summary statistics on the characteristics of the media consumers. Year is 2018. The observations are at the individual level. Data are from the *2018 Digital News Report* (Reuters Institute, 2018).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.6: Characteristics of the media consumers: Television, Radio, and Pure online media

(a) **TV & Radio**

	BFM TV	France 24	France Info	LCI	TF1	Private radio
	Mean	Mean	Mean	Mean	Mean	Mean
Age	49.3	46.7	52.0	52.0	46.1	47.0
=1 if male (%)	54.6	56.3	59.5	57.1	37.2	61.1
=1 if Bachelor degree or more (%)	36.6	44.6	38.4	26.2	26.4	31.7
=1 if Annual income $\geq$ €30 000 (%)	45.9	42.3	46.4	49.2	39.4	42.9
=1 if leaves in Paris region (%)	24.0	21.8	18.9	20.6	17.1	14.8
=1 if highly interested in news (%)	65.3	64.4	68.9	71.4	57.4	65.7
=1 if places itself on the Left (%)	28.6	41.4	50.0	42.9	31.8	38.0
=1 if uses Twitter (%)	27.0	34.5	25.8	33.3	29.5	25.9

(b) **Pure players**

	Buzzfeed	Huffington Post	L'Internaute	Mashable	Mediapart
	Mean	Mean	Mean	Mean	Mean
Age	32.8	44.5	46.9	34.4	46.9
=1 if male (%)	55.7	64.3	49.5	70.0	65.0
=1 if Bachelor degree or more (%)	45.8	59.2	43.6	43.8	48.0
=1 if Annual income $\geq$ €30 000 (%)	41.1	53.7	48.4	33.3	51.2
=1 if leaves in Paris region (%)	27.9	28.6	22.3	35.0	22.6
=1 if highly interested in news (%)	62.3	73.5	63.1	65.0	76.6
=1 if places itself on the Left (%)	60.7	54.6	50.5	40.0	69.3
=1 if uses Twitter (%)	36.1	33.7	28.2	45.0	33.6

Notes: The table gives summary statistics on the characteristics of the media consumers. Year is 2018. The observations are at the individual level. Data are from the *2018 Digital News Report* (Reuters Institute, 2018).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.7: Validity of the instrument: Topic of the event

(a)

	(1) Crime	(2) Politics	(3) Conflicts	(4) Economy	(5) Labour	(6) Sport	(7) Arts	(8) Human interest
main								
Low pressure * Centrality	0.08	0.01	0.01	-0.05	-0.10	-0.01	-0.01	-0.10
	(0.09)	(0.10)	(0.15)	(0.07)	(0.24)	(0.06)	(0.07)	(0.27)
Low pressure	0.54	0.14	0.67	-0.35	-1.77	-0.14	-0.10	-0.74
	(1.02)	(1.08)	(1.59)	(0.83)	(2.72)	(0.71)	(0.85)	(2.94)
Centrality	-0.03	0.01	0.17	-0.02	-0.04	0.05	-0.01	0.10
	(0.08)	(0.10)	(0.15)	(0.07)	(0.22)	(0.06)	(0.07)	(0.25)
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Event-level controls	✓	✓	✓	✓	✓	✓	✓	✓
Observations	3,904	3,904	3,904	3,904	3,904	3,904	3,904	3,904
Mean DepVar	0.10	0.15	0.02	0.20	0.01	0.20	0.14	0.02
Sd DepVar	0.30	0.35	0.15	0.40	0.11	0.40	0.34	0.14

(b)

	(1) Weather	(2) Disaster	(3) Environment	(4) Religion	(5) Science	(6) Education	(7) Society	(8) Health
main								
Low pressure * Centrality	0.70	0.08	-0.09	0.27	-0.33	-0.07	-0.08	0.04
	(0.76)	(0.19)	(0.19)	(0.22)	(0.21)	(0.18)	(0.17)	(0.13)
Low pressure	9.23	1.01	-1.31	2.69	-2.74	-0.53	-1.29	0.47
	(9.61)	(2.39)	(2.12)	(2.58)	(2.27)	(2.11)	(1.89)	(1.52)
Centrality	-0.75	-0.10	-0.04	0.03	0.31	0.06	0.01	-0.08
	(0.67)	(0.19)	(0.16)	(0.20)	(0.19)	(0.15)	(0.15)	(0.13)
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Event-level controls	3,439	3,904	3,904	3,904	3,904	3,904	3,904	3,904
Observations	0.01	0.02	0.02	0.01	0.02	0.01	0.02	0.03
Mean DepVar	0.07	0.14	0.14	0.10	0.13	0.11	0.15	0.17

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using a logistic regression. 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 calendar-month fixed effects, as well as event-level controls as defined in equation (1) (the seed's number of followers at the time of the event and its squared value, and an indicator variable equal to one if the first tweet in the event is emitted during the night and to zero otherwise). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.8: Validity of the instrument: Named entities

	Places			Organizations			Individuals		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Low pressure * Centrality	-3.03	-2.78	-0.04	3.20	1.72	2.34	7.60	5.56	4.45
	(7.16)	(6.99)	(6.76)	(5.14)	(4.84)	(4.95)	(7.79)	(7.30)	(7.45)
Low pressure	-16.07	-24.52	10.29	63.22	38.85	45.26	57.01	46.20	30.64
	(79.65)	(78.50)	(74.76)	(58.25)	(55.02)	(56.62)	(84.50)	(79.12)	(81.23)
Centrality	-3.58	-7.23	-8.81	-0.98	1.25	0.81	-3.67	-3.34	-1.28
	(5.85)	(5.97)	(5.66)	(4.39)	(4.16)	(4.25)	(6.51)	(6.27)	(6.45)
# words in the event	0.02***	0.02***	0.02***	0.01***	0.01***	0.01***	0.02***	0.02***	0.02***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓	✓
Event-level controls		✓	✓		✓	✓		✓	✓
Drop media & journalist			✓			✓			✓
Observations	3,904	3,904	3,773	3,904	3,904	3,773	3,904	3,904	3,773
Mean DepVar	419.7	419.7	419.7	269.4	269.4	268.6	369.0	369.0	369.5
Sd DepVar	938.8	938.8	946.5	584.2	584.2	586.4	1,023.9	1,023.9	1,034.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using OLS (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. The dependent variable of interest is the sum of the total number of references to places in the articles published in the event in Columns (1) to (3), to organizations in Columns (4) to (6), and to individuals in Columns (7) to (9). All specifications include day-of-the-week and calendar-month fixed effects, and in Columns (2)-(3), (5)-(6) and (8)-(9) we also control for event-level controls as defined in equation (1) (the seed's number of followers at the time of the event and its squared value, an indicator variable equal to one if the first tweet in the event is emitted during the night and to zero otherwise, and measures of the topic of the event). Columns (1)-(2), (4)-(5) and (7)-(8) report the estimates for all the events that appear first on Twitter; in Columns (3), (6) and (9) we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.9: Validity of the instrument: Reliability of the sites (using Decodex)

	Reliable sites		Unreliable sites		False sites		Don't know	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Low pressure * Centrality	-0.08	-0.18	-0.05	-0.06	-0.01	0.01	0.14	0.24
	(0.42)	(0.40)	(0.07)	(0.07)	(0.09)	(0.09)	(0.39)	(0.37)
Low pressure	-0.50	-1.86	-0.37	-0.47	-0.38	-0.26	1.25	2.58
	(4.74)	(4.43)	(0.79)	(0.78)	(1.00)	(0.99)	(4.41)	(4.08)
Centrality	0.28	0.31	0.06	0.05	-0.04	-0.05	-0.30	-0.30
	(0.39)	(0.37)	(0.06)	(0.06)	(0.09)	(0.09)	(0.36)	(0.34)
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Event-level controls		✓		✓		✓		✓
Observations	3,903	3,903	3,903	3,903	3,903	3,903	3,903	3,903
Mean DepVar	82.2	82.2	1.1	1.1	1.2	1.2	15.5	15.5
Sd DepVar	15.8	15.8	3.0	3.0	2.6	2.6	15.3	15.3

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using OLS (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 calendar-month fixed effects, as well as in the even columns event-level controls as defined in equation (1) (the seed's number of followers at the time of the event and its squared value, an indicator variable equal to one if the first tweet in the event is emitted during the night and to zero otherwise, and measures of the topic of the event). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.10: Media-level approach, Reduced form and First stage estimates

	Reduced form				First stage	
	(1)	(2)	(3)	(4)	(5)	(6)
	# articles	# articles	# articles	# articles	Log # tweets	Log # tweets
Instrument						
Low pressure * Centrality			0.02**	0.02**	0.16***	0.13***
			(0.01)	(0.01)	(0.05)	(0.05)
Controls						
Centrality	0.01**		-0.00	-0.00	-0.19***	-0.16***
	(0.01)		(0.01)	(0.01)	(0.05)	(0.05)
Low pressure		0.02	0.21**	0.24**	1.93***	1.56***
		(0.01)	(0.10)	(0.10)	(0.57)	(0.59)
Log # seed's followers						
# seed's followers	-0.00	-0.00	-0.00	-0.01	0.08***	0.16***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.02)	(0.05)
# seed's followers-squared	0.00	0.00	0.00	0.00	-0.00***	-0.00***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night			0.01	0.01	-0.27***	-0.25***
			(0.02)	(0.02)	(0.05)	(0.05)
Media FEs	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist				✓		✓
Observations	658,917	658,917	658,917	636,844	658,917	636,844
Clusters (events)	3,904	3,904	3,904	3,773	3,904	3,773
Mean DepVar	0.3	0.3	0.3	0.3	4.1	4.1
Sd DepVar	2.0	2.0	2.0	2.0	1.6	1.6

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using OLS. 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. The dependent variable is the number of media articles published in the event in Columns (1) to (4) (reduced form estimates) and the logarithm of the number of tweets in Columns (5) and (6) (first stage estimates). All specifications include day-of-the-week, calendar-month and media fixed effects, and we also control for the topic of the event ("Topic of the event"). Columns (1) to (3) and (5) report the estimates for all the events that appear first on Twitter; in Columns (4) and (6), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.11: Media-level approach, IV estimates (Second stage), Conditional on covering the event

	Conditional on covering the event					
	(1) # articles	(2) # articles	(3) Length	(4) Length	(5) Originality	(6) Originality
Log(Number of tweets)	0.8	1.0	-55.0	-45.9	0.0	0.0
	(0.5)	(0.6)	(96.4)	(118.0)	(0.0)	(0.0)
Low pressure	-0.022	-0.030	-44.190	-46.290	-0.009	-0.006
	(0.202)	(0.219)	(41.480)	(41.904)	(0.009)	(0.009)
Centrality	0.137*	0.150*	-11.476	-12.052	-0.000	0.000
	(0.072)	(0.079)	(9.453)	(10.177)	(0.002)	(0.002)
# seed's followers	-0.085	-0.218	13.166	17.108	0.000	0.002
	(0.055)	(0.149)	(11.194)	(29.255)	(0.002)	(0.008)
# seed's followers-squared	0.002	0.004	-0.466*	-0.536	0.000	0.000
	(0.001)	(0.003)	(0.238)	(0.563)	(0.000)	(0.000)
=1 if first tweet during night	0.394*	0.462*	-27.148	-18.275	-0.008	-0.010
	(0.215)	(0.252)	(41.985)	(46.471)	(0.009)	(0.010)
Media FEs	✓	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist		✓		✓		✓
Observations	64,806	62,572	64,806	62,572	64,806	62,572
Clusters (events)	3,904	3,773	3,904	3,773	3,904	3,773
F-stat for Weak identification	11.3	7.6	11.3	7.6	11.3	7.6
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0	0.0
Mean DepVar	2.6	2.6	2,570.3	2,563.9	0.4	0.4
Sd DepVar	5.9	5.9	1,424.4	1,423.5	0.4	0.4

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a media-news event, and only the media outlets that devote at least one article to the event are included. Standard errors are clustered at the event level. The dependent variable is the number of articles in Columns (1) and (2), their length in Columns (3) and (4), and their originality in Columns (5) and (6). The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week, calendar-month and media fixed effects, and we also control for the topic of the event ("Topic of the event"). Odd columns report the estimates for all the events that appear first on Twitter; in even columns, we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.12: Naive estimates: Media-level approach, Depending on the number of journalists with a Twitter account

	Low nb journalists Twitter			High nb journalists Twitter		
	(1)	(2)	(3)	(4)	(5)	(6)
Log(Number of tweets)	0.01***	0.01***	0.00	0.04***	0.03***	0.04***
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)
# seed's followers	-0.00**	0.00	-0.00	-0.00	-0.02	-0.02
	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.02)
# seed's followers-squared	0.00	-0.00	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.01	0.01*	0.02*	0.04	0.04	0.05
	(0.01)	(0.01)	(0.01)	(0.03)	(0.03)	(0.04)
Total number of journalists			0.00***			0.00***
			(0.00)			(0.00)
Media FEs	✓	✓		✓	✓	
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Drop media & journalist		✓	✓		✓	✓
Observations	240,394	232,378	52,016	274,942	265,687	182,497
Clusters (events)	3,904	3,773	3,773	3,904	3,773	3,773
Number of media outlets included	71	71	71	73	73	73
Mean DepVar	0.1	0.1	0.1	0.5	0.5	0.7
Sd DepVar	0.8	0.8	0.7	3.0	3.0	3.5

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using a linear-log model. 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 and calendar month fixed effects, and we also control for the topic of the event ("Topic of the event"). In Columns (1)-(2) and (4)-(5), we also control for media fixed effects (in Columns (3) and (6) we instead control for the "total number of journalists" that is invariant at the media level). Columns (1) and (4) report the estimates for all the events that appear first on Twitter; in Columns (2)-(3) and (5)-(6), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). 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, defined with respect to the median (25). The number of tweets is computed before the first news article in the event appears. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.13: Media-level approach, IV estimates (Second stage), Depending on the number of times the media outlets tweeted in the event

	Low nb tweets in event		High nb tweets in event	
	(1)	(2)	(3)	(4)
Log(Number of tweets)	0.02	0.03	0.99*	1.27*
	(0.02)	(0.02)	(0.59)	(0.77)
Low pressure	0.00	0.00	0.03	0.05
	(0.01)	(0.01)	(0.25)	(0.27)
Centrality	0.00**	0.00**	0.17*	0.19*
	(0.00)	(0.00)	(0.10)	(0.11)
# seed's followers	-0.00	-0.01*	-0.09	-0.28
	(0.00)	(0.01)	(0.06)	(0.19)
# seed's followers-squared	0.00	0.00	0.00	0.01
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.01*	0.01*	0.60**	0.69**
	(0.01)	(0.01)	(0.29)	(0.34)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	618,102	597,538	40,809	39,299
Clusters (events)	3,904	3,773	3,724	3,595
F-stat for Weak identification	10.9	6.7	10.5	7.4
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.1	0.1	2.6	2.6
Sd DepVar	0.8	0.8	7.1	7.2

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month and media fixed effects. In Columns (1) and (2) (respectively (3) and (4)), we consider the media with a relatively low (respectively relatively high) number of Tweets in the event (defined using the median number of tweets tweeted by the media outlets in each event). In Columns (2) and (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.14: Media-level approach, IV estimates (Second stage), Depending on the media outlets' propensity to tweet in events

	Low tendency to tweet		High tendency to tweet	
	(1)	(2)	(3)	(4)
Log(Number of tweets)	0.03*	0.04	0.17*	0.23*
	(0.02)	(0.03)	(0.10)	(0.13)
Low pressure	0.00	0.00	-0.00	0.00
	(0.01)	(0.01)	(0.03)	(0.04)
Centrality	0.00	0.00	0.03**	0.03**
	(0.00)	(0.00)	(0.01)	(0.01)
# seed's followers	-0.00	-0.01	-0.02*	-0.05*
	(0.00)	(0.01)	(0.01)	(0.03)
# seed's followers-squared	0.00	0.00	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.01**	0.02**	0.07*	0.08*
	(0.01)	(0.01)	(0.04)	(0.04)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	302,053	291,965	356,864	344,879
Clusters (events)	3,904	3,773	3,904	3,773
Number of media outlets included	94	94	93	93
F-stat for Weak identification	10.4	6.4	12.0	7.5
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.1	0.1	0.4	0.4
Sd DepVar	0.6	0.6	2.7	2.7

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Columns (1) and (2) (respectively (3) and (4)), we consider the media with a relatively low (respectively relatively high) tendency to tweet during our period of interest (defined with respect to the median value of the overall number of tweets made by the media outlets in events between August 1st 2018 and November 30th 2018). In Columns (2) and (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.15: Media-level approach, IV estimates (Second stage), Depending on the reliance on advertising revenues

	1st quartile	2nd quartile	3rd quartile	4th quartile
	(1)	(2)	(3)	(4)
	# articles	# articles	# articles	# articles
Log(Number of tweets)	0.00	0.13	0.09*	0.08*
	(0.01)	(0.09)	(0.05)	(0.05)
Low pressure	0.01	-0.01	0.00	-0.01
	(0.00)	(0.03)	(0.02)	(0.01)
Centrality	0.00	0.02*	0.02***	0.02***
	(0.00)	(0.01)	(0.01)	(0.01)
# seed's followers	-0.00	-0.02**	-0.01**	-0.01**
	(0.00)	(0.01)	(0.01)	(0.01)
# seed's followers-squared	-0.00	0.00*	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.01*	0.06*	0.02	0.03
	(0.00)	(0.04)	(0.02)	(0.02)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Observations	78,849	130,437	110,984	112,522
Clusters (events)	3,904	3,904	3,904	3,904
Number of media outlets included	26	35	31	31
F-stat for Weak identification	10.8	11.7	12.2	11.0
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.0	0.4	0.2	0.2
Sd DepVar	0.4	1.7	1.3	1.2

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. All specifications include day-of-the-week, calendar-month, and media fixed effects. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. Column (1) includes the media outlets in the first quartile of the reliance on advertising (defined as the share of advertising revenues over total revenues) distribution, Column (2) those in the second quartile, Column (3) those in the third quartile, and Column (4) those in the fourth quartile. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.16: Media-level approach, IV estimates (Second stage), Private vs. public media outlets

	Private media		Public media	
	(1)	(2)	(3)	(4)
	Nb articles	Nb articles	Nb articles	Nb articles
Log(Number of tweets)	0.07*	0.10*	0.16	0.24
	(0.04)	(0.06)	(0.12)	(0.16)
Low pressure	-0.00	0.00	-0.00	0.01
	(0.01)	(0.02)	(0.05)	(0.05)
Centrality	0.01**	0.01**	0.03**	0.04**
	(0.01)	(0.01)	(0.01)	(0.02)
# seed's followers	-0.01**	-0.02*	-0.02	-0.06*
	(0.00)	(0.01)	(0.01)	(0.03)
# seed's followers-squared	0.00*	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.03*	0.03*	0.08*	0.10*
	(0.02)	(0.02)	(0.05)	(0.06)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	619,877	599,114	31,232	30,184
Clusters (events)	3,904	3,773	3,904	3,773
Number of media outlets included	177	177	8	8
F-stat for Weak identification	11.3	7.0	12.3	7.6
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.2	0.2	0.5	0.5
Sd DepVar	1.2	1.2	2.1	2.1

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. All specifications include day-of-the-week, calendar-month and media fixed effects. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. Columns (1) and (2) include the private media outlets, and Columns (3) and (4) the public ones. In the even columns, we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.17: 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).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.18: Unrepresentativeness of the Twitter users: Characteristics of the news-consuming citizens depending on whether they share news on Twitter

	Does not share news on Twitter	Share news on Twitter	Diff/se
Age	50	43	7.6***
			(1.3)
=1 if male (%)	46	61	-14.4***
			(4.2)
=1 if Bachelor degree or more (%)	32	46	-14.2***
			(4.2)
=1 if Annual income $\geq$ €30 000 (%)	45	49	-3.4
			(4.4)
=1 if leaves in Paris region (%)	18	24	-6.0
			(3.2)
=1 if highly interested in news (%)	50	77	-26.6***
			(4.1)
=1 if places itself on the Left (%)	33	49	-16.5***
			(3.9)
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 share news on Twitter. Column (2) presents the results for the individuals who share news Twitter. In Column (3) we perform a t-test on the equality of means (robust standard errors are in parentheses).

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.19: Media-level approach, IV estimates (Second stage): Heterogeneity depending on the media consumers' use of Twitter

	Low use of Twitter		High use of Twitter	
	(1)	(2)	(3)	(4)
Log(Number of tweets)	0.17	0.30	0.25*	0.32*
	(0.16)	(0.21)	(0.13)	(0.18)
Low pressure	-0.04	-0.02	0.01	0.02
	(0.06)	(0.06)	(0.05)	(0.05)
Centrality	0.05**	0.05**	0.04**	0.04**
	(0.02)	(0.02)	(0.02)	(0.02)
# seed's followers	-0.04**	-0.08*	-0.03*	-0.06
	(0.02)	(0.04)	(0.02)	(0.04)
# seed's followers-squared	0.00*	0.00*	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.11*	0.14*	0.08	0.10*
	(0.07)	(0.08)	(0.05)	(0.06)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	54,656	52,822	66,368	64,141
Clusters (events)	3,904	3,773	3,904	3,773
Number of media outlets included	14	14	17	17
F-stat for Weak identification	12.3	7.6	12.3	7.6
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.8	0.8	0.5	0.5
Sd DepVar	2.4	2.5	1.9	1.9

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Columns (1) and (2) ("Low use of Twitter"), we include the media outlets whose consumers have a relatively low use of Twitter, and in Columns (3) and (4) ("High use of Twitter"), the media outlets whose consumers have a relatively high use of Twitter (the information on the consumers' use of Twitter is from the *2018 Digital News Report* (Reuters Institute, 2018)). In Columns (2) and (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.20: Media-level approach, IV estimates (Second stage): Heterogeneity depending on the media consumers' age (using Reuters' survey data)

	Relatively young readership		Relatively old readership	
	(1)	(2)	(3)	(4)
Log(Number of tweets)	0.26*	0.35*	0.11	0.20
	(0.16)	(0.21)	(0.10)	(0.14)
Low pressure	-0.01	0.00	-0.02	-0.01
	(0.05)	(0.06)	(0.04)	(0.04)
Centrality	0.05**	0.05**	0.03**	0.03**
	(0.02)	(0.02)	(0.01)	(0.01)
# seed's followers	-0.04**	-0.07*	-0.02*	-0.06*
	(0.02)	(0.04)	(0.01)	(0.03)
# seed's followers-squared	0.00*	0.00	0.00	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.11*	0.13*	0.07	0.09*
	(0.06)	(0.07)	(0.05)	(0.05)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	85,888	83,006	35,136	33,957
Clusters (events)	3,904	3,773	3,904	3,773
Number of media outlets included	22	22	9	9
F-stat for Weak identification	12.3	7.6	12.3	7.6
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.7	0.7	0.5	0.5
Sd DepVar	2.3	2.3	1.9	1.9

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Columns (1) and (2) ("Relatively young readership"), we include the media outlets whose consumers are relatively young, and in Columns (3) and (4) ("Relatively old readership"), the media outlets whose consumers are relatively old (the information on the consumers' age is from the *2018 Digital News Report* (Reuters Institute, 2018); we use the average age of the readers to define the two groups). In Columns (2) and (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.21: Media-level approach, IV estimates (Second stage): Heterogeneity depending on the media consumers' age (using ACPM's survey data)

	Relatively young readership		Relatively old readership	
	(1)	(2)	(3)	(4)
Log(Number of tweets)	0.13*	0.18*	0.13	0.17
	(0.08)	(0.11)	(0.09)	(0.12)
Low pressure	-0.00	0.00	-0.00	0.00
	(0.03)	(0.03)	(0.03)	(0.03)
Centrality	0.03***	0.03**	0.02*	0.02*
	(0.01)	(0.01)	(0.01)	(0.01)
# seed's followers	-0.02**	-0.04*	-0.02*	-0.05**
	(0.01)	(0.02)	(0.01)	(0.02)
# seed's followers-squared	0.00*	0.00*	0.00*	0.00*
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.04	0.05	0.06*	0.07*
	(0.03)	(0.04)	(0.03)	(0.04)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Drop media & journalist		✓		✓
Observations	93,278	90,149	92,055	88,959
Clusters (events)	3,904	3,773	3,904	3,773
Number of media outlets included	24	24	25	25
F-stat for Weak identification	12.1	7.5	12.1	7.5
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.4	0.4	0.4	0.4
Sd DepVar	1.5	1.5	1.8	1.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Columns (1) and (2) ("Relatively young readership"), we include the media outlets whose consumers are relatively young, and in Columns (3) and (4) ("Relatively old readership"), the media outlets whose consumers are relatively old (the information on the consumers' age is from the ACPM's OneNext survey data, and we define the two groups using the share of the readers who are between 15 and 34 years old – below or above the median). In Columns (2) and (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table C.22: Media-level approach, IV estimates (Second stage), Depending on the size of the newsroom

	Low nb of journalists				High nb of journalists			
	(1) Nb articles	(2) Nb articles	(3) Length	(4) Originality	(5) Nb articles	(6) Nb articles	(7) Length	(8) Originality
Log(Number of tweets)	0.09*	0.10*	-38.75	0.04	0.36*	0.37*	-19.35	0.00
	(0.05)	(0.06)	(169.41)	(0.03)	(0.20)	(0.21)	(84.60)	(0.02)
Low pressure	0.00	0.00	-85.30	-0.01	-0.03	-0.03	-20.41	-0.01
	(0.02)	(0.02)	(71.09)	(0.01)	(0.07)	(0.08)	(40.51)	(0.01)
Centrality	0.01*	0.01*	-13.94	0.00	0.07**	0.07**	-10.89	0.00
	(0.01)	(0.01)	(15.77)	(0.00)	(0.03)	(0.03)	(8.76)	(0.00)
# seed's followers	-0.01**	-0.01**	6.01	-0.00	-0.04*	-0.04*	13.65	-0.00
	(0.01)	(0.01)	(20.26)	(0.00)	(0.02)	(0.03)	(10.64)	(0.00)
# seed's followers-squared	0.00*	0.00*	-0.21	0.00	0.00	0.00	-0.53**	0.00
	(0.00)	(0.00)	(0.48)	(0.00)	(0.00)	(0.00)	(0.23)	(0.00)
=1 if first tweet during night	0.04**	0.04**	-27.74	-0.00	0.15*	0.16*	-20.66	-0.01
	(0.02)	(0.02)	(73.08)	(0.01)	(0.08)	(0.08)	(37.92)	(0.01)
# journalists with Twitter account		0.00***				0.00***		
		(0.00)				(0.00)		
Media FEs	✓		✓	✓	✓		✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓	✓	✓
Observations	126,773	118,965	13,579	13,579	127,609	123,705	35,679	35,679
Clusters (events)	3,904	3,904	3,538	3,538	3,904	3,904	3,838	3,838
Number of media outlets included	33	33	33	33	33	33	33	33
F-stat for Weak identification	12.0	12.0	8.6	8.6	12.3	12.3	14.6	14.6
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.2	0.2	2,891.2	0.3	0.9	0.9	2,428.8	0.4
Sd DepVar	1.1	1.1	1,542.4	0.4	4.1	4.2	1,322.3	0.4

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles in Columns (1)-(2) and (5)-(6), and, conditional on covering an event, their length in Columns (3) and (7) and their originality in Columns (4) and (8). The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month and media fixed effects. In Columns (1) to (4) (respectively (5) to (8), we consider the media with a relatively low (respectively relatively high) number of journalists (defined with respect to the median). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

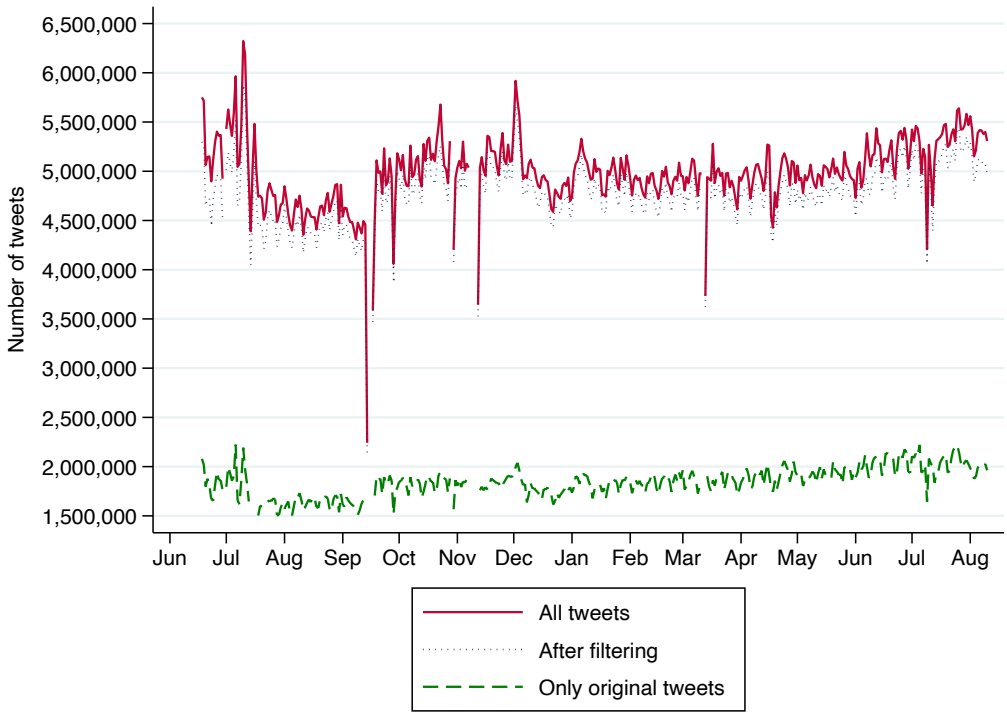
Table C.23: Media-level approach, IV estimates (Second stage), Depending on the reliability of the sites

	Reliable sites	Unreliable sites	False sites	Unreliable or False sites
	(1)	(2)	(3)	(4)
	# articles	# articles	# articles	# articles
Log(Number of tweets)	0.20*	0.01	0.02	0.02
	(0.11)	(0.03)	(0.05)	(0.03)
Low pressure	-0.00	0.01	-0.01	0.00
	(0.04)	(0.01)	(0.01)	(0.01)
Centrality	0.03**	0.01	0.00	0.00
	(0.01)	(0.00)	(0.00)	(0.00)
# seed's followers	-0.02*	-0.00	-0.01	-0.00
	(0.01)	(0.00)	(0.01)	(0.00)
# seed's followers-squared	0.00	0.00	0.00	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.08**	0.01	0.00	0.01
	(0.04)	(0.02)	(0.02)	(0.01)
Media FEs	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓
Observations	319,287	19,520	20,080	39,600
Clusters (events)	3,904	3,904	3,904	3,904
Number of media outlets included	86	5	6	11
F-stat for Weak identification	11.81	12.27	7.83	10.17
Underidentification (p-value)	0.0	0.0	0.0	0.0
Mean DepVar	0.5	0.1	0.1	0.1
Sd DepVar	2.8	0.7	0.6	0.7

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ $\times$ low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month and media fixed effects. In Column (1), we consider the media outlets whose site is rated as "reliable" by Decodex, in Column (2), those whose site is rated "unreliable", in Column (3) those whose site is rated "false", and finally in Column (4) we pulled together those whose site is rated either "unreliable" or "false". More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

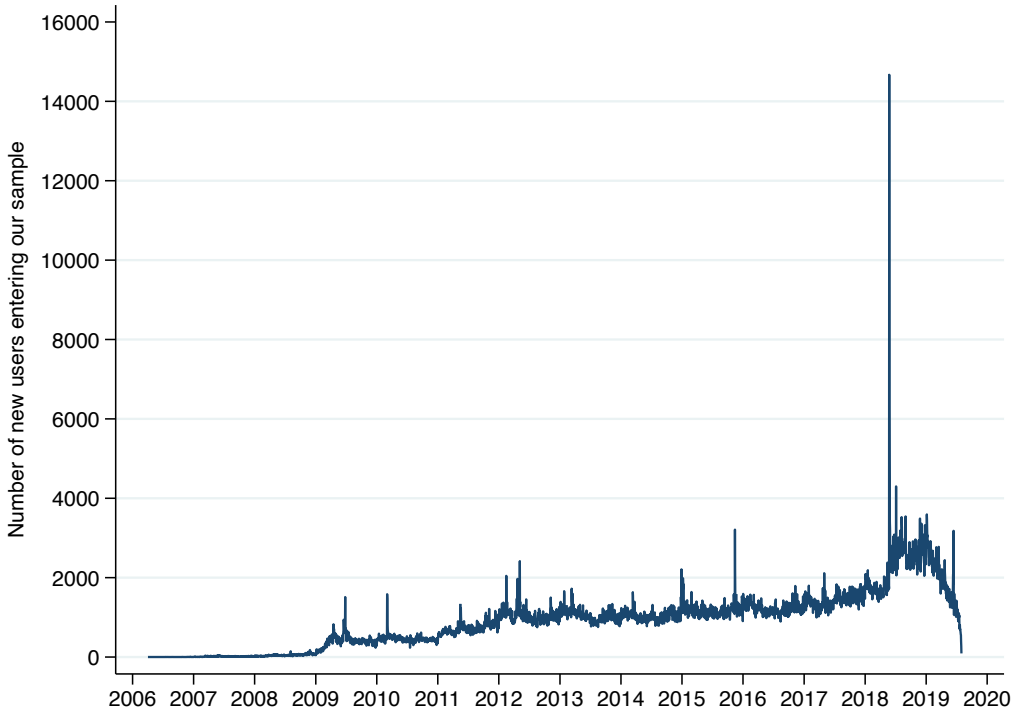
D Additional figures



Notes: The Figure plots the daily number of tweets included in our dataset. The red line plots all the tweets, the dot blue line these tweets once we apply the filter, and the green dashed line only the original tweets. Time period is June 18, 2018 - August 10, 2019. The few number of days without information comes from exceptional days when the server collapsed and we were thus unable to capture the tweets in real time.

Figure D.1: Daily distribution of the number of tweets in the sample

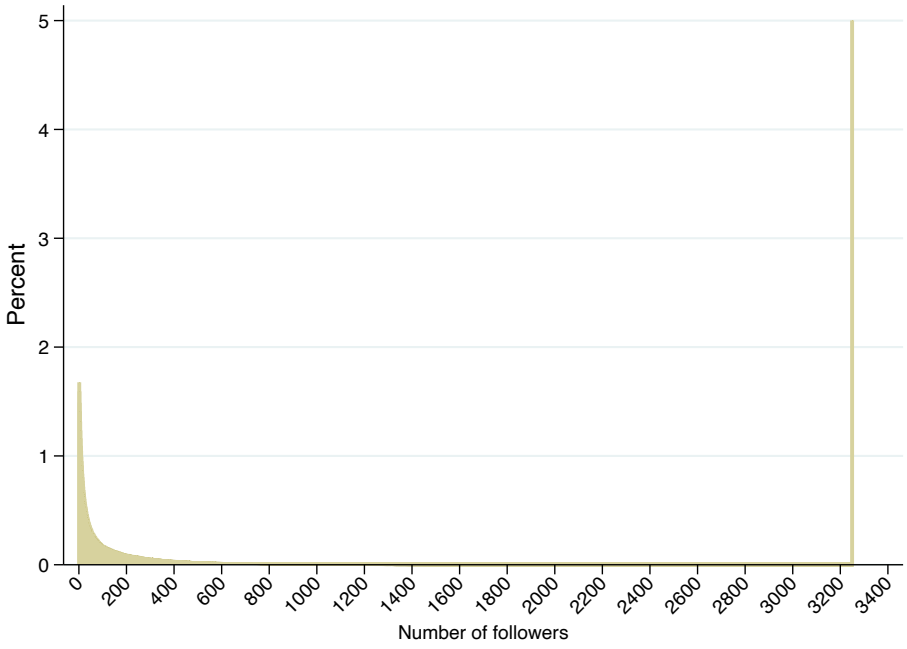
Electronic copy available at: https://ssrn.com/abstract=3663899



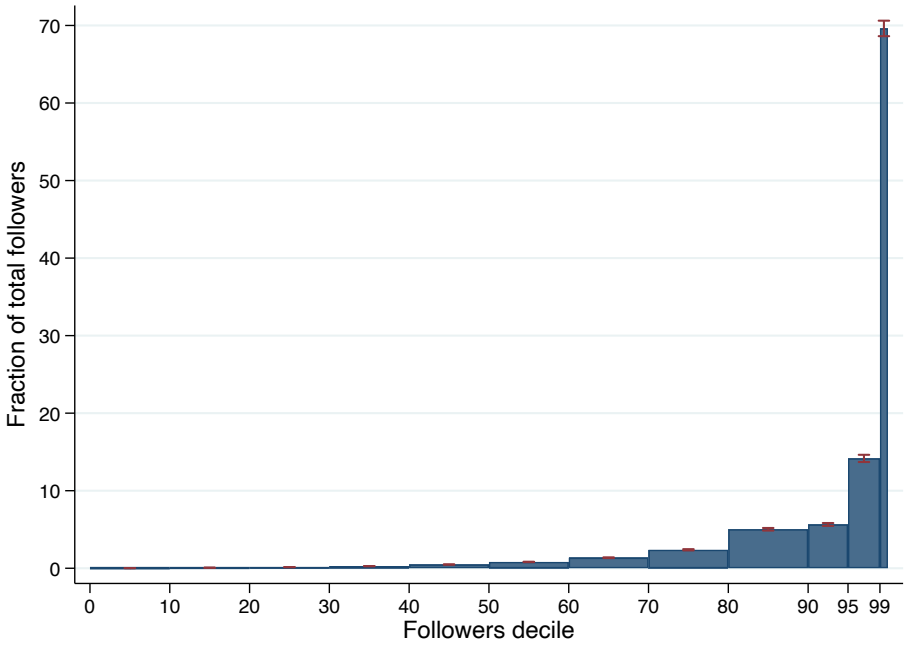
Notes: The figure plots the number of users entering our sample depending on the date of their Twitter account creation.

Figure D.2: Twitter users: Number of users depending on the date of their account creation

Electronic copy available at: https://ssrn.com/abstract=3663899



(a) **Distribution of the number of followers**

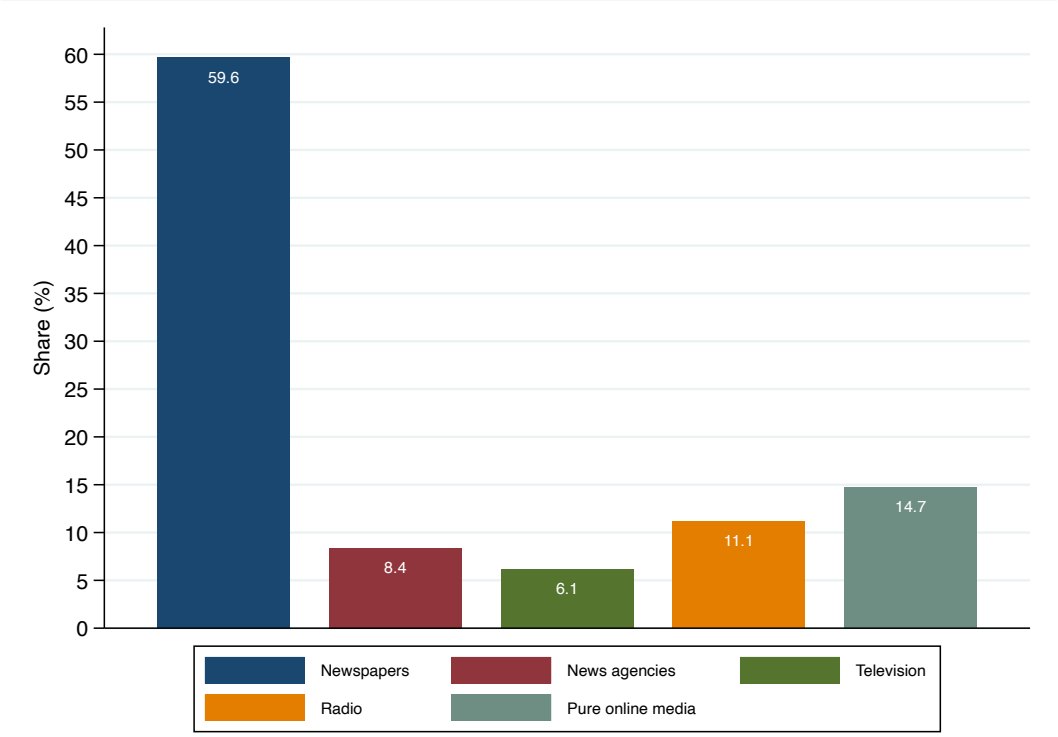


(b) **Cumulative distribution of the number of followers**

Notes: The upper Figure D.3a plots the distribution of the number of followers (winsorized at the 95th percentile, i.e. 3,355 followers) of the Twitter users in our dataset. The bottom Figure D.3b plots the cumulative distribution of the number of followers.

Figure D.3: Twitter users: Distribution of the number of followers

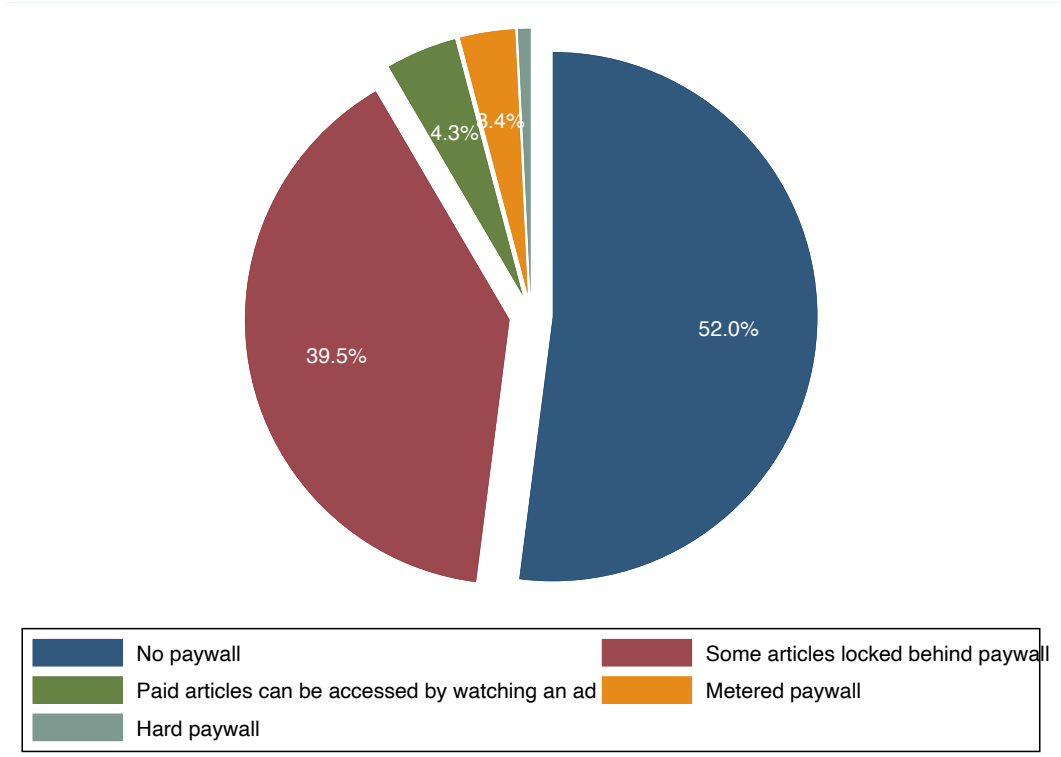
Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The figures plot the share of the documents depending on the offline format of the media outlet. The list of the media outlets included in each category is given in Section A.

Figure D.4: Share of the documents by offline format

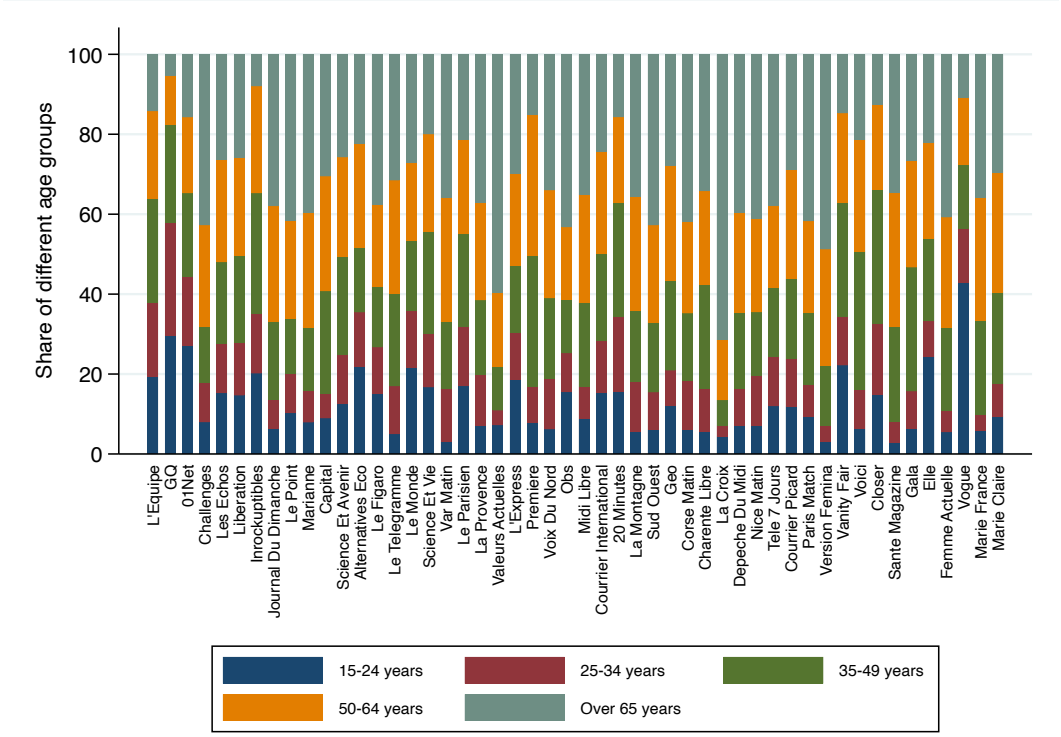
Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The Figure reports the share of the media outlets in our sample depending on their online business model. 52% of the media in our sample do not have a paywall ("no paywall"), and 4.3% 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 D.5: News editors' business model

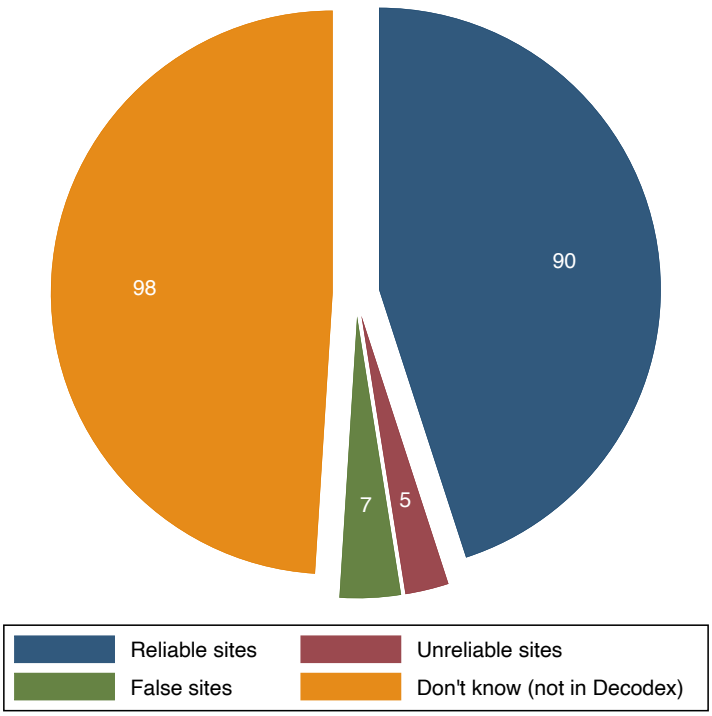
Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The figure provides the share of readers by age groups (15-24 years, 25-34 years, 35-49 years, 50 to 64 years, and over 65 years) for the 50 media outlets that are covered in the ACPM's OneNext study.

Figure D.6: Audience characteristics: Evidence from the ACPM's OneNext Study

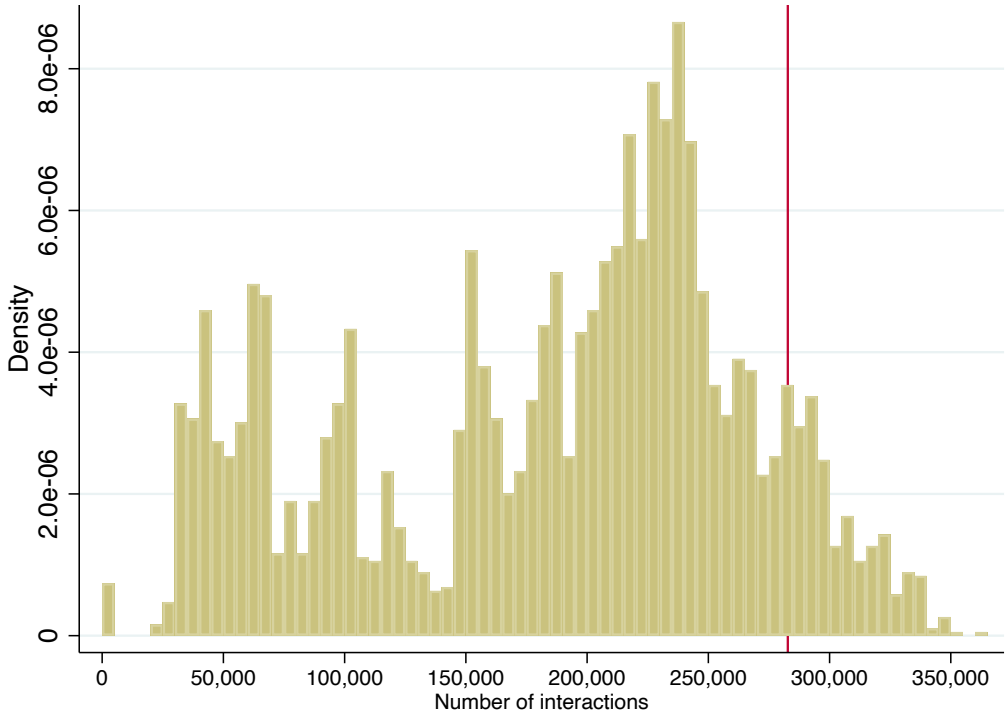
Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The Figure reports the share of the media outlets in our sample depending on their reliability. We rely on "Decodex", the *Le Monde*'s fact-checking product to estimate the "reliability" of the media outlets included in our dataset.

Figure D.7: Reliability of the websites

Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The figure plots the distribution of the number of interactions generated by the tweets posted in the hour preceding the event (with bins equal to $5,000$ interactions).

Figure D.8: Distribution of the number of interactions generated by the tweets posted in the hour preceding the event

Electronic copy available at: https://ssrn.com/abstract=3663899

E Robustness checks

Table E.1: Event-level approach, IV estimates (Second stage), Robustness: Not-news related measure of social media pressure

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	22.0*	18.0*	17.7*	18.2*	25.1*
	(11.9)	(9.9)	(10.1)	(10.3)	(14.6)
Low pressure	3.315	2.999	2.571	-0.514	-0.176
	(3.101)	(2.900)	(2.858)	(3.552)	(3.938)
Centrality	3.249**	2.634**	3.019**	3.156**	3.394**
	(1.270)	(1.203)	(1.279)	(1.312)	(1.448)
Log # seed's followers		0.657			
		(0.560)			
# seed's followers			-2.410*	-2.251*	-5.470*
			(1.264)	(1.220)	(2.987)
# seed's followers-squared			0.045*	0.042	0.098*
			(0.027)	(0.026)	(0.058)
=1 if first tweet during night				7.217*	8.780*
				(3.905)	(4.688)
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	10.0	12.4	11.8	11.5	7.0
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	43.5	43.5	43.5	43.5	43.5
Sd DepVar	81.1	81.1	81.1	81.1	81.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. All specifications include day-of-the-week and calendar-month fixed effects. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$×low news pressure$_e$ (see equation (3)). low news pressure$_e$ is measured by the number of interactions generated by all the tweets except the tweets generated by the Twitter accounts of journalists and of media outlets. The number of tweets is computed *before* the first news article in the event appears. Columns (1) to (4) report the estimates for all the events that appear first on Twitter; in Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.2: Media-level approach, IV estimates (Second stage), Robustness: Not-news related measure of social media pressure

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.14*	0.11*	0.11*	0.11*	0.15*
	(0.08)	(0.06)	(0.06)	(0.06)	(0.09)
Low pressure	0.02	0.02	0.02	-0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Centrality	0.02**	0.02**	0.02**	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.01*	-0.01*	-0.03*
			(0.01)	(0.01)	(0.02)
# seed's followers-squared			0.00*	0.00	0.00*
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.04*	0.05*
				(0.02)	(0.03)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	658,917	658,917	658,917	658,917	636,844
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	9.1	11.6	11.0	10.6	6.5
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.3	0.3	0.3	0.3	0.3
Sd DepVar	2.0	2.0	2.0	2.0	2.0

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). low news pressure$_e$ is measured by the number of interactions generated by all the tweets except the tweets generated by the Twitter accounts of journalists and of media outlets. The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.3: Event-level approach, IV estimates (Second stage), Robustness: Number of original tweets

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of original tweets)	26.2*	21.8*	21.3*	21.8*	28.5*
	(14.0)	(11.8)	(11.9)	(12.1)	(15.9)
Low pressure	4.349	3.949	3.439	0.734	1.216
	(3.055)	(2.855)	(2.810)	(3.329)	(3.654)
Centrality	4.174**	3.346**	3.741**	3.881**	4.303**
	(1.657)	(1.506)	(1.597)	(1.633)	(1.841)
Log # seed's followers		0.769			
		(0.544)			
# seed's followers			-2.420*	-2.280*	-5.389*
			(1.259)	(1.223)	(2.900)
# seed's followers-squared			0.045*	0.042	0.096*
			(0.026)	(0.026)	(0.056)
=1 if first tweet during night				6.386*	7.584*
				(3.573)	(4.070)
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	10.9	13.5	12.9	12.6	8.4
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	43.5	43.5	43.5	43.5	43.5
Sd DepVar	81.1	81.1	81.1	81.1	81.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of original tweets and is instrumented by centrality$_e$ ×low news pressure$_e$ (see equation (3)). The number of original tweets tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week and calendar-month fixed effects. Columns (1) to (4) report the estimates for all the events that appear first on Twitter; in Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.4: Media-level approach, IV estimates (Second stage), Robustness: Number of original tweets

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of original tweets)	0.16*	0.13*	0.13*	0.13*	0.17*
	(0.09)	(0.07)	(0.07)	(0.07)	(0.10)
Low pressure	0.03	0.03	0.02	0.01	0.01
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Centrality	0.02**	0.02**	0.02**	0.02**	0.03**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.01*	-0.01*	-0.03*
			(0.01)	(0.01)	(0.02)
# seed's followers-squared			0.00*	0.00	0.00*
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.04*	0.05*
				(0.02)	(0.02)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	658,917	658,917	658,917	658,917	636,844
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	10.3	12.9	12.4	12.1	8.1
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.3	0.3	0.3	0.3	0.3
Sd DepVar	2.0	2.0	2.0	2.0	2.0

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of original tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of original tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.5: Media-level approach, IV estimates (Second stage), Robustness: Only media outlets located in France

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.13*	0.11*	0.11*	0.11*	0.15*
	(0.07)	(0.06)	(0.06)	(0.06)	(0.09)
Low pressure	0.02	0.02	0.02	-0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Centrality	0.02***	0.02**	0.02**	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.01*	-0.01*	-0.03*
			(0.01)	(0.01)	(0.02)
# seed's followers-squared			0.00	0.00	0.00*
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.04*	0.05*
				(0.02)	(0.03)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	614,928	614,928	614,928	614,928	594,332
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	9.7	12.2	11.6	11.2	7.0
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.3	0.3	0.3	0.3	0.3
Sd DepVar	2.1	2.1	2.1	2.1	2.1

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. Only the media outlets that are located in France are included. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. In Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.6: Event-level approach, IV estimates (Second stage), Robustness: Additional controls

	Number of articles					
	(1)	(2)	(3)	(4)	(5)	(6)
Log(Number of tweets)	17.7*	17.9*	17.9*	18.0*	18.4*	18.9*
	(9.9)	(10.2)	(10.2)	(10.3)	(10.5)	(10.8)
Low pressure	-0.311	-0.017	-0.000	0.099	-0.023	-0.125
	(3.478)	(3.468)	(3.474)	(3.461)	(3.480)	(3.503)
Centrality	3.130**	2.885**	2.889**	2.909**	2.921**	2.856**
	(1.299)	(1.264)	(1.267)	(1.285)	(1.290)	(1.259)
# seed's followers	-2.207*	-3.538**	-3.524**	-3.565**	-3.657**	-5.142**
	(1.194)	(1.467)	(1.457)	(1.493)	(1.534)	(2.502)
# seed's followers-squared	0.041	0.064**	0.064**	0.065**	0.067**	0.093*
	(0.025)	(0.030)	(0.030)	(0.031)	(0.032)	(0.048)
=1 if first tweet during night	7.051*	8.166**	8.120**	7.811**	8.003**	8.491**
	(3.817)	(3.884)	(3.900)	(3.759)	(3.810)	(3.892)
Log # tweets seed has liked		-0.543	-0.547	-0.483	-0.415	-0.403
		(0.505)	(0.506)	(0.514)	(0.529)	(0.533)
Log # users seed's account is following		0.914	0.924	0.872	0.731	0.870
		(0.828)	(0.831)	(0.822)	(0.826)	(0.843)
Log # seeds' public lists		2.069*	2.105*	2.135*	1.504	0.554
		(1.154)	(1.189)	(1.203)	(1.062)	(0.946)
Log # seed's tweets		-0.150	-0.167	-0.077	-0.304	-0.024
		(0.775)	(0.790)	(0.767)	(0.830)	(0.732)
=1 if seed is located in France			-1.035	-0.502	-0.711	-0.791
			(2.954)	(2.838)	(2.858)	(2.889)
=1 if user language is French				-4.450	-5.417	-5.464
				(5.463)	(5.879)	(5.900)
Date of account creation					-1.067	-1.001
					(0.732)	(0.704)
= 1 if verified account						20.717
						(17.225)
Month & DoW FEs	✓	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓	✓
Observations	3,904	3,904	3,904	3,904	3,904	3,904
F-stat for Weak identification	12.2	11.8	11.8	11.7	11.5	11.4
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0	0.0
Mean DepVar	43.5	43.5	43.5	43.5	43.5	43.5
Sd DepVar	81.1	81.1	81.1	81.1	81.1	81.1

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week and calendar-month fixed effects. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.7: Media-level approach, IV estimates (Second stage), Robustness: Additional controls

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.11*	0.11*	0.11*	0.11*	0.11*
	(0.06)	(0.06)	(0.06)	(0.06)	(0.07)
Low pressure	0.00	0.00	0.00	0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Centrality	0.02**	0.02**	0.02**	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
# seed's followers	-0.01*	-0.02**	-0.02**	-0.02**	-0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
# seed's followers-squared	0.00	0.00**	0.00**	0.00**	0.00**
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
=1 if first tweet during night	0.04*	0.05**	0.05**	0.05**	0.05**
	(0.02)	(0.02)	(0.02)	(0.02)	(0.02)
Log # tweets seed has liked		-0.00	-0.00	-0.00	-0.00
		(0.00)	(0.00)	(0.00)	(0.00)
Log # users seed's account is following		0.01	0.01	0.01	0.00
		(0.00)	(0.00)	(0.00)	(0.00)
Log # seeds' public lists		0.01*	0.01*	0.01*	0.01
		(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's tweets		-0.00	-0.00	-0.00	-0.00
		(0.00)	(0.00)	(0.00)	(0.00)
=1 if seed is located in France			-0.01	-0.00	-0.00
			(0.02)	(0.02)	(0.02)
=1 if user language is French				-0.03	-0.03
				(0.03)	(0.04)
Date of account creation					-0.01
					(0.00)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event	✓	✓	✓	✓	✓
Observations	658,917	658,917	658,917	658,917	658,917
Clusters (events)	3,904	3,904	3,904	3,904	3,904
F-stat for Weak identification	11.3	11.0	11.0	10.8	10.7
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.3	0.3	0.3	0.3	0.3
Sd DepVar	2.0	2.0	2.0	2.0	2.0

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears, and we only consider the subset of news events that appear first on Twitter. All specifications include day-of-the-week, calendar month, and media fixed effects. More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.8: Event-level approach, IV estimates (Second stage), Robustness check: Dropping events whose seed has a verified Twitter account

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	18.1*	18.5*	18.4*	18.8*	22.3*
	(10.8)	(10.7)	(10.8)	(10.9)	(12.7)
Low pressure	2.819	3.177	3.015	-1.220	-1.034
	(3.010)	(3.022)	(3.019)	(3.622)	(3.788)
Centrality	2.498***	2.437***	2.460***	2.600***	2.691**
	(0.918)	(0.931)	(0.948)	(0.979)	(1.045)
Log # seed's followers		0.935			
		(0.661)			
# seed's followers			17.755	21.066	16.688
			(19.674)	(19.313)	(21.240)
# seed's followers-squared			-4.683	-5.284	-4.538
			(3.813)	(3.748)	(4.061)
=1 if first tweet during night				9.622**	10.678**
				(4.064)	(4.481)
Month & DoW FEs	✓	✓	✓	✓	✓
Drop media & journalist					✓
Observations	3,693	3,693	3,693	3,693	3,653
F-stat for Weak identification	8.1	8.3	8.1	8.0	6.7
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	42.6	42.6	42.6	42.6	42.5
Sd DepVar	71.7	71.7	71.7	71.7	71.8

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week and calendar-month fixed effects. Columns (1) to (4) report the estimates for all the events that appear first on Twitter except those whose seed has a verified Twitter account; in Column (5), we further drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.9: Event-level approach, IV estimates (Second stage), Robustness check: Winsorized measure of the number of tweets

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	18.3*	15.7*	15.4*	15.8*	20.8*
	(9.5)	(8.3)	(8.4)	(8.6)	(11.1)
Low pressure	2.995	2.750	2.285	-0.301	0.150
	(2.924)	(2.805)	(2.765)	(3.373)	(3.624)
Centrality	3.153**	2.567**	2.943**	3.060**	3.233**
	(1.224)	(1.175)	(1.245)	(1.266)	(1.353)
Log # seed's followers		0.754			
		(0.546)			
# seed's followers			-2.093*	-1.957*	-3.952*
			(1.106)	(1.076)	(2.108)
# seed's followers-squared			0.037	0.035	0.068*
			(0.023)	(0.022)	(0.040)
=1 if first tweet during night				6.030*	7.014*
				(3.396)	(3.735)
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	3,865	3,865	3,865	3,865	3,735
F-stat for Weak identification	14.9	17.4	16.8	16.4	10.9
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	43.5	43.5	43.5	43.5	43.5
Sd DepVar	81.4	81.4	81.4	81.4	82.2

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. An observation is a news event. Robust standard errors are reported between parentheses. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week and calendar-month fixed effects, and in Columns (2) to (5) we also control for the topic of the event ("Topic of the event"). Columns (1) to (4) report the estimates for all the events that appear first on Twitter; in Column (4), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.10: Media-level approach, IV estimates (Second stage), Robustness check: Selected sample – Only media outlets that produce at least 10 articles in events during our period of interest

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.15*	0.13*	0.12*	0.13*	0.17*
	(0.08)	(0.07)	(0.07)	(0.07)	(0.10)
Low pressure	0.03	0.02	0.02	-0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
Centrality	0.02***	0.02**	0.02**	0.02**	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.02*	-0.02*	-0.04*
			(0.01)	(0.01)	(0.02)
# seed's followers-squared			0.00*	0.00	0.00*
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.05*	0.06*
				(0.03)	(0.03)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	551,997	551,997	551,997	551,997	533,472
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	10.2	12.8	12.2	11.8	7.3
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.3	0.3	0.3	0.3	0.3
Sd DepVar	2.2	2.2	2.2	2.2	2.2

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event; only the media outlets that produce at least 10 articles in events during our period of interest are included. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week, calendar-month and media fixed effects, and in Columns (2) to (5) we also control for the topic of the event ("Topic of the event"). Columns (1) to (4) report the estimates for all the events that appear first on Twitter; in Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

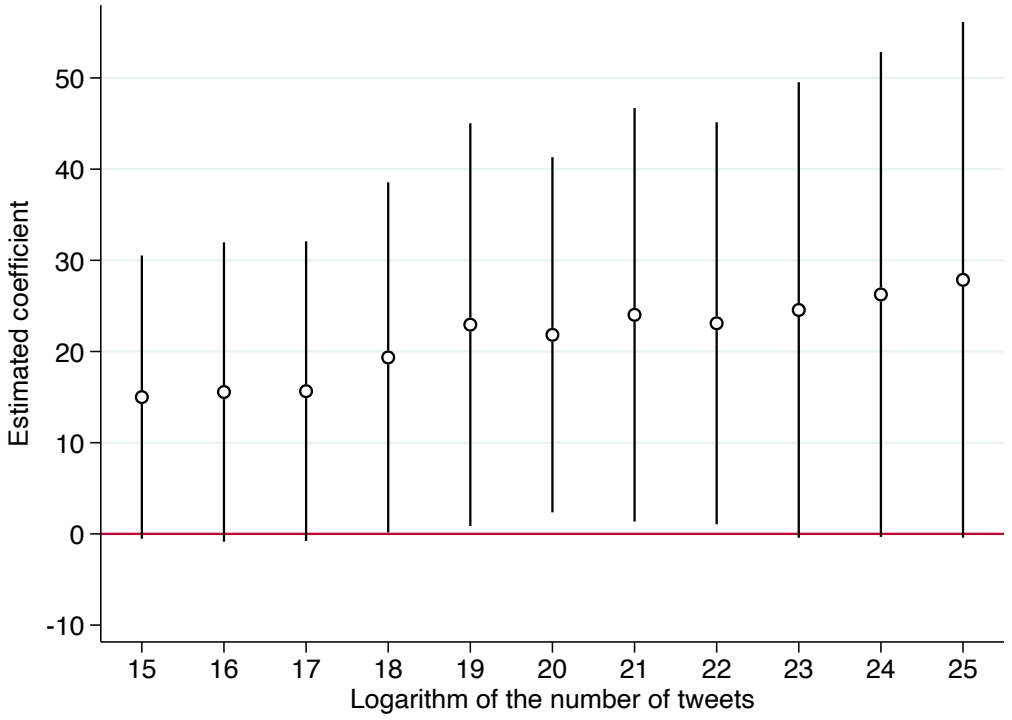
Electronic copy available at: https://ssrn.com/abstract=3663899

Table E.11: Media-level approach, IV estimates (Second stage), Robustness check: Excluding the news agencies

	Number of articles				
	(1)	(2)	(3)	(4)	(5)
Log(Number of tweets)	0.09*	0.08*	0.07*	0.08*	0.11*
	(0.05)	(0.04)	(0.04)	(0.05)	(0.06)
Low pressure	0.02	0.01	0.01	-0.00	0.00
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)
Centrality	0.01**	0.01**	0.01**	0.01**	0.01**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Log # seed's followers		0.00			
		(0.00)			
# seed's followers			-0.01**	-0.01**	-0.02*
			(0.01)	(0.01)	(0.01)
# seed's followers-squared			0.00*	0.00*	0.00
			(0.00)	(0.00)	(0.00)
=1 if first tweet during night				0.03*	0.04*
				(0.02)	(0.02)
Media FEs	✓	✓	✓	✓	✓
Month & DoW FEs	✓	✓	✓	✓	✓
Topic of the event		✓	✓	✓	✓
Drop media & journalist					✓
Observations	651,109	651,109	651,109	651,109	629,298
Clusters (events)	3,904	3,904	3,904	3,904	3,773
F-stat for Weak identification	9.7	12.3	11.7	11.3	7.0
Underidentification (p-value)	0.0	0.0	0.0	0.0	0.0
Mean DepVar	0.2	0.2	0.2	0.2	0.2
Sd DepVar	1.2	1.2	1.2	1.2	1.2

Notes: * p<0.10, ** p<0.05, *** p<0.01. The time period is August 1st 2018 - November 30th 2018. Models are estimated using an instrumental variable model. Standard errors are clustered at the event level. An observation is a media-news event; the news agencies AFP and Reuters are not included. The dependent variable is the number of articles. The endogenous explanatory variable is the logarithm of the number of tweets and is instrumented by centrality$_e$ × low news pressure$_e$ (see equation (3)). The number of tweets is computed *before* the first news article in the event appears. All specifications include day-of-the-week, calendar-month and media fixed effects, and in Columns (2) to (5) we also control for the topic of the event ("Topic of the event"). Columns (1) to (4) report the estimates for all the events that appear first on Twitter; in Column (5), we drop the events whose seed is the Twitter account of a media outlet or journalist ("Drop media & journalist"). More details are provided in the text.

Electronic copy available at: https://ssrn.com/abstract=3663899



Notes: The figure plots the coefficient associated to "Log(Number of tweets)" in equation (4) depending on the measure of centrality used, varying from the first 15 to the first 25 users. More details are provided in the text.

Figure E.1: Robustness check: Different measures of centrality

Electronic copy available at: https://ssrn.com/abstract=3663899

References

Aggarwal, Charu C, *Machine learning for text*, Vol. 848, Springer, 2018.

Aldecoa, Rodrigo and Ignacio Marin, "Deciphering Network Community Structure by Surprise," *PloS one*, 2011, *6.*

Allan, James, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang, "Topic Detection and Tracking Pilot Study Final Report," in "In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop" 1998, pp. 194–218.

_ , Stephen Harding, David Fisher, Alvaro Bolivar, Sergio Guzman-Lara, and Peter Amstutz, "Taking Topic Detection From Evaluation to Practice," in "Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04" HICSS '05 IEEE Computer Society Washington, DC, USA 2005.

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.

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.

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.

Newman, Mark E J and Michelle Girvan, "Finding and Evaluating Community Structure in Networks," *Physical review E*, 2004, *69* (2), 26113.

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

Salton, Gerard M., Andrew K. C. Wong, and Chungshu Yang, "A Vector Space Model for Automatic Indexing," *Commun. ACM*, 1975, *18* (11), 613–620.

Traag, Vincent A, Rodrigo Aldecoa, and J-C Delvenne, "Detecting communities using asymptotical surprise," *Physical Review E*, 2015, *92* (2), 22816.

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.

Electronic copy available at: https://ssrn.com/abstract=3663899

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

Electronic copy available at: https://ssrn.com/abstract=3663899