The Production of Information in an Online World: Is Copy Right?*

Julia Cagé†[1], Nicolas Hervé[2], and Marie-Luce Viaud[2]

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

April 4, 2019

Abstract

This paper documents the extent of copying and estimates the returns to originality in online news production. We build a unique dataset combining all the online content produced by French news media during the year 2013 with new micro audience data. We develop a topic detection algorithm that identifies each news event, trace the timeline of each story, and study news propagation. We unravel new evidence on online news production. First, we document high reactivity of online media: one quarter of the news stories are reproduced online in under 4 minutes. Second, we show that this comes with extensive copying: only 33% of the online content is original. Third, we investigate the cost of copying for original news producers. Using article-level variations and media-level daily audience combined with article-level social media statistics, we find that readers partly switch to the original producers, thereby mitigating the newsgathering incentive problem raised by copying.

Keywords: Internet, Information spreading, Copyright, Social media, Reputation
JEL No: L11, L15, L82, L86

*We are grateful to the Editor (Nicola Gennaioli) and to five anonymous referees for insightful comments that substantially improved the paper. We gratefully acknowledge the many helpful comments and suggestions from Charles Angelucci, Yasmine Bekkouche, Filipe Campante, Lucien Castex, Etienne Fize, Matthew Gentzkow, Sergei Guriev, Emeric Henry, Ali Hortacsu, Elise Huillery, Laurent Joyeux, Benjamin Marx, Petra Moser, Aurélie Ouss, Arnaud Philippe, Thomas Piketty, Andrea Prat, Valeria Rueda, Agnès Saulnier and Katia Zhuravskaya. We are grateful to participants at the Barcelona GSE Summer Forum, the Big Data for Media Analysis Conference, the CEPR Public Economics Annual Symposium, the Economics of Media and Communication Conference, the IEA World Congress, the NBER Political Economy Meeting, the NET Institute Conference on Network Economics, and SIOE 2017, and to seminar participants at Banque de France, the Paris School of Economics, Sciences Po Paris, the Toulouse School of Economics, and the University Carlos III of Madrid. We thank Jérôme Fenoglio and Pierre Buffet for sharing *Le Monde*'s data on the number of views per article, Andrew Montalenti for giving us access to Parsely's data, and Richard Fletcher for providing us the survey data from the 2013 and 2018 Reuters' Digital News Reports. Edgard Dewitte, Anais Galdin, Béatrice Mazoyer, Lucile Rogissart and Jeanne Sorin provided outstanding research assistance. This research was generously supported by the NET Institute, the Paris School of Economics, the Banque de France, Sciences Po's Scientific advisory board (SAB), and the French National Research Agency (ANR-17-CE26-0004). Since November 2015, Julia Cagé has been a Board member of the Agence France Presse; this paper does not reflect the views of the AFP and responsibility for the results presented lies entirely with the authors. An online Appendix with additional empirical material is available here.

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

1 Introduction

While online media have dramatically increased access to information, the impact of the Internet on news coverage has spurred concerns regarding the quality of news that citizens have access to. The switch to digital media has indeed affected the news production technology. The production of information is characterized by large fixed costs and increasing returns to scale (Cagé, 2017). Historically, newspapers have been willing to bear such a fixed cost in order to reap a profit from the original news content they provided (Schudson, 1981; Gentzkow and Shapiro, 2008). But in today's online world, utilizing other people's work has become instantaneous.[1] This makes it extremely difficult for news content providers to distinguish, protect and reap the benefits of the news stories they produce.[2] From a theoretical perspective, the impact of copying on media newsgathering incentives is ex-ante uncertain. Yet, understanding the different mechanisms at play has implications for the modern media industry and may help inform ongoing debates about the quality of 21st-century journalism. It also has clear relevance for other industries whose traditional revenue sources are collapsing due to new technologies. Digitization is indeed disrupting several industries beyond the news media, whose products are non-rival and can be copied at almost no cost, including books, music, and movies (see e.g. Waldfogel, 2017).

In this paper, we document the extent of copying online and estimate the returns to originality in online news production. Despite the intrinsic policy significance of the news industry and the growing importance of online news consumption, there is very little empirical evidence, particularly at the micro level, on the production of online information. We attempt to open up this black box by using new micro data and relying on a machine-learning approach. To do so, we build a unique dataset on online news production. More precisely, we examine the main French news media – including newspapers, television channels, radio stations, pure online media, the French news agency Agence France Presse (AFP), and Reuters' dispatches in French – and track every piece of content these outlets produced online in 2013. Our dataset contains 2.5 million documents.[3] To the extent of our knowledge, it is the first time that such a transmedia approach has been adopted to study the production of information, covering the entirety of the content produced by media online, whatever their offline format.[4]

[1] While print editions have simultaneous daily updates, online editions can be updated anytime. Moreover, not only do we observe an increase in the ease to "steal content" from competitors, but also an increase in the ease to "steal consumers". Increased consumer switching is indeed an essential distinguishing feature of online news consumption (Athey et al., 2013).

[2] According to Hamilton (2004), in the internet era, *"competitors' ability to confirm and appropriate a story once an idea is circulated reduces the incentives for journalists to spread large amounts of time on original, investigative reporting."*

[3] The reason for using French media in 2013 is mostly data driven. Content data for this research were indeed constructed as part of the OTMedia research project, a unique data collection program conducted by the French National Audiovisual Institute. Moreover, the French media market has the advantage of being by and large very similar to other Western media markets.

[4] Other studies have taken a transmedia approach to investigate media consumption patterns. See in par-

Using the content produced by news media, we perform a topic detection algorithm to construct the set of news stories. Each document is placed within the most appropriate cluster, i.e. the one that discusses the same event-based story. We obtain a total number of 25,000 stories. We then study the timeline of each story. In particular, for each story, we determine first the media outlet that breaks the story, and then analyze the propagation of the story, second-by-second. We investigate the speed of news dissemination and the length of the stories, depending on the topic and other story characteristics. Covering a news story does not necessarily imply providing original reporting on this story. We study how much each media outlet contributes to a story. More precisely, we develop a plagiarism detection algorithm to quantify the originality of each article compared to all the articles previously published within the event. The algorithm tracks small portions of text (verbatim) that are identical between documents. We distinguish between content copied from articles published by news agencies (to which media outlets subscribe) and content copied from competing media outlets.

Furthermore, we investigate the extent to which verbatim copying comes with acknowledgments. To do so, we develop a media reference detection algorithm to compute the number of citations received by each media outlet. A citation here is a reference to a news organization as the source of the story (e.g. "as revealed by *The New York Times*"). We study citation patterns at the event level.

Finally, in order to estimate the returns to originality in online news production, we collect audience data that we merge with the content data. For each website, we compute daily-level information on the number of unique visitors and the total number of page views and, for each article, we compute the number of times it has been shared on Facebook and on Twitter. We use this social media information to construct an audience measure at the article level and to investigate whether more original articles get relatively more views (regression analysis using event, date and media fixed effects).

Our main findings are as follows. First we find high reactivity of online media. E.g. we show that on average news is delivered to readers of different media outlets 169 minutes after first being published on the website of the news breaker, but in less than 4 minutes in 25% of the cases. The reaction time is the shortest when the news breaker is a news agency, and the longest when it is a pure online media.

Next, we show that high reactivity comes with high verbatim copying. We find that only 32.6% of the online content is original, and that moving from the first ventile to the last ventile of the reactivity distribution nearly doubles the originality rate. In effect, every time an original piece of content is published on the Internet, it is actually published three times:

ticular Prat (2018) who builds a media consumption matrix using survey data on the US covering television, radio, printed media, websites, and social media. See also Kennedy and Prat (2019) who similarly consider news consumption across platforms.

once by the original producer, and twice by media outlets who simply copy-and-paste this original content. Obviously, in practice we often observe an even larger number of media outlets copying part of the content of an original article: we show that more than 73% of the documents classified in events present at least some external copy.[5] But in terms of numbers of original characters copied, this is equivalent to a situation where each piece of original content is published three times.

The copied content partly comes from the press dispatches published by the news agencies to which the media outlets subscribe. If we exclude the content copied from these news agencies, we find that the average external copy rate is equal to 16%. Given the limitations of our plagiarism detection algorithm, which captures only exact verbatim copying but not rewording, this should be taken as a lower bound. Furthermore, while the media outlets are permitted to use the news agencies' material, the content reproduced from the press dispatches, insofar as it tends to be used by multiple media outlets (6 on average), cannot be considered "original" from the point of view of the consumers. Moreover, despite the substantiality of copying, media outlets hardly ever name the sources they copy: once we exclude copy from the news agencies, we show that only 3.4% of the documents mention the competing news organization they copy as the source of the information.

This new evidence sets the stage to investigate the implications of extensive copying. In particular, the scale of copying online might potentially negatively affect media outlets' newsgathering incentives. In the event that online audience was distributed randomly across the different websites and regardless of the originality of the articles, our results would imply that the original news producer captures only 33% of the audience and of the economic returns to original news production, which as a first approximation can be assumed to be proportional to audience. Advertising pricing on the Internet is indeed based on audience, and advertising is the largest contributor to publishers' online revenues (Anderson, 2012).[6]

From a theoretical perspective, the impact of copying on newsgathering incentives is relatively uncertain. In particular, it depends on a number of different parameters, including readers' mobility across media outlets, the quality of the copy with respect to the original, and consumers' valuation of originality. By using survey data on patterns of online readership, we first show that most consumers tend to consume news on multiple outlets online, thereby suggesting that switching behavior can play an important role. We present a very stylized theoretical framework to understand the different forces at play. On the one hand, readers have a preference for a specific media to which they tend to be loyal (e.g. for ideological reasons). On the other hand, they also have a preference for original news production. Depending on

[5] Verbatim copying can be either internal, if a media outlet copies-and-pastes content from documents it has itself previously published, or external if it reproduces content written by a competitor.

[6] Note however that the objective function of the media cannot be confined to the number of reads or their profitability. In particular, media owners may also derive utility from non-monetary profits, e.g. their political influence, as we will discuss in Section 4.3.

the relative strength of these individual-specific parameters, and depending on the extent to which the copying media offers a lower-quality coverage than the original news producer, they might decide to switch to the online media that has produced original information, either at the daily level or on a longer-term basis.

We attempt to estimate some of the model parameters in the following way. First, we present evidence showing that copying is of lower quality than the original. In particular, copying tends to be incomplete. On average, when an article is copied, "only" 9.4% of its content is reproduced; this means that nearly nine tenths of the content of the original article is missing from the copy. Furthermore, besides incomplete copying, there might be other reasons why copying leads to lower-quality articles, e.g. because the original is bundled with additional information that is absent the copy.

Second, using article-level variations (with event, date and media fixed effects), we show that a 50-percentage-point increase in the originality rate of an article leads to a 40% increase in the number of times it is shared on Facebook, and to a 17% increase in the number of Tweets. We discuss a number of possible channels that may help rationalize these findings. Our preferred explanation is that consumers may favor originality. In particular, investigating whether the returns to originality vary depending on the characteristics of the media outlets, we show that originality has a stronger positive effect for the outlets which are in a more competitive environment, and so more subject to switching. We also find that the returns to originality are lower for the media that are more copied by their competitors. These heterogeneous effects are consistent with the predictions of our simple theoretical framework.

With the data at our disposal, we are not able to decompose how much of the extra audience comes from consumers' social networks (i.e. the Facebook and Twitter shares of their friends), and how much comes from other mechanisms, e.g. the fact that consumers might browse across news sites and pick the one offering the best coverage for a given story. In order to further investigate this issue, one would need information on the websites' traffic sources (ideally at the article level) and specific survey data. To the extent of our knowledge, such micro-level audience data is not available to the researcher. In any case, the point is that these mechanisms are sufficient to redirect a substantial fraction of the online audience to the original producer, which in some ways is reassuring as regards the media's incentives to produce original news. Furthermore, we should also stress that although these effects seem to be quite strong, they only include switching behavior at the short-run level. It is possible that longer-run reputation effects allow original producers to recoup an even larger share of the audience. We provide some indicative evidence of such a reputation effect.

Finally, we combine media-level daily audience data and article-level social media statistics (number of Facebook and of Twitter shares) to obtain an audience measure (number of views) at the article level. We first assume a simple linear relationship between the number of shares

on social media and the number of article views. We then use a unique data set on the number of views and Facebook shares at the article level from *Le Monde* (covering the period April to August 2017) to characterize the joint distribution of the number of Facebook shares and the number of visitors. We use these different estimates to obtain a lower and an upper bound of the number of times each article is viewed. We show that a 50-percentage-point increase in the originality rate of an article leads to a 45% increase in its number of predicted readers. Lastly, depending on the specification we use, we find that the original content represents between 45.4 and 61.4% of online news consumption, i.e. much more than its relative share in total online content (32.5%). In other words, media outlets with a larger fraction of original content tend to receive a higher audience.

Of course, our results do not imply that reputation effects and consumers' preference for originality alone can solve plagiarism issues. Greater intellectual property protection could also play a role in raising the incentives for original news production, and we certainly do not mean to downplay the extent of this problem. However, our results suggest that in order to effectively address this issue, it is important to study reputation effects and how viewers react to the newsgathering investment strategies of media outlets.

Related literature Using micro data, Gentzkow (2007) estimates the relationship between the print and online newspapers in demand.[7] Our paper is complementary to his. We investigate the production of original content and document the benefits of original information production. Franceschelli (2011) was the first to assess empirically the impact of the Internet on news coverage.[8] Using a dataset that includes every article published by the two main Argentinean newspapers, he reconstructs the typical timeline of a news story in the online world.[9] Compared to this previous work, our contribution is threefold. First, we construct the set of news stories and study their timeline using the entire universe of French news media online, rather than two newspapers. To the extent of our knowledge, we are the first to study simultaneously the content produced by all the news media, whatever their offline format. Moreover, we identify the stories that result from original reporting by a news organization. Second, while Franceschelli (2011) relies restrictively on the mention of proper nouns to identify the news stories, we develop and run a state-of-the-art algorithm relying on word frequency without any restriction. Hence, our paper also contributes to the existing literature from a methodological point of view: in particular, we develop a new event detection algorithm that could be of use in the future for other researchers interested in text analysis and clustering. Third and most importantly, we quantify the importance of plagiarism online and

[7] On the effect of the Internet on the demand for traditional media, see also George (2008).

[8] Salami and Seamans (2014) also study the effect of the Internet on newspaper content, and in particular newspaper readability. But they examine the production of content offline, not online.

[9] Boczkowski (2010) has conducted an ethnographic study of editorial work at these two Argentinean newspapers.

combine this new evidence from the production side with article-level information on news consumption using social media data. This allows us to estimate the returns to originality in online news production.

Our results also complement a growing empirical literature on copyright (MacGarvie and Moser, 2014; Biasi and Moser, 2015; Giorcelli and Moser, 2015; Li et al., 2018). Most of the literature on copyright online has centered on digitization and piracy within the music industry (Rob and Waldfogel, 2006; OberholzerGee and Strumpf, 2007; Waldfogel, 2012, 2015).[10] With the exception of Chiou and Tucker (2017), there is little evidence on copying and intellectual property regarding online news media. Yet, the modern news media industry shares a number of important characteristics with the cultural industry online; in particular, digital products just like news articles are non-rival, non-excludable, and can be copied at almost no cost (see e.g. Bae and Choi, 2006; Peitz and Waelbroeck, 2006a). We contribute to this literature by providing new empirical evidence on the extent of copying online and estimating the returns to originality. Our paper is a unique attempt to understand who is producing news, the character of what is produced and the propagation of information in the online world.[11]

The rest of the paper is organized as follows. In Section 2 below, we describe the media universe and the content data we use in this paper, and review the algorithms we develop to study the production and propagation of information online. Section 3 provides new evidence on the speed of news dissemination and the importance of copying online, discusses heterogeneity in the copying behavior and media outlets' reputation, and quantifies verbatim copying without acknowledgement. In Section 4, we discuss the mechanisms at play and the theoretical framework that we use to analyze the impact of originality and copying on consumer behavior. In Section 5, we use article-level variations to investigate the relationship between originality and online audience and estimate the returns to originality in online news production. Section 6 performs a number of robustness checks and discuss the external validity of our main findings. Finally, Section 7 concludes.

2 Data and algorithms

2.1 Media universe

Our dataset covers 86 general information media outlets in France: 1 news agency; 59 newspapers (35 local daily, 7 national daily, 12 national weekly, 2 national monthly, and 3 free newspapers); 10 pure online media (i.e. online-only media outlets); 9 television channels; and

[10] Recent work has also investigated the effect of digitization projects like Google Books (Reimers, 2019; Nagaraj, 2018). For an assessment of the impact of copyright laws on the magazine industry in America during the 18th and 19th centuries, see Haveman and Kluttz (2014) and Haveman (2015).

[11] Sen and Yildirim (2015) investigate how popularity of online news stories affect editors' decisions. Athey et al. (2013) provide a model of advertising markets for news media.

7 radio stations. The news agency is the Agence France Presse (AFP), the third largest news agency in the world (after the Associated Press and Reuters). Moreover, our dataset also includes all the dispatches published in French by Reuters.[12] For each of these media outlets, we gather all the content they published online in 2013.[13]

There was only a small number of online-only news media in France in 2013, the main ones being Mediapart, a website created in 2008 with a hard paywall model, which attracted nearly 600,000 unique visitors a month, the freely-accessible Rue89 (1.479 million unique visitors), Atlantico (1.258 million), Slate (966,000), the Huffington Post (the French version launched in 2012), and Agoravox. As highlighted by the Open Society Foundations (2013) in its study of the French digital media market in 2013, these pure online media *"provide general information, mostly similar to that occurring in offline and online editions of traditional media, but they also try to establish their own independent editorials as well as comment on political, social, and economic issue".* For example, Mediapart built a reputation through investigative journalism.

The complete list of the media outlets included in our dataset is provided in the online Appendix Section C.1.[14] The 86 media outlets included in our sample are by far the main French news media both during our period of interest (2013) and still today.[15] The choice of 2013 France is data driven: the content data was collected as part of the OTMedia research project conducted by the INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French radio and television audiovisual archives). To the best of our knowledge, there is no equivalent dataset for other countries and time periods. This allows us to provide unique evidence on the propagation and verbatim copying of news stories online.[16]

We choose a "transmedia" approach because, on the Internet, there is a tendency for different media to converge (see e.g. Peitz and Reisinger, 2016). One cannot infer the offline format of a media by visiting a website, as illustrated in the online Appendix Figure G.1. On

[12] We do not have data for the AP's dispatches in French, which do not seem to be used significantly by any of the major French media. In contrast, all the media outlets included in our sample subscribed in 2013 to either or both the AFP and Reuters.

[13] However, we do not consider their offline news production, e.g. the content of the news bulletins only broadcast on television.

[14] In the online Appendix Section C.1, we also indicate the name of the companies that own each of these media outlets. Compared to other countries, the market for online news is very competitive in France, with a low concentration (Noam, 2016; Kennedy and Prat, 2019; Cagé, 2017).

[15] Only missing are those local daily newspapers that had no websites at the time, and some very small digital news media that could not be considered important information providers in 2013. Also not included in our analysis is Wikipedia. While Wikipedia is the largest encyclopedia on the web, and an important source of news on ongoing events, it relies entirely on free contributions (see e.g. Greenstein and Zhu, 2012; Greenstein et al., 2016; Algan et al., 2016). Our interest in this paper is rather on traditional media and on their incentives to invest in original news production. Note moreover that contributors on Wikipedia have to source the information they provide, and often use traditional media as a source.

[16] Moreover, as we will see in Section 6.3, the French media market is by and large very similar to other Western media markets, and it has been far from upset since 2013.

the web, media all offer texts, videos and photos. We include the AFP even though it does not deliver news straight to individual consumers[17] because it is a key provider of original information in the online world. For the same reason, we incorporate the dispatches published in French by the news agency Reuters. We think it is essential to consider news agencies when investigating newsgathering and copying online. To the extent of our knowledge, we are the very first to perform such an inclusive empirical analysis of original news production.[18]

Using their RSS feeds, we track every piece of content news media produced online in 2013. For the media outlets whose RSS feeds were not tracked by the INA, we complete the OTMedia data by scrapping the Sitemaps of their website. Finally, we get all the AFP and Reuters dispatches directly from the agencies. Merging these datasets, we obtain the universe of all the articles published online by French news media in 2013 (the only year for which the data is available). The articles we use in our database contain text and often photos, as well as videos. Our focus here is on text.[19]

Our dataset contains $2,552,442$ documents for the year 2013; around $7,000$ documents on average per day. Figure 1 plots this number on a daily basis. On average, more documents are published during the week, and we observe a drop in this number during the weekends.[20] Note however that, interestingly, while media outlets do not face the same space constraint online that they face offline[21], the total amount of content produced on a daily basis is relatively stable through time. While space online is technically infinite, media outlets indeed still face an implicit space constraint which is the limited attention of the readers. Hence online, media outlets may be wise to publish less rather than more content.[22]

70.9% of the documents are from the websites of the print media; 4.5% from radio; 6.4% from television; 15.1% from the AFP and Reuters and the remaining documents from the pure online media (online Appendix Figure G.2a). On average, these documents are $2,058$ characters long.[23] Table F.1 in the online Appendix provides summary statistics for the entire

[17] News agencies are based on a Business-to-Business model (they sell news to other media outlets), not on a Business-to-Consumer model.

[18] We do not consider news aggregators and curators, however, nor do we investigate information dissemination on social media. Doing so is well beyond the scope of this paper whose focus is on original news *producers*. On the effect of aggregators, see Athey and Mobius (2012); George and Hogendorn (2012, 2013); Chiou and Tucker (2017); Calzada and Gil (2016).

[19] We do not study the online production of videos and photos. Analyzing the propagation of photos and videos online require different technical tools and algorithms than those we develop here and will be the topic of future research.

[20] The drop in the number of documents we observe in July is due to a combination of two factors. First, fewer journalists work in July and so less information is produced due to the summer vacation. Second, because of a heatwave, a number of servers broke down at the INA in July; as this happened during the summer vacation, it took more time than usual to fix them and we (unfortunately) lost a number of documents.

[21] See e.g. Eisensee and Strömberg (2007) who have documented the fact that media operate under constraints of limited space and time, which may lead to the crowding out of some newsworthy piece of news.

[22] As emphasized by Peitz and Reisinger (2016), *"if a user has a limited attention span for news and is unable or unwilling to push herself to read more news, a media platform (...) can emphasize the most relevant news items"*.

[23] Online Appendix Figure G.3 plots the distribution of the length of the articles. For the reader to have

sample, as well as by media format (print media, television, radio, pure online media and news agencies).

[Figure 1 about here.]

In the rest of this section, we briefly review the algorithms we develop to study the production and propagation of information online. We provide more details in the online Appendix Section C and perform a number of robustness checks in Section 6. In the online Appendix Section E, we illustrate these different algorithms by taking the example of a specific news event.

2.2 Event detection

Event detection algorithm Using the set of documents previously described, we perform an event detection algorithm to detect media events. This category of algorithm is often referred to as Topic Detection and Tracking (TDT) in the computer science community. These algorithms are based on natural language processing methods. The goal of online topic detection is to organize a constantly arriving 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. Each document is described by a semantic vector which takes into account both the headline and the text.[24] A semantic vector represents the relative importance of each word of the document compared to the full dataset. A standard scheme is TF-IDF.[25]

2. 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).

in mind an order of magnitude, opinion pieces by Paul Krugman in the *New York Times* are around 4,000 characters long.

[24]Vectorization is an embedding technique which aims to project any similarity computation between two documents. Describing documents by a semantic vector is usual in the computer science literature nowadays. But, to the extent of our knowledge, it is an improvement compared to what has been done so far in the economic literature, e.g. Franceschelli (2011) considering only proper nouns.

[25]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. We describe the TF-IDF weight more formally in the online Appendix Section C.2. In Section 6.2, we show that our main findings are robust to using an alternative embedding scheme.

3. 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.

4. A cluster is finalized if it does not receive any new document for a given period of time. We use a one-day window.[26]

Finally, to ensure consistency, we only keep the events with documents from at least two different media outlets, and with more than 10 documents in our preferred specification. In Section 6, we relax this condition and investigate the extent to which it affects our main results. We also discuss alternative embedding techniques for news events and compare their clustering performances.

Performance of the algorithm This event detection algorithm can be compared to other detection systems by its ability to put all the stories in a single event together. To ensure the performance of our algorithm, we perform two robustness checks.[27]

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.[28] We find that the performance of our algorithm is as good as the one of the state-of-the-art algorithms. In particular, our implementation has performances that clearly outperform the best online algorithm of Allan et al. (1998).[29]

[26]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.

[27]We have also looked at the GDELT project (https://www.gdeltproject.org/) that extracts events from news articles. Two important things need to be highlighted. First, for the year 2013, the GDELT's coverage of the French news media is very low (e.g. their dataset only includes 3 articles from *20 Minutes*, 2 from *Challenges*, 22 from Europe1, 10 from France Info, etc.). Furthermore, the GDELT project uses a different definition of what a news event is; their focus is on the identification of the people involved in the news event, and on the categorization of the events in a given taxonomy. Hence, they define each news event based on only one article from which they extract the people involved, the location of the event, etc. On the contrary, we define events from the clustering of multiple news articles dealing with the same topic. This allows us to study news propagation and identify copy online; such an analysis could not be performed with the GDELT's data. Hence, the GDELT's dataset cannot be used to assess our results or the quality of our event detection algorithm.

[28]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. More details are provided in the online Appendix Section C.2.

[29]We provide details of the statistical measures of the performance of the algorithm in the online Appendix Section C.2. Note that using the TDT dataset as a benchmark may suffer from a number of shortcomings that have been highlighted by Allan et al. (2005). In the context of our study, the main potential shortcoming is the occurrence of "garbage" clusters containing a very large number of stories. In the online Appendix Table F.3, we present summary statistics on the number of documents per event. The maximum number of documents in an event is $2,393$, which can indeed be interpreted as a garbage cluster. However, we have very few occurrences of such large clusters. The 99th percentile is indeed equal to 269. In the robustness Section 6, we show that dropping these few large clusters does not affect our main results.

As an additional robustness check, we compare our events to those obtained by the Europe Media Monitor (EMM) NewsExplorer.[30] The EMM NewsExplorer provides on a daily basis the top 19 stories of the day. With our event detection algorithm, we match 92% of the stories in their sample.

2.3 News events

We obtain a total number of 25,215 news events. Events can last more than one day; on average, they last 41 hours.[31] The average number of documents per event is 34 and, on average, 15 media outlets refer to an event (online Appendix Table F.3). There are 182 events per day on average, with 69 new events beginning every day. These events are roughly equally distributed during the year. Online Appendix Figure G.5 plots the total number of events per day, as well as the number of new events.

Out of the 2,552,442 documents in the dataset, 851,864 (33.4%) are classified in events (for a daily plot of this ratio, see Figure G.6 in the online Appendix). The remaining 66.6% of the documents are not classified in events. Note however that the classified documents represent, in terms of characters, 40% of the total content produced in 2013. Classified documents are indeed longer on average (online Appendix Table F.1). Note moreover that relaxing the "10 documents condition" to define an event increases the share of articles classified in events (we then classify 47.2% of the articles in events, and these articles represent more than 53% of the total content), but does not affect our main results (Section 6).

The fact that we leave out as much as 50 to 60% of the total online content as unclassified documents (depending on the specification) is clearly an important limitation of our analysis. At the same time, we should stress that unclassified documents raise a number of special issues, and that most of them can be considered as relatively less central from the viewpoint of information production and diffusion. In particular, unclassified documents come mostly from local daily newspapers (while local newspapers represent 55.7% of the documents in our dataset, they account for 66.8% of the unclassified documents). Local newspapers indeed cover very local affairs that are not covered either by other local outlets (whose market differs) nor by national outlets.[32] We think that leaving these articles out of our event analysis is not a major problem.

Other unclassified documents correspond to one-off reports, to what Schudson (2015) calls

[30]The EMM NewsExplorer is an initiative of the European Commission Research Centre. http://emm.newsexplorer.eu/NewsExplorer/home/fr/latest.html

[31]Note that what we define here as the length of an event is the length of the event *coverage* – the time interval between the first and the last article covering the event – not the length of the actual event. Online Appendix Figure G.4 plots the distribution of this length.

[32]While more than 50% of the documents published by national newspapers, radio, and TV are classified in events, only 20% of those published by local newspapers are (see online Appendix Figure G.7 for a plot of these ratios).

"contextual reporting"[33], as well as to editorial and opinion pieces. This last category can actually be of importance to public debate and information. However, we feel that these documents would require a specific analysis. Furthermore, when they are published in the middle of a news sequence, these editorial and opinion pieces will generally be classified in a news event (and so appear in our analysis as part of the classified documents).

Note also that, on average, unclassified documents are much less popular in terms of social media audience than the classified ones. Online Appendix Table F.2 presents statistical differences between the articles classified and not classified in events: on average, documents that are not classified have approximately half as many Facebook and Twitter shares than those that are classified.[34]

Hence, in this paper, given that our subject of interest is the propagation of news stories online and the importance of copying, we focus our main analysis on the $851,864$ articles classified in our $25,215$ events.[35] Table 1 provides summary statistics on these articles.[36]

[Table 1 about here.]

Topic of the events We classify the events according to their topic. In order to do so, we rely on the metadata associated with the AFP dispatches included in the event. There is at least one AFP dispatch in nearly 95% of our events (we do not define the topic of the remaining events). The AFP uses the 17 IPTC classes to classify its dispatches.[37] These top-level media topics are: (i) Arts, culture and entertainment; (ii) Crime, law and justice; (iii) Disaster and accidents; (iv) Economy, business and finance; (v) Education; (vi) Environment; (vii) Health; (viii) Human interest; (ix) Labour; (x) Lifestyle and leisure; (xi) Politics; (xii) Religion and belief; (xiii) Science and technology; (xiv) Society; (xv) Sport; (xvi) Conflicts, war and peace; and (xvii) Weather.

Figure 2 plots the share of events associated with each media topic (given that some events are associated with more than one topic, the sum of the shares is higher than 100%). Nearly one third of the events are about "Politics", 29% about "Economy, business and finance" and around 23% about "Crime, law and justice". "Sport" comes fourth, appearing in 13% of the

[33] Fink and Schudson (2014) classify news articles into five possible categories: investigative, contextual, conventional (conventional stories focus on one-time activites or actions that have occured or will occur within 24 hours), social empathy (social empathy stories describe a person or group of people not often covered in news stories) and other. In earlier work, Tuchman (1980) defined five categories of news: hard, soft, spot, developing, and continuing.

[34] Facebook and Twitter shares data is described in details in Section 2.7 below.

[35] In Section 6, the number of classified articles increases to $1,203,521$ when relaxing the "10 documents condition" to define an event.

[36] Table 1 also includes summary statistics on the number of shares on Facebook and on Twitter which we will further describe below.

[37] More precisely, to define the subject, the AFP uses URI, available as QCodes, designing IPTC media topics (the IPTC is the International Press Telecommunications Council). These topics are defined more precisely in the online Appendix Section C.5. An event can be associated with more than one top-level media topic.

events. The other topics like "Weather", "Education" or "Science and technology" have much less importance. This does not mean that there is no article related to these topics, but that these topics are not associated with *events*.

[Figure 2 about here.]

We then trace the timeline of each story and study news propagation.

2.4 Timeline and plagiarism detection

Timeline More precisely, for each event, we order the documents depending on the timing of their publication, determine the media outlet that breaks the story, and then rank the other outlets. Using the publication time, we also document how long it takes each media outlet to cover the story.

The fact that a media outlet is talking about a story does not necessarily mean that it is providing original reporting on that story, however. We thus study how much each media outlet contributes to a story. To measure this contribution, we develop a plagiarism detection algorithm in order to quantify the original content in each document compared to the content of all the documents published earlier in the event.

Plagiarism detection algorithm The plagiarism detection algorithm efficiently tracks identical portions of text between documents.[38] For each document, we determine the portions of text that are identical to content previously published by all the documents out earlier in the event, and isolate the original content in the document. The originality rate of a document is defined as the share of the document's content (in number of characters) that is original.

Moreover, we trace back each portion of text to its first occurrence in the event. It allows us to determine for each document the number of times it is copied and the share of the document which is ultimately copied.

2.5 Exclusive vs. non-exclusive news events

In the case of a government press release giving rise to a number of articles, the first media outlet covering the story cannot be considered a news breaker providing exclusive news. We may also overestimate verbatim copying by attributing the release to the first media outlet and counting as copy the reproduction of the release by other outlets. To deal with this issue, we manually code all the news stories in our sample to isolate the stories that are the results

[38] Technically, the algorithm is based on *hashing* techniques of n-grams (the n-grams consist in sets of n consecutive words, we use 5-grams) and a threshold on the minimal length of a shared text portion to consider there is a copy (we use 100 characters). We use an *hashing*-based technique to save processing times (see e.g. Stein, 2007). For more details, see online Appendix Section C.3. We focus on exact (verbatim) copying only.

of a piece of original reporting by (at least) one outlet.[39] We call these stories exclusive news events. The remaining stories are either non-exclusive news events or short news items with multiple witnesses.

To distinguish between these three types of news events, we investigate the nature of the information issuer. More precisely, we define as non-exclusive news events those news events where the original information can be considered to be in the public domain, and was not produced by the media outlet itself. This includes news events where the information issuer is the government, the police, companies or non-governmental organizations, as well as cultural and sport events. In the online Appendix Section D, we provide more details on the different kind of information issuers and additional descriptive statistics.

We then define two categories of exclusive news events: investigative stories and (non-investigative) reporting stories. Investigative stories involve substantial in-depth reporting, whereby media outlets are playing watchdog (e.g. the NSA spying scandal revealed by *Le Monde* on October 21, 2013 and described in the online Appendix Section E). (Non-investigative) reporting stories are stories for which the news originates from the presentation of facts by the media outlet, but with limited in-depth reporting. Typically, these are facts that nobody tries to hide and which a media outlet decides to present to the public.

Finally, short news items with multiple witnesses are news events for which there are multiple witnesses: e.g. a public protest; a murder in the public space; a terrorist attack; a plane crash.[40] The journalists may not be the first to report the story – e.g. due to breaking news alerts on social media – but they are the first to provide "reliable" information on the story.

We find that non-exclusive news events represent 84.5% of the events, as illustrated in Figure 3. Short news items with multiple witnesses account for 8.4% of the events, and exclusive news events for 7%. Finally, only 1.3% of the events in our sample can be considered as investigative stories. While this number may seem low, it is in fact in line with previous findings in the literature. E.g. examining a sample of front-page stories at the *Milwaukee Journal Sentinel*, the *New York Times*, and the *Washington Post*, Fink and Schudson (2014) find that investigative reports only represent 1% of the stories in 2003 (see also Rosenstiel et al., 2007).[41]

[Figure 3 about here.]

[39] To ensure consistency, all the stories have been coded twice, by two different Research Assistants. The classification of the stories for which the Research Assistants first disagree has then been discussed at length by the authors.

[40] We include weather-related events in this category (but they only represent 0.6% of the events).

[41] These two examples are described in details in Hamilton (2016).

2.6 Citation detection

Do media outlets obey the formal procedures for citing and crediting when they copy? To answer this question, we finally develop an algorithm to detect media citations in the documents. Citations are references to a news organization as the source of the information, e.g. *"as revealed by Le Monde"*. In particular, we distinguish when a media is referred to as the source of the information from when the information is about the media outlet itself (e.g. appointment, takeover,...) This algorithm is described in the online Appendix Section C.4.

In every document in our sample, we identify all the citations to media outlets as the source of the information. It is indeed not unusual to have references to more than one media in a document, e.g. when a scoop is revealed by a media outlet and commentated by a politician on the website of another outlet, or when a scoop is revealed by a media outlet and gives rise to an AFP dispatch reproduced by other outlets. We study citation patterns in Section 3.4.

2.7 Audience data

Lastly, we collect audience data that we merge with the content data.

Daily-level audience data First, we measure online audience for the media outlets in our sample using data from the OJD (the French press organization whose aim is to certify circulation and audience data): for a subset of websites – 58 out of the 85 media outlets in our sample[42] – we have information on the number of unique visitors, the number of visits and the number of page views.[43] This information is available at the daily level. The average daily number of page views is around 1.6 million. Table 2 provides summary statistics for these variables.

[Table 2 about here.]

Facebook shares Furthermore, we collect information on the number of times each article has been shared on Facebook. We do so by using the Facebook Graph API (Application Programming Interface) (see online Appendix Section C.7).

We obtain information on this variable for all the documents in our sample, with the exception of the articles published by the AFP and Reuters that are not available online to the general audience. On average, articles are shared 64 times on Facebook; however, half of the articles are not shared. The distribution of the number of Facebook shares is skewed to the right (the standard deviation is equal to 956 and the maximum is $240,450$ while the 99th

[42] The AFP being based on a Business-to-Business model, it does not deliver news to individual consumers on its website. Similarly, there is no audience data for Reuters' dispatches.

[43] Websites whose audience is very small are not monitored by the OJD.

percentile is "only" 1,017). We discuss below a number of empirical strategies to deal with this issue. In Table 1, we present summary statistics for this variable using both the raw data and a winsorized version of the Facebook shares variable at the 99th percentile.

When winsorized at the 99th percentile, articles classified in events are shared on average 37 times on Facebook. Note moreover that some topics tend to generate more shares than others. In particular, whether we consider the average number of shares received by the articles depending on the topic of the event in which they are classified, or the total number of shares of all the articles in an event depending on the topic of the event, articles dealing with "Sport" and with "Economy, business and finance" generate much less attention on social media than articles about "Crime, law and justice" and about "Arts, culture and entertainment" (online Appendix Figure G.8).[44]

If the data on the number of Facebook shares can be used as a proxy for the "popularity" of each article, it also suffers from a number of caveats. In particular, Facebook shares do not directly reflect consumer demand since they are filtered through the Facebook News Feed algorithm. Hence we collect social media statistics at the article level from an additional source, namely Twitter.

Twitter As opposed to Facebook, Twitter does not provide a specific API to measure the popularity of given web pages on the social media. However, the Twitter Search API gives access to tweets containing specified keywords, as long as they were published in the past seven days. We use this feature to collect, for each article, all the tweets containing the article's URL and thereby obtain information on the number of times articles are shared on Twitter. We provide more details on the procedure we follow in the online Appendix Section C.8.

For each article, we have eight different measures of the number of times it its "shared" on Twitter: (i) the number of direct tweets; (ii) the number of direct retweets; (iii) the number of direct likes; (iv) the number of direct replies; and then computing the statistics on the retweets and the replies (v) the indirect number of tweets; (vi) the indirect number of retweets; (vii) the indirect number of replies; and (viii) the indirect number of likes. Obviously, all these different measures are very strongly correlated (online Appendix Table C.1). For the sake of simplicity, in our preferred specification, we consider an aggregate measure of the number of shares. More precisely, for each article, the total number of times it is shared on Twitter is defined as the sum of the values for these eight measures.[45]

Lastly, note that there is a positive relationship between the number of shares on Facebook

[44] Note that this effect is not driven by differences in average originality rates. Articles classified in "Sport" events have indeed a higher originality rate than the average, while articles about "Crime, law and justice" have a lower originality rate than the average.

[45] As a robustness check, we show that our results are robust to instead considering each of these measures separately (see below).

and the number of shares on Twitter thus defined, as illustrated in the online Appendix Figure G.9. In Section 5, we use these social media statistics as a proxy for the number of views. To test the accuracy of this proxy, we rely on evidence from *Le Monde* newspaper and, using five months of data from its website, show that the relationship between the number of views and the number of social media shares is almost perfectly linear. Although this is an argument in favor of using such a proxy, it must be kept in mind that not all the online readers share articles on social media. To investigate the extent to which people who share articles on social media (mainly Facebook and Twitter) are selected compared to the population that read news but do not share online, we use survey data from the *2013 Digital News Report* (Reuters Institute, 2013).[46] Among the individuals consuming news online, 16% share information on social networks such as Facebook and Twitter while 84% do not.[47] In the online Appendix Table F.4, we compare the characteristics of these two groups. Not surprisingly, we find that the readers who share news on social media tend to be significantly younger than the others, with a statistically significant difference, and that they have a higher probability to live in Ile-de-France. They also have a higher probability to have children under the age of 18 (but this may be due to the fact that they are younger on average). We find no other statistically significant differences between these two groups however, whether we consider house status (tenant or owner), type of city (village, small or big city), education, gender, income, marital status, or work status. Hence, while the readers who share news articles on social media are selected, this selection seems to come mostly from age.

3 Empirical analysis

3.1 The speed of news dissemination

In this Section, we study the speed of news dissemination online.[48] I.e. we investigate how quickly news is delivered to readers of different media outlets after being published first on the website of the news breaker.[49]

Studying the speed of news dissemination is of interest because the commercial value of a news item may depend on how long a news media retains exclusive use of it. We first study the time interval between the publication of the first document covering a story and the second one. We find that on average, it takes 169 minutes for some information published

[46] These data are described in more details in Section 4.2 when documenting readers' mobility across the media.

[47] The question asked is: *"During an average week in which, if any, of the following ways do you share or participate in news coverage? Share a news story via social network (like Facebook, Twitter, or Reddit)"*, and the surveyed individuals answer by "Yes" or "No".

[48] In the online Appendix Section J, we provide additional evidence on the temporal pattern of news publication.

[49] Unfortunately, we do not have information on when the actual news event takes place; the only information we have is the exact time at which the event is reported for the first time by a media outlet in our sample.

by a media outlet to be published on the website of another outlet. But this average masks considerable heterogeneity. In half of the cases, it takes less than 22 minutes, of which less than 243 seconds in 25% of the cases and less than 6 seconds in 10% of the cases.

Table 3 reports the average reaction time depending on the offline format of the news breaker. If a news agency (the AFP or Reuters) is the first media outlet to publish some information, then the reaction time is shorter. When a news agency is the news breaker, we find that the second media outlet covers it after 116 minutes on average, but after only 11 minutes in half of the cases and in 1 second or less in 5% of the cases. This rapidity comes from the fact that media outlets receive the news directly from the news agency; they don't have to monitor it the way they monitor what is published on their competitors' website. Furthermore, a number of media outlets have automatized the posting of prepackaged AFP content. In other words, AFP content of their choice is automatically integrated into their website.

The average reaction time is longer when the news breaker is a media outlet other than the news agency. This appears clearly in Figure 4, which represents the Kaplan-Meier survival functions depending on whether the news breaker is a news agency or another media outlet. In the online Appendix Table F.5, we report the results of a Cox proportional hazards regression where we relate a "news agency news breaker" indicator variable (equal to one if the news breaker is a news agency and to zero otherwise) to survival time, i.e. in our context, to the reaction time of the second media outlet covering the event. We obtain a 0.65-unit increase in the expected log of the relative hazard for the events whose news breaker is a news agency as compared to the news events whose news breaker is not a news agency. This difference is statistically significant at the one-percent level and robust to controlling for date fixed effects.

[Figure 4 about here.]

We find that the reaction time is the highest when the news breaker is a pure online media. Even if demonstrating this lies beyond the scope of this article, a possible explanation is that pure online media may suffer from a lower reputation. Hence legacy media may want to wait for multiple sources before covering an event broken by these new media. An alternative hypothesis is that the news provided by pure online media are of less interest and/or are of lower quality. Hence, other media outlets might be less interested in publishing the corresponding news stories, and additionally might monitor pure online media to a lesser extent. The "waiting behavior" of the news media when the news breaker is not a news agency may also be driven by some strategic considerations: given that some readers have a taste for original news production, news media may decide to take the time to add original content to the primary news story, even if that comes at the expense of reactivity (we will see below that there is a negative correlation between reactivity and originality). In the online Appendix

Figure G.10, we plot the Kaplan-Meier survival estimates separately for each of the offline format of the news breaker. Consistently with Table 3, it appears clearly that the probability of surviving past 169 minutes (the average reaction time) is much higher if the news breaker is a pure online media than for other offline formats (while there is no statistically significant differences between television, radio and print media).

[Table 3 about here.]

We also investigate how the reaction time varies depending on the nature of the news events. We show that the reaction time is the shortest for the short news items with multiple witnesses, and that the differences are statistically significant.[50] Finally, in the online Appendix Section J.3, we provide some additional evidence on the profile of the news events.

3.2 The importance of copying online

We now turn to an estimation of the originality of the articles published online in 2013. This is a key question because the high reactivity of the media discussed above may actually come from the use of plagiarism, and the use of plagiarism may negatively affect newsgathering incentives.

Originality rate We first use our plagiarism detection algorithm to determine for each document the portions of text that are identical to content previously published by all the documents released earlier in the event, and isolate the original content in the document. By definition, the originality of the first article in the event is 100%.

On average, the originality rate of the documents classified in events is equal to 36.5%.[51] In Figure 5a, we plot the distribution of the originality rate. The distribution is bimodal with one peak for the articles with less than 1% of original content (nearly 17% of the documents) and another peak for the 100%-original articles (nearly 22% of the documents). The median is 14%. In other words, with the exception of the documents which are entirely original, the articles published within events consist mainly of verbatim copying: 54.6% of the articles classified in events have less than 20% originality.

We study how the originality varies with the nature of the news event. Figure 5b plots the Kernel density estimates. We find that articles published in non-exclusive news events tend to have a lower originality rate. This is not surprising: non-exclusive news events are

[50] In the online Appendix sub-Section J.2, we document how the reaction time varies with the publication time of the breaking news.

[51] Given that documents are of different lengths, we also compute the ratio of original content in the dataset over the total content. We find that the share of original content is equal to 32.6%. In other words, nearly 70% of online information production is copy-and-paste. This finding is consistent with the results obtained by Boczkowski (2010) who highlights the rise of homogeneization in the production of news stories online by two Argentinean newspapers.

indeed events derived from information that is in the public domain (e.g. a government press release) and media outlets tend to reproduce this information as it is.

[Figure 5 about here.]

In the online Appendix, we further document how the originality rate varies depending on the characteristics of the articles. In particular, online Appendix Figure G.11 illustrates how the average originality rate varies depending on the document length: short articles tend to rely less on copy than longer ones.

Originality and reaction time Figure 6 shows the average originality rate of the articles for each ventile of the reactivity distribution. Moving from the first to the last ventile of the reactivity distribution increases the average originality rate from around 26% to around 40%. In other words, on average, the longer it takes for a media outlet to cover an event, the higher the originality rate of the article. This finding is robust to dropping the articles published by the news agencies (online Appendix Figure G.12), and to computing the reactivity distribution at the media-outlet level (online Appendix Figure G.13).

[Figure 6 about here.]

Where does the copied content come from? We trace back each "identical portion" of text to its first occurrence in the event. Hence, for each document, we determine: (i) the original content, (ii) the number of documents copied (including documents published by the media outlet itself), and (iii) for each document copied, the number of characters copied. (Obviously, if a media outlet reproduces content that has already been published by more than one outlet previously in the event, we cannot determine from which document the copying outlet has actually copied the content. It might indeed not have reproduced it from the original content provider. However, assuming that media outlets copy content from its first occurrence seems to be the most sensible assumption.)

Table 4 presents the results. Variables are values for documents, and we consider all the documents classified in events, with the exception of the press dispatches published by the AFP and Reuters. We find that, on average, documents include content from 4.1 documents previously published in the event.

Internal vs. external copying Verbatim copying can be either "internal" or "external". A media outlet can indeed copy and paste content from documents it has itself previously published (in particular when it is updating previous versions of the same article, for example adding new elements). Conditional on publishing at least one document related to the event, half of the media outlets publish at least 2 documents in the event.

We find that out of the 851,864 documents classified in events, 620,760 (72.8%) present at least some external copy. On average, documents include content from 3.9 documents previously published in the event by competing media outlets. If we sum up the external copied content, we obtain an external copy rate of 61% (78.7% conditional on copying). In Section 5, when we estimate the returns to originality, we focus on external copy only.

Excluding the content copied from the news agencies When considering the returns to originality, one needs to distinguish between content copied from the news agencies and content copied from other media outlets. All the AFP' clients are indeed allowed to reproduce the AFP content in its entirety, and the business model of the news agency is based on the reproduction of its content by other media outlets (similarly for Reuters). We show that on average, documents include content from 2.3 documents published by competing media outlets other than the news agencies.[52] If we exclude content copied from the news agencies, we find that the average external copy rate is 15.9% (25.7% conditional on copying).

Share of the original story that is copied Finally, we compute the share of each document which is copied. On average, each document is copied by 3.9 documents, 3.3 if we exclude internal verbatim copying. If we focus on external verbatim copying and sum up the portions of the documents that are reproduced by at least one external media outlet, we find that on average the share of a document that is copied is equal to 9.4%. The majority of the documents are not copied, however. If we restrict our analysis to copied documents, we find that the share of a document that is copied by at least one external media outlet is 24.1% on average.

This share varies strongly depending on the publication rank of the document. Online Appendix Figure G.14 plots the average share of a document that is copied by at least one external media outlet depending on its publication rank. We find that for breaking news documents, this share is 60%, 25% when we exclude documents published by the news agencies. It then decreases to nearly 25% (12%) for the second document and converges rapidly to around 5%.

[Table 4 about here.]

3.3 Copying behavior and reputation

Overall, on average, media outlets tend to rely a great deal on plagiarism. But do all the news media display the same copying behavior? In this section, we first study the copying behavior

[52] In other words, while we note earlier that documents include content copied from 3.9 documents previously published in the event, this can be broken down as follows: on average, articles include content reproduced on the one hand from 1.6 news agencies' dispatches and on the other hand from 2.3 documents published by media other than the news agencies.

of the media depending on their type (newspapers, television stations, etc.). We then rank all the news outlets in our sample depending on their reliance on copying, and draw from this analysis a tentative typology of the media. Finally, we investigate the extent to which some outlets are more copied than others. The combination of a low reliance on copying and a high probably of being copied by other outlets may be considered as a proxy for the reputation of a given media outlet.

Copying behavior depending on the media type Does the copying behavior vary depending on the media type? We compute the average copy rate separately for local newspapers, national newspapers, television stations, radio channels, and pure online media. Figure 7 presents the results. It appears clearly that local newspapers tend to produce less original content online than other types of media outlets. A possible explanation is that, while local newspapers may compete on local news to attract readers, they may rely on copying regarding national news. Coherently with this assumption, we find that local newspapers tend to strongly rely on content produced by the news agencies.

We also find that pure online media tend to be on average more original than other media outlets. This may come from the fact that these pure online media *"seek to offer distinctive voices"* (Nicholls et al., 2016). Note however that pure online media only account for 3% of the documents in our dataset[53], and that we should not overstate these differences in editorial priorities. French pure online media indeed *"closely approximate an online newspaper"*, and most of the pure online media *"are committed to forms of professional journalism that are broadly the same as many legacy media"* (Nicholls et al., 2016).[54]

[Figure 7 about here.]

Classifying the media depending on their copying behavior While there is heterogeneity in the copying behavior of the media outlets depending on their type, are there also differences between news media within types? Figure 8 plots the external copy rate excluding content copied from the news agencies of each of the media outlets in our sample.[55] We rank the outlets in ascending order – from those which rely the least on external copy to those which rely the most – and organize them into four quartiles.[56]

[53] Moreover, only 25% of the pure online media documents are classified in events.

[54] As highlighted in Nicholls et al. (2016), *"all see themselves as offering news and focus in large part on public affairs."* Furthermore, Nicholls et al. (2018) note that *"Mediapart offers general coverage as a boutique online version of the print newspaper tradition"*.

[55] To compute this average, we weighted the articles by their size.

[56] Online Appendix Figure G.15 presents the same numbers but reports the error bars. Furthermore, in the online Appendix Figure G.16, we perform a similar analysis but at the media company level (i.e. rather than considering each of the media outlets in our sample separately, for each outlet, we first determine the identity of the company owning it, and then compute the average copy rate for each of these companies). Consistently with the results of Figure 7, it appears clearly that media companies specialized in radio – such as RTL Group

From this figure, we can classify the outlets into three different categories. First, the "niche media", e.g. *Le Monde Diplomatique* or *Témoignage Chrétien*, which aim to offer general news coverage but with a perspective that appeals to a relatively specific group of readers (e.g. *Témoignage Chrétien* aims to offer a Christian perspective; *Le Monde Diplomatique* a mostly alter-globalization one; etc.) These media do not rely on copying: the first four media in the first quartile have an average copy rate that is below 1%. Interestingly, nor do they rely on content produced by the news agencies: on average, less than 1% of the content of the articles published by Non Fiction, *Le Monde Diplomatique*, *Témoignage Chrétien*, *VSD*, *Courrier International*, Slate and France Culture comes from the AFP or Reuters. However, while the content produced by these media is mostly original, only a relatively small share of the articles they publish is classified into events and, as we will see below, they are hardly ever used by other outlets as a source of information. Moreover, these niche media tend to have a very small audience (e.g. around 28,000 visits a day for *Le Monde Diplomatique* compared to 2.9 million for *Le Monde*).

The second category is composed of the "mass media" (by opposition to the niche media previously defined) which tend not to resort too much to copying, e.g. TF1, *Le Figaro*, *Le Monde* or France Info. At least in relative terms, they may be considered media outlets with a relatively good behavior or "reputation". They need to be distinguished from our third category, composed of the mass media which tend to resort a great deal to external copying, e.g. RMC or *La Provence* and which can be considered "low-reputation" media outlets from the point of view of their copying behavior.

[Figure 8 about here.]

Classifying the media depending on their probability of being copied Needless to say, its copying behavior may not be sufficient to capture the reputation of a media outlet. In particular, high-reputation media may not only "behave well" but also receive relatively more attention from their competitors. Hence, we complement this copying behavior dimension with a second dimension: the probability of being copied. Figure 9a (respectively Figure 9b) reports for each of the media outlets in our sample their average copy rate[57], and the total content they produced that was copied by at least one other media outlet in 2013 (respectively the average share of the content they produced that was copied by at least one other outlet).

From Figure 9, we can classify the media into four broad categories: (i) the media outlets that do not rely on copy but are not used as a source of information by other outlets (bottom left corner); (ii) the media outlets that do not rely on copy but are used as a source of

and Lagardère Active – and in local news production – e.g. the Crédit Mutuel – rely on average more on copy-and-paste than companies owning pure online media or national weekly newspapers.

[57] Measured as before by the external copy rate excluding content reproduced from the news agencies.

information by other outlets (i.e. other outlets tend to copy their content relatively more) (top left corner); (iii) the media outlets that rely on copy and are not used as a source of information by other outlets (bottom right corner); and (iv) the media outlets that rely on copy but are used as a source of information by other outlets (top right corner). Even if such a classification raises a number of issues regarding how the "reputation" of a media outlet should be defined, one can consider the outlets in the top left corner as the high-reputation outlets and those in the bottom right corner as the low-reputation outlets. Included in the first category, whether we consider the total content or the share of content copied, are three national daily newspapers (*Le Monde*, *Le Parisien*, and *Les Echos*), two national weekly newspapers (*Le Point* and *Le JDD*), one local daily newspaper (*Ouest France*), two television stations (BFM TV and France Television), and two radio stations (RTL and France Info). Moreover, if we focus on the share of the content that is copied, we see a number of smaller media outlets appearing in the high-reputation category, in particular pure online media (Rue 89 and Arrêt sur Images). These small outlets produce little content (they tend to have a very small newsroom) but a relatively large share of the content they produce is reproduced by their competitors.

In the online Appendix Table F.6, we provide summary statistics for the media outlets depending on their "reputation" as defined using their reliance on copy and the use of their content by others. Interestingly, the 15 media outlets in the top left corner of Figure 9a (low reliance on copy and highly copied by others) have an average audience of $580,782$ unique visitors a day, compared to only $130,469$ visitors, i.e. more than four times less, for the outlets in the bottom right corner. Even if imperfect, the average annual audience of the different media outlets may be seen as an alternative proxy for the long-term reputation of a media. We find a similar difference (statistically significant at the five-percent level or more) if we consider the number of shares on social media. E.g. the average total number of shares on Facebook received by the "high-reputation" media outlets in 2013 is equal to $3,5$ million compared to $757,000$ for the "low-reputation" outlets. Obviously, these differences do not imply that reputation has a causal effect on sharing on social media; but it is interesting to note that overall in 2013, the media outlets that rely relatively less on copy and whose content is relatively more used by their competitors tend to have a higher audience and benefit from more reposting on Facebook and Twitter.

[Figure 9 about here.]

Finally, one may also proxy the reputation of the media outlets by the number of "citations" they receive (i.e. the number of times their competitors refer to them as the source of the information). We compute this measure in the next section.

3.4 Credit and citation patterns

In France, under certain conditions, media outlets are allowed to reproduce content originally published by their competitors, but the "right to quote" is subject to the mention of the source. In this section, we first compute the total number of citations received by each of the media outlets in 2013, and then study the extent to which the occurrences of verbatim copying we identified above come with acknowledgment. In other words, we analyze whether media outlets tend to name the outlets they copy.

On average, the media outlets in our sample received $8,054$ citations in 2013, $3,645$ if we exclude the citations received by the news agencies. Online Appendix Figure G.17 ranks the media outlets depending on the number of citations they received. 20 outlets received more than $5,000$ citations in 2013, among which all the national media outlets included in the top left corner of Figure 9a, with the sole exception of France Television. Among the most referenced media outlets one can also observe the presence of Mediapart, a pure online media that has broken some of France's biggest scandals involving politicians.[58]. Moreover, Mediapart hardly ever relies on content produced by others. From these viewpoints, it can be considered a high-reputation media outlet. The fact that despite the high number of references it receives it is not part of the most copied outlets in our sample may be rationalized by the use of exact verbatim copying to measure copy. The articles published by Mediapart tend to be much longer on average than those published on its competitors' websites; hence the need to reformulate for the media outlets that want to cover a story broken out by this pure online media.

If we focus on verbatim copying, to what extent do media outlets tend to name the outlets they copy? In Figure 10, we plot the share of the documents crediting the copied media depending on the copy rate. We do so both including and excluding the news agencies as a copied media. We find that not only do media outlets hardly evet name the media they copy, but that their propensity to do so scarcely increases with the extent of copying. Once the news agencies are excluded as a copied media, the share of crediting documents is always below 5%. If we also consider documents copied from the news agencies, we show that this share increases from 4% to 32% with the importance of copying. Media outlets most probably credit the news agencies more than the other outlets because they are not competing directly with them.

[Figure 10 about here.]

Note that the public exposure of a copying behavior may hurt the reputation of the

[58] E.g. at the end of 2012, Mediapart revealed that the French budget minister avoided paying tax in France on sums deposited in undeclared Swiss bank accounts. Following Mediapart's allegations, a tax legal investigation was opened into the tax fraud accusations and Jérôme Cahuzac resigned before being charged with tax fraud.

responsible journalist or the media she works for. Such exposure tends to increasingly take place on social media such as Twitter where journalists who are victims of plagiarism denounce these blameworthy behaviors, in particular when the original source is not quoted. This may explain the importance of maintaining a reputation for "good journalism" that avoids plagiarism matters.

4 Copying and newsgathering incentives: unraveling the mechanisms

In the previous section, we have quantified the speed of news propagation online and the importance of copying. What is the impact of copying on media profitability and newsgathering incentives? This is a key question, not only from an economic perspective but also because changes in the market for news affect political outcomes (see among others Cagé, 2017; Gavazza et al., 2018; Gentzkow, 2006; George and Waldfogel, 2006; Snyder and Stromberg, 2010).

From a theoretical perspective, the impact of copying on newsgathering incentives is relatively uncertain. One possibility is that readers are sparsely mobile across media outlets. If so, being original or being copied should have little impact on a media outlet's audience.[59] Another possibility is that readers are mobile and shop for the best news across media outlets. This potentially raises the incentives for original news production, but this also makes copying more problematic. An important parameter is the extent to which copied articles are of lower quality than the original (which will depend on copyright law and other factors). In some cases, copies could also benefit original news articles through an exposure or sampling effect, a mechanism that has been well documented in the piracy literature (see e.g. Peitz and Waelbroeck, 2006a,b).

Finally, in order for copying to harm original news producers through an audience-stealing mechanism and thus decrease their incentives to invest in original news production, the media outlets' investment in newsgathering should be driven by an audience motive, e.g. because audience is positively correlated with advertising revenues and thus profitability. But the objective function of media outlets may include other non-profit driven motives.

In this section, we first document the extent of consumers switching across media outlets. We then discuss readers' valuation of originality, and the limits of the sampling effect argument in the context of the news media industry. Finally, we examine the link between audiences and the objective function of the media outlets. In the next section, we use article-level variations to provide estimates of the returns to originality in terms of audience.

[59] Note that in the extreme case that all consumers were loyal to their preferred media and there was no switching between media outlets, then the demand for each of the media would not depend on their investment in newsgathering and the media would have no incentive to produce investigate journalism.

In the online Appendix Section A, we present a simple theoretical framework on copying and returns to originality whose objective is to highlight the mechanisms through which copying may negatively affect newsgathering incentives and aid interpreting the empirical results. The three key parameters of this simple framework are the consumers' loyalty to a particular media, the consumers' taste for originality, and the quality of the copy with respect to the original. Consumers are heterogeneous with respect to their taste for originality, and face a trade-off between their loyalty to their preferred media and their taste for originality, depending on the quality of the copy. As long as the ratio of the average taste for originality over loyalty is high enough compared to the relative quality of the copy, at least some readers will switch across the media. Furthermore, we show that when media outlets are more "isolated", there are lower returns to originality, a prediction that we will test in the next section, where we use media-level daily audience and article-level social media statistics to quantify the returns to originality. We summarize below some of the main forces and parameters at play, and refer the reader to the online Appendix for this very simple theoretical framework.

4.1 Readers' mobility across the media

Recent studies of audience news consumption behavior have indicated that news users increasingly rely on multiple news media (see e.g. Pew Research Center, 2016; Reuters Institute, 2017). Given that *"people have more power to navigate the news content they want to use, when, where and how"* (Swart et al., 2017), they seem to shop for the best news across outlets online. As a consequence, they follow the news on multiple media platforms (Picone et al., 2015; Yuan, 2011). In a context where copying online is substantial, such a behavior may be rationalized by the imperfections of the copy (see sub-section 4.2 below), and by the fact that the ratio of the average taste for originality over consumers' loyalty is high enough compared to the relative quality of the copy. It has been well-documented that the Internet has reduced loyalty to any one outlet, in particular for technological reasons (Athey et al., 2013). It is revealing that online, when coming on a news website from search or social media, most of the users cannot recall the name of the website's news brand after their visit (Reuters Institute, 2017).

In this paper, we use survey data on patterns of online readership to document the extent of readers' mobility in France. We rely on survey data from the *2013 Digital News Report* (Reuters Institute, 2013).[60] The sample includes 1,016 individuals for France for the year 2013. Among the survey questions, respondents are asked whether they followed different media outlets online.[61] Out of the 9 television channels included in our sample, 5 are covered

[60] Similar data has been used by Kennedy and Prat (2019) in their study of news consumption across platforms in 2015.

[61] *"Which, if any, of the following have you used to access news in the last week via online platforms (web, mobile, tablet, e-reader)?"*.

by this question regarding online news consumption[62]; 13 national newspapers[63] (out of 24); and 5 pure online media[64] (out of 10). Furthermore, radio stations in our sample are gathered into two categories: private radio (RTL, Europe1, RMC, etc.) and public radio (France Inter, France Culture, France Info, etc.). Finally, from the "other" category, we compute a measure of the online consumption of local newspapers.

If we first consider the 26 media outlets (considering "public radio", "private radio", and "local newspapers" as a media outlet) for which we have information on consumption via online platforms, we see that nearly two thirds of the surveyed individuals consume at least one media outlet online. Among those who consume at least one news media, the average number of outlets consumed is equal to 2.35; in other words, users spread their news consumption over multiple platforms online.

The most popular media outlet is *20 Minutes*, with a 17.8% reach, followed by *Le Monde*, *Le Figaro*, and *TF1*. The least popular brands in terms of penetration are *La Croix*, Slate and Atlantico (see online Appendix Figure G.18). How many other media outlets do the respondents who consume each of these media also access online? We compute this number for each of the news media considered separately. Figure 11 presents the results. While the extent of competition varies depending on the media outlets, the first important thing to note is that none of the media outlets in our dataset seem to work "in isolation", with captive users. Even respondents who consume news from TF1 – which is the "most isolated" media outlet in our sample – access news on average from three different online platforms, i.e. consume news online from two media outlets other than TF1. Hence all the media outlets may suffer from consumer switching when their content is copied and pasted by their competitors.[65]

[Figure 11 about here.]

We use this survey data to build a matrix of proximity across media outlets. Figure 12 presents the results visually: the size of the circles represents the audience of the media outlets (the larger the circle, the higher the audience), and the size of the rows between two websites shows the probability that a respondent accessing one website also accesses the other (the thicker the row, the higher this probability). Three different media ensembles appear: there are depicted in red (including *Le Monde*, *Libération*, Mediapart, etc.), green (*20 Minutes*, France TV, the private radio, etc.), and blue (*Le Figaro*, *Le Point*, *L'Express*, etc.). In other words, respondents reading *Le Monde* also tend to consume news from *Libération*,

[62] TF1; BFM TV; I>TELE; LCI; and France Television.

[63] *Le Monde*; *Le Figaro*; *Libération*; *Les Echos*; *La Croix*; *20 Minutes*; *Metrofrance*; *Direct Matin*; *Le Point*; *Le Nouvel Observateur*; *L'Express*; *Marianne*; and *Courrier International*.

[64] Slate; Atlantico; Rue89; Médiapart; and the Huffington Post.

[65] Note also that on average more successful media outlets (as measured by the percentage of weekly usage) seem to be more isolated than less successful ones.

while readers of TF1 rely more on *Direct Matin* or private radio as an alternative source of information.

[Figure 12 about here.]

While, as we just saw above, none of the media outlets work in isolation, some media outlets (such as *Le Monde*, *Le Figaro* or *L'Express*) are in a more "competitive" environment than others. Another way to see it is to compute a correlation matrix where we use the consumption pattern of each of the survey respondents and report pairwise correlations between the online consumption of each of the media outlets in our sample (online Appendix Figure G.19 and Table F.7). Consistently with our previous findings, while the probability of consuming *Libération*, Slate, *La Croix*, or *Marianne* is strongly correlated with the probability of consuming other media outlets online (e.g. *Le Monde* for *Libération* and Slate), media outlets such as TF1 or BMFTV are more "isolated" online.

We highlight these differences in terms of the "competitiveness" of the environment in which the media outlets work online, because the competitive environment of an outlet may affect its returns to originality. One may indeed expect more "isolated" outlets to have lower returns to originality than outlets that are more subject to consumers' switching. This is a prediction of our very simple theoretical framework: when media outlets work more in isolation – consumers' loyalty to their preferred media is higher – then the returns to originality are lower. The intuition for this result is as follows: in terms of audience, the returns to originality for a media depend on the share of the consumers who are loyal to its competitors but nonetheless switch to its website to read the original articles. This share decreases with the strength of loyalty. We show that this is indeed the case in Section 5.2.4 below when investigating heterogeneity in media outlets' returns to originality.

4.2 Valuation of originality

Hence, according to the survey data on patterns of online readership, and in line with previous findings of the existing literature, consumers seem to be mobile across media outlets. This implies that they will shop for their preferred news across media outlets. For such a consumption pattern to have a negative impact on media incentives to provide original content, at least some readers need to be indifferent as to whether they consume the copy or the original, despite the fact that the copied articles may be of lower quality than the original one.[66]

We try to capture these two opposite forces in the very simple theoretical framework we present in the online Appendix to this paper. We proxy the fact that the copy is of lower quality than the original by a parameter $\lambda \in]0, 1[$. Despite this lower quality, a fraction of

[66] If all the consumers were indifferent between the copy and the original, then they will all read their preferred media independently of the identity of the news breaker, and the returns to originality will be null.

the consumers read the copy rather the original due to their "loyalty" to the media publishing the copy, a parameter we call $\bar{u}$ and which corresponds to the utility consumers derive from reading their preferred media (e.g. because it is better fitted to their political stance or they prefer the tone of voice used). In our very simple theoretical framework with only two competing media outlets, A and B, and a continuum of consumers i of mass one, a consumer i loyal to media A will read the copy rather than the original published on the website of media B iff $\bar{u} + \lambda v_i > v_i$, where v_i is consumer i's taste for originality. If we assume that v is uniformly distributed with unit density over the interval $[0, 2\bar{v}]$, where $\bar{v}$ is the average taste for originality, then the fraction of switchers is given by $1 - \frac{1}{2(1-\lambda)}\frac{\bar{u}}{\bar{v}}$, which can be interpreted as the "returns to originality" for media B. The higher the quality of the copy with respect to the original, the lower these returns.

Why are the copied articles of lower quality? In the piracy literature, the original digital product is often considered of higher quality than the copy, in particular because it is bundled with other non-digital components, e.g. a printed manual for software or a CD case for music CDs (see Bae and Choi, 2006; Peitz and Waelbroeck, 2006a). In the case of the news media, following a similar line of reasoning, we can say that the original is of higher quality than the copy because it is bundled with additional information that is absent from the copy. First, copying media outlets tend not to reproduce the articles they copy in their entirety. On average, when an article is copied, "only" 9.4% of its content is reproduced; it means that nine-tenths of the content of the original article is missing from the copy.

Second, online, articles tend to be published along with photographs, videos or other kinds of illustrations, e.g. data visualizations, visual stories and graphics. In this article, we only consider text. However, having "manually" analyzed the websites of a number of media outlets and discussed the question with a number of publishers, our educated guess is that while plagiarism is a common practice regarding text, it is very uncommon to reproduce alongside the illustrating images, in particular when it turns to visualizations. While text plagiarism falls in a grey area in terms of copyright enforcement (due to the right to quote exemption, the issue of the substantiality of copying and of the originality of the copied work, etc.), photos and data visualizations are more clearly copyrighted, and it is much easier to identify the infringement. In other words, the original may be of higher quality than the copy because the original is bundled with photographs, visual stories and graphics, etc. that are absent from the copy.

Moreover, an article is not published in isolation on the website of a media outlet. Most often, the reader can find links to "Related coverage" on the outlet's website, i.e. articles dealing with the same broad topic and of potential interest to the reader. Sometimes, media outlets also offer a list of additional content "Recommended for you". This related content may be more relevant when provided by the news breaker that has invested in newsgathering

and whose journalists may have a better sense of what the event is about, than when provided by the copying outlets.

Finally, the copied articles may be of lower quality than the original ones if copy is a manifestation of lousy journalism. While we do not measure the "quality" of the articles in this paper, we may nevertheless assume that on average news article that contain copy-and-paste material may be poorly written compared to original news articles. In other words, the degree of copying and other quality characteristics of the news articles, in particular in terms of writing, may be positively correlated.

The above arguments help to rationalize the positive relationship we describe below between originality and news consumption as proxied by the number of shares on social media. Consumers may favor original content over copy because the original content is of higher quality. Moreover, original news producers may also benefit from an increase in their audience through a sampling effect. In the context of the music industry, this effect corresponds to the fact that *"downloaders use the downloaded files for sampling in order to make more informed purchasing decisions"* (Peitz and Waelbroeck, 2006b). In the case of the news media industry, this could take the form of readers discovering a new media outlet through reading its original content reproduced on the website of its competitors. However, for such a positive impact of copy to be at play, the copying media outlets should mention the identity of the news media where the copy comes from. Yet, we see in Section 3.4 that it tends not to be the case. Note finally that despite consumers' valuation of originality, some readers may consume the copy as long as the ratio of the average taste for originality over the consumers' loyalty to media brands is high enough (in our very simple theoretical framework, $\frac{\bar{v}}{\bar{u}} > \frac{1}{2(1-\lambda)}$). Ultimately, quantifying the returns to originality is an empirical issue.

4.3 Audience, advertising revenues, and the objective function of the media

Web traffic and advertising revenues In the next section, we estimate the economic returns to originality in terms of audience. We do so because these returns can, as a first approximation, be assumed to be proportional to audience, in particular via online advertising revenues. As highlighted by Anderson (2012), *"the core business model for effective financing of web content for many sites is through advertising"*. Advertising is indeed the largest contributor to publishers' online revenues.[67] Yet, advertising pricing on the Internet is based on audience (see e.g. Anderson, 2012; Zhu and Wilbur, 2011).[68] According to Deloitte (2016), a

[67] Another source of revenues being subscriptions but, with the exception of Mediapart, the pay models were not very developed online in France in 2013. Note moreover that, even if pay models are becoming an important part of the business of digital news nowadays, in most countries including France, there is still only a minority of news lovers who pay for online news (Cornia et al., 2017). We come back to the issue of the external validity of our findings in Section 6.3.

[68] Either through measures of expected impressions (the so-called cost-per-thousand views or cost-per-mille rates) or through the actual performance of ads (pay-per-click or cost-per-click pricing).

10% increase in overall web traffic to newspaper publishers' sites leads to an estimated 0.64% increase in their overall revenues. The average value of a web visit is estimated between €0.04 and €0.08.[69]

Note moreover that even public-service broadcasters in France, while they are mostly funded through license-fees, also depend on advertising funding.[70] In other words, audiences also enter their objective function.

Non-market profits and the objective function of the media That being so, the objective function of public-service broadcasters of course includes other dimensions than the audience size alone. To begin with, public-service broadcasters, unlike their commercial competitors, have public service obligations. Where private media companies work primarily in the interest of their shareholders, public-service broadcasters *"are obliged to serve the whole society by enhancing, developing and serving social, political and cultural citizenship"* (Digital Strategy Group of the European Broadcasting Union, 2002). In particular, there may be a trade-off between investing in program variety serving many audiences and investing in quality popular content for large-scale audiences. In other words, the public service motive entering the public-service broadcasters' objective function may come at the expense of audience size maximization. Moreover, public-service broadcasters may decide to invest in original content as part of their public-service mission despite low returns of originality in terms of audience. While the share of original content in our dataset is equal to 32.6% overall, it is equal to 31.6% for private media but to 39.5% for public media.

Furthermore, note that public-service broadcasters are not the only media outlets whose objective function may include other dimensions than the monetary motive. In particular, a number of media outlets may have political motives entering their production function: media owners may derive utility from influencing the political tastes of their readers. In the theoretical literature, a number of important papers have introduced political motives in the objective function of media owners (Duggan and Martinelli, 2011; Anderson and McLaren, 2012; Balan et al., 2014).[71] But if media owners care about consumer actions then, despite the economic losses due to the decrease in readership caused by online copying, they may find the verbatim copying of their content useful for spreading their editorial line across different outlets. To use the terminology of Prat (2018), verbatim copying may in this case increase their "power", as defined by their ability to induce voters to make electoral decisions through biased reporting.

[69] The Deloitte (2016) study uses data from 66 newspaper publishers with both online and offline publications in France, Germany, Spain, and the UK between 2011 and 2013.

[70] The mix between license-fees and advertising funding is not specific to France; it also characterizes the funding of public-service broadcasters in countries like Germany and Ireland.

[71] See also Gentzkow et al. (2015a) for a review of the theoretical literature on the market determinants of media bias.

To what extent are the political motives more important than the profit motives for media owners? Empirically, the jury is still out. While Gentzkow and Shapiro (2010), in the first large-scale empirical study of the determinants of political slant in the news, find little evidence that the identity of a media outlet's owner affects its slant, Puglisi and Snyder (2011), studying the coverage of political scandals by newspapers in the United States between 1997 and 2007, find evidence of some supply-side factors driving bias. Estimating the importance of the political aspirations of the media in our sample is beyond the scope of this paper. However, we can distinguish between different types of media outlets. In particular, the identity of the owner may directly affect the importance given to the non-monetary motives in the objective function. Hence, we classify on the one hand the media outlets whose owners are media companies, i.e. companies that have their core activity in the media industry, and on the other hand the media outlets that are owned by companies involved in other activities, either commercial or industrial, constituting their main sources of revenue. We call the first category of owners the "media owners" and the second category the "non-media owners". Our assumption is that owners that have their core activity in sectors other than the media – the "non-media owners" – may care relatively less about monetary profits (given that the media are not their main source of revenues) and be more motivated by political motives[72] than the "media owners".

While the different weight put on the monetary vs. the political motives in the objective function of the media owners should not affect their returns to originality in terms of audience and monetary profits, it may nonetheless impact their incentives to produce original content. In particular, everything else being equal, despite the audience-cost of copying, owners with political aspirations may prefer to produce original content to spread their editorial line across different outlets through the verbatim copying of their content. However, this preference for originality should only happen for "political content" that they can slant and that is likely to affect political opinions, but not for "softer content" like "human interest" content or weather-related content, to use the classification of our events by topics (see Section 2.3). To investigate whether this is the case, we compute the average originality rate of the articles in our sample depending on the nature of the owner (media or non-media owner) and on the topic of the event in which the article is classified. Online Appendix Figure G.20 presents the results. While for weather events, the average originality rate of the articles published by the "media owners" is higher than that of the articles published by the "non-media owners" (as well as for "lifestyle and leisure" events, but the difference is not statistically significant), it is much lower for articles dealing with "Religion and belief" (the average originality rate of the articles published by the "non-media owner" is 8.5 percentage-point higher than that of the articles

[72] This will be the case in particular if their media acquisition has been driven by a political agenda (see e.g. Cagé and Gadenne, 2015).

published by the "media owners"), "Politics" (7.5 percentage-point gap), and "Sciences and technology"[73], as well as for "Economy, business and finance" (4.1) and "Labour", albeit to a lower extent. These results suggest that, compared to "media owners", "non-media owners" seem to invest more in originality but mainly for the articles that may affect (broadly speaking) consumer actions from a political point of view. On the contrary, the economic motive may dominate (hence a higher reliance on copy-and-paste) for articles related to topics with no political dimension.

Obviously, these findings are only suggestive and, in the remainder of the analysis, we will mainly focus on the monetary profits of the different media outlets (measured by their audience), and abstract from the political motives (given that our regressions include event, date, and media fixed effects, this omission should not affect our estimations of the returns to originality in terms of audience). But it is nonetheless important to bear in mind that the media may take non-monetary factors into account when deciding the amount of original content they want to produce. Furthermore, it is also important to highlight that, even if the media were entirely profit-driven, monetary profits can take a number of different dimensions. E.g. Gentzkow et al. (2015b), in their study of the effect of party control of state governments on the press in the United States between 1869 and 1928, assume that newspaper profits come from two sources: market profits (that increase with audience) and a political bounty paid to any newspapers affiliated with the party in power. Similarly, Besley and Prat (2006) assume that the media face two possible sources of profit: commercial profits and profits from collusion with government. Unfortunately, we cannot measure the importance of this second kind of monetary profits with the data we have. Note however that overall, despite these limitations, the number of reads plays an important role in the objective function of the media – in particular through their impact on advertising revenues online – and can as a first approximation be used as a relevant measure of the returns to originality.

Propagation of audience Finally note that, while publishing an original article may allow a media outlet to attract additional readers (e.g. because readers have a preference for original content), it may also generate an higher audience for the other articles published on its website. The reading of a given article may indeed generate the reading of other articles on the website of the same media outlet. We obtain data from Similar Web (a website traffic statistics company) on the average number of pages visited during a visit on the websites of 37 out of the 85 media outlets in our sample in 2013. According to these data, on average, once on a website, readers read 3 articles (the median is 2.7). In other words, one may assume that attracting a reader by one news article may generate the reading of two additional articles for a media outlet.

[73]However, there are only 134 events related to this category.

In sum, online copying may have a negative effect on the copied media's market profits because of audience stealing and associated losses in advertising revenues. But original news producers nonetheless partly benefit from their investment in newsgathering, on the one hand owing to non-market profits (public service motive and political aspirations), and on the other hand through an audience effect. First, consumers whose taste for originality is high enough switch across the media (we document empirically that news users rely on multiple news media), providing original news producers with additional viewers. Second, long-term reputation provides another way to understand why media invest in information production (Gentzkow and Shapiro, 2008). In the next section, we investigate the relationship between the production of original information and the audience of the websites.

5 Online audience and the returns to originality

This section provides tentative estimates of the returns to originality. Unfortunately, our main dataset does not include article-level information on the number of visitors, but only aggregated information on web traffic at the daily level for the media outlets (all articles combined). This is an important limitation of our dataset of which we are fully aware. We attempt to overcome it by using alternative article-level information that we collect from two different social media, namely Facebook and Twitter. Furthermore, we use an additional dataset to relate article-level Facebook and Twitter shares and article-level numbers of views.

In this section, we first document the relationship between the number of times an article is viewed and the number of times it is shared on social media (5.1). We then provide estimates of the returns to originality using our different proxies for article-level audience (5.2). Finally, we compute an audience-weighted measure of the importance of original content (5.3), and provide some orders of magnitude as to the returns to originality (5.4).

5.1 Social media statistics and number of views

5.1.1 Evidence from *Le Monde*'s data

What is the relationship between the number of times an article is viewed and the number of times it is shared on Facebook? Answering this question is of particular importance for us given that our approach uses this relationship to compute number of views per article statistics.[74] To understand the mapping between article views and number of Facebook shares, we use data from the French daily newspaper *Le Monde*. More precisely, we obtained access to data on the number of views for each article published by *Le Monde* between April

[74] A number of articles in the literature simply assume that exposure is proportional to Facebook shares (see e.g. Allcott and Gentzkow, 2017). However, such an assumption can be questioned and this is why we made the choice here to document this relationship empirically.

and August 2017, as well as the URL of the articles. We use the URL to compute as before the number of shares on Facebook. On average, during this time period, each *Le Monde*'s article is viewed by $19,656$ unique visitors and shared $1,015$ times on Facebook.

Online Appendix Table F.8 provides detailed summary statistics for these two variables. Both distributions are skewed to the right, and have heavy tails (with a large Kurtosis). The skewness of the distribution is higher for the number of shares on Facebook. This also appears in the online Appendix Figure G.21 that plots the skewness function versus the spread function for these two variables; both distributions are remarkably right skew (note that the right skewness of the shares and views distributions is not specific to *Le Monde*'s data, and is the reason why we apply a logarithmic transformation when estimating the returns to originality below).

Figure 13 plots the relationship between the number of views and the number of shares on Facebook at the article level for the $17,314$ articles published by *Le Monde* between April and August 2017 (sub-Figure 13a). Specifically, we characterize the joint distribution of the number of Facebook shares and the number of unique visitors at the daily level, and use a rank-rank specification with 20 quantile categories. We find that the relationship between the number of views and the number of shares is almost perfectly linear. A 10 percentile point increase in the number of Facebook shares is associated with a 7.3 percentile point increase in the number of views on average.[75] Hence, for each article a published by the media n on a given date d, we can use its Facebook rank (P_{FBadn}) to compute its rank in the number of visitors distribution (P_{Vadn}). This relationship can be summarized with only two parameters: a slope and an intercept.

Given that in our main dataset we only have aggregated information on the total audience at the daily level for each media outlet, the second step consists in investigating the average number of visitors in each rank of the number of visitors distribution. For each article a published by media n on date d, we normalize its number of visitors (V_{adn}) by the average number of visitors received by the articles published by the media outlet on this given date ($\overline{V_{dn}}$). We call this ratio R_{adn} ($R_{adn} = \frac{V_{adn}}{\overline{V_{dn}}}$). We then compute the average value of this ratio ($\overline{R_{adn}}$) for each rank of the distribution. Figure 13b shows the results. We approximate the relationship between the rank in the number of visitors distribution (P_{Vadn}) and the average number of visitors (as a multiple of the mean number of daily visitors) by a polynomial of degree six (so as to obtain the best possible fit). We also use alternative non-linear specifications and show that this has a limited impact on our main results (see below).

[Figure 13 about here.]

[75] In the online Appendix Section J.4, we provide additional evidence of such a linear relationship using a dataset we obtain from Parsely, a technology company that provides web analytics. This dataset covers the year 2017 and contains information on the number of clicks and on the number of shares on social media originating from the United States for $1,363,308$ articles published in English.

5.1.2 Article-level estimation of the audience

As we already highlighted, we do not have information in our main dataset on the number of article-level visitors. To offset this downside of our data, we develop a number of strategies combining information on the daily number of page views (equivalently of the number of articles read) with our social media statistics (the number of Facebook shares and the number of Tweets) available at the article level. This allows us to obtain an article-level estimation of the audience.

Naive (media-level) approach From the content data, we know on a daily basis the total number of articles published by each media outlet. If, on a given day, all the articles published on the website of an outlet were "equally successful", then to obtain the number of views per article we would just have to divide the total number of page views by the number of articles published (naive approach). This is the first approach we follow.

Social media approach, assuming linear relationship All the articles are not equally successful, however. We use the information on the number of Facebook shares (respectively on the number of Tweets) to obtain a less naive measure of the audience of each article. More precisely, we compute for each media/day the total number of Facebook shares (total number of Tweets) and then attribute a number of views to each article as a function of its relative number of shares on Facebook (relative number of Tweets). We do this both by using the raw number of Facebook shares (of Tweets) and the winsorized version of the variable.

Obviously, this approach is also imperfect: e.g. even those articles that are not shared on Facebook (on Twitter) may nonetheless attract some views. Moreover, as we saw in Section 2.7, readers who share articles on social media tend to be younger than those who do not. Finally, this approach relies on the assumption that the relationship between the number of shares on Facebook (on Twitter) and the number of article views is linear.

Social media approach, using estimates from *Le Monde* (rank-rank approach) Hence, to improve our estimation, we ultimately use the estimated parameters from *Le Monde*'s article-level data to approximate the number of views of each article. For the sake of robustness, we use two different methodologies: a rank-rank approach and a blinder approach simply regressing the share of the total number of daily views represented by each article on its share of the total number of Facebook shares.

The rank-rank approach relies on the findings described above (Section 5.1.1). First, for each article, we compute its rank in the Facebook shares distribution (P_{FBadn}) and then use the estimated coefficients from *Le Monde* (slope equal to 0.73 and intercept equal to 14.20) to impute its rank in the number of visitors distribution ($\widehat{P_{Vadn}}$). Then, from the total number

of views received by the media outlet n on date d, we estimate the number of views of each article by using the parameters obtained when estimating the following relationship using *Le Monde*'s data: $\overline{R_{adn}} = \alpha + \beta_1 P_{Vadn} + \beta_2 P_{Vadn}^2 + \beta_3 P_{Vadn}^3 + \beta_4 P_{Vadn}^4 + \beta_5 P_{Vadn}^5 + \beta_6 P_{Vadn}^6 + \epsilon_{adn}$. Doing so, we obtain an estimated value of the number of views received by each article.

Social media approach, using estimates from *Le Monde* (non-linear shares-shares approach) As an alternative non-linear strategy, still using *Le Monde*'s data, we perform the following estimation:

$$\begin{aligned}\text{Share Visits}_{adn} = \delta + \gamma_1 \text{Share Facebook}_{adn} + \gamma_2 \text{Share Facebook}_{adn}^2 + \gamma_3 \text{Share Facebook}_{adn}^3 \\ + \gamma_4 \text{Share Facebook}_{adn}^4 + \gamma_5 \text{Share Facebook}_{adn}^5 + \gamma_6 \text{Share Facebook}_{adn}^6 + \epsilon_{adn}\end{aligned}$$

where $\text{Share Visits}_{adn}$ is the share of the total views received by media n on date d represented by article a, and $\text{Share Facebook}_{adn}$ is similarly the share of the total number of Facebook shares received by media n on date d represented by article a. We use the estimated parameters to compute in our main dataset the number of views received by each article from the number of times it has been shared on Facebook.

5.2 Originality and news use across social media platforms: article-level estimation

To estimate the returns to originality using article-level estimations, we proceed in three steps. First, we investigate how the number of times an article is shared on Facebook varies with its originality and reactivity. Second, we perform a similar estimation but using the number of Twitter shares. Third, we consider our predicted number of readers per article. Finally, we investigate whether the returns to originality vary depending on the characteristics of the media outlets, which allows us to rationalize the positive relationship between originality and news consumption we unravel.

5.2.1 Number of Facebook shares

We use article-level data to investigate how the number of times an article is shared on Facebook varies with its originality and reactivity. Given that the distribution of the number of Facebook shares is right-skewed, we perform a log-linear estimation. Equation (1) describes our preferred identification equation (the observations are at the article level):

$$\text{Facebook shares}_{aedn} = \alpha + \mathbf{Z'_{aedn}}\beta + \lambda_e + \gamma_n + \delta_d + \epsilon_{aedn} \quad (1)$$

where a index the article, n the media, e the event and d the publication date of the article (an event can last more than one day), and we use the log of the dependent variable.[76]

$\mathbf{Z}'_{\mathbf{aedn}}$ is a vector that includes the characteristics of the article a published by media n on date d and included in the event e. λ_e, γ_n and δ_d denote fixed effects for event, media outlet and date, respectively. In other words, we use within media outlet-event-date variation for the estimation. Standard errors are clustered by event.

The vector of explanatory variables includes (i) the publication rank of the article (the rank of the breaking news article is equal to 1, it is equal to 2 for the article published next in the event, then to 3,...); (ii) the reaction time (which is equal to 0 for the breaking news article and is then a measure of the time interval between the publication time of the considered article and that of the breaking news article); (iii) the originality rate of the article (in percentage: the variable varies from 0 to 100%); (iv) the length of the article (total number of characters in thousands); (vi) the original content (also by number of thousand characters); and (vii) the non-original content. Alternatively, we use an indicator variable equal to one for the breaking news article, and to zero otherwise, and then only control for the length of the article. Regarding the rank and reactivity measures, we are expecting a negative sign for the estimated coefficients: by construction, the higher the reaction time, the longer it takes the media to cover the event (similarly for the publication rank). In contrast, we are expecting a positive sign for our measures of originality (the originality rate and the original content), as well as for the breaking news indicator variable.

Colums (1) to (3) of Table 5 present the results. Regarding originality, we find that an increase of $1,000$ in the number of original characters leads to a 22% increase in the number of Facebook shares. If we instead consider the originality rate, we show that a 50-percentage-point increase in the originality rate of an article (e.g. moving from an article with no original content to an article with 50% originality) leads to a 40.5% increase in the number of Facebook shares. If we now turn to reactivity, we find that both the publication rank and the reaction time matter. The effect is economically small, however: taking 41 hours (which is about the average length of an event) to cover an event rather than writing about it from the beginning decreases the number of Facebook shares by around 10.9%. Yet, we observe high returns from being the news breaker: according to our estimates, being the breaking news article more than doubles the number of Facebook shares received by an article.

[Table 5 about here.]

Robustness In order to take into account nonlinear effects, we define 20 categorical variables depending on the originality rate of the articles (less than 5%; between 5% and 10%;...;

[76] More precisely, because the number of Facebook shares can take a value of zero, we use the log of (1 + Facebook shares).

between 95% and 100%). We then estimate equation (1) using as independent variables these categorical variables rather than the continuous originality rate measure. Figure 14 plots the estimates of the coefficients from the specification (articles with an originality rate lower than 5% are the omitted category). The results show that the number of times an article is shared on Facebook increases continuously with the originality rate of the article. Articles whose originality rate is between 25% and 40% receive twice as many shares on Facebook than articles for which it is below 5%.

[Figure 14 about here.]

Equation (1) uses the publication rank of the article as a measure of reactivity. However, different news events exhibit a different number of articles; hence a publication rank of 10 means something different for a news event with 10 or 100 articles. To deal with this issue, we run a robustness check where rather than using the absolute rank of the articles in the event, we use their percentile rank (with 20 quantile categories). Online Appendix Table F.9 presents the results: moving from the 5th to the 10th percentile rank of the publication distribution decreases the number of times an article is shared on Facebook by 0.61 to 0.65% depending on the specification. Moreover, the effect is statistically significant at the one-percent level, and the coefficients on the different measures of originality are unchanged.

Finally, as an alternative strategy to deal with the skewness of the Facebook shares variable distribution, we use a winsorized version of the variable at the 99th percentile. We then perform a linear estimation. Online Appendix Table F.10 presents the results which are consistent with the ones we obtain when performing the log-linear estimation. E.g., we show that a one-thousand increase in the number of original characters leads to 11.6 additional shares of the article on Facebook (the effect is statistically significant at the one-percent level).

5.2.2 Number of Twitter shares

As a measure of the returns of original news production, the number of times an article is shared on Facebook suffers from a number of caveats, in particular the fact that this number is partially filtered through the Facebook News Feed algorithm. While we cannot directly correct for this filtering, we show that our findings are robust to the use of other proxies for individual readers' demand, namely the number of shares on Twitter. Columns (4) to (6) of Table 5 present the results of the estimation of equation (1) where rather than considering the number of Facebook shares as the dependent variable, we use the number of shares on Twitter.

The results we obtain are consistent with the findings using the number of Facebook shares. On the one hand, social media audience increases with the number of original characters: an

increase of 1,000 in the number of original characters leads to a 11.4% increase in the number of Tweets. If we instead consider the originality rate, a 50-percentage-point increase in the originality rate of an article leads to a 17.3% increase in the number of Tweets. Moreover, as before, both the publication time and the publication rank matter regarding reactivity. E.g., an increase by one in the publication rank leads to a 0.2% decrease in the number of shares on Twitter.

We have constructed the number of times an article is shared on Twitter variable as the sum of different measures (number of direct tweets, number of direct reweets, number of direct likes, etc.) By aggregating these correlated measures, we may overemphasize the extent to which a story is likable on Twitter. In the online Appendix Table F.11, we show as a robustness check that the magnitude and statistical significance of the coefficients is unchanged if we instead consider independently as a dependent variable the number of (direct) tweets (Columns (1) to (3)), the number of (direct) retweets (Columns (4) to (6)), and the number of likes (Columns (7) to (9)). E.g. we find that a 50-percentage-point increase in the originality rate of an article increases the number of times it is tweeted by 13.3%.

No more than the number of shares on Facebook, the number of Tweets is a perfect measure of the audience of an article. However, the consistent findings we obtain by using both measures seem to reveal the fact that consumers favor original content and reactivity. In Section 5.3 below, we combine social media and audience statistics to build an audience-weighted measure of the importance of original content.

5.2.3 Predicted number of readers

Finally, in Columns (7) to (9) of Table 5, we present the results of the estimations when we use the number of times an article is viewed (using the Facebook approach detailed above) rather than the number of shares on social media as a dependent variable. Although this measure is imperfect – it is a predicted measure of the number of readers based on the estimates we obtain from *Le Monde* data rather than the actual number of readers – it may be considered as the more telling variable to estimate the returns to originality in terms of audience.

The signs of the coefficients are consistent with those we obtain for the number of shares on Facebook and on Twitter. In terms of magnitude, an increase of 1,000 in the number of original characters leads to a 23.2% increase in the number of times this article is viewed, and a 50-percentage-point increase in the originality rate of an article leads to a 44.8% increase in this number.

5.2.4 Heterogeneity of the effects

In this section, we investigate whether the returns to originality vary depending on the characteristics of the media outlets and of the events they cover. We consider different dimensions

of heterogeneity: first, the competitiveness of the media environment; second, the extent to which the media outlets are copied by other outlets; and finally, the topic of the events (e.g. sport or economy) and their "general interest". Doing so allows us to improve our understanding of the mechanisms at play behind the positive returns to original news production.

Competitiveness of the media environment We estimate equation (1) with an interacted "high competition" indicator variable equal to one for the media outlets that are in a "more competitive" media environment and to zero for those that are in a "less competitive" media environment. The competitiveness of the environment is measured with respect to the average number of other media outlets consumed by the readers who access a given media (see Figure 11 above).[77] In the spirit of the very simple theoretical framework we present in the online Appendix to the paper (Section A), a highly competitive environment is an environment in which $\bar{u}$ – the utility users derive from consuming a specific media – is low, while a less competitive environment is an environment where consumers' loyalty to certain media brands is high. Obviously, as we have highlighted, none of the media outlets is "in isolation" online, but it is nonetheless of interest to exploit the heterogeneity of their competitive environment.

Table 6 presents the results (in columns (1) and (2) we report the number of Facebook shares, in columns (3) and (4) the number of Tweets, and in columns (5) and (6) the predicted number of views). Regardless of the outcome we use, we find that both the coefficient for the "Originality rate" and the coefficient for the interaction between the originality rate and the high-competition indicator variable ("Originality rate * High competition") are positive and statistically significant at the one-percent level. In other words, given that we observe consumer switching across media outlets for all the media in our sample, originality always matters; but originality has a stronger positive effect for the outlets which are in a more competitive environment (i.e. with fewer captive users), and so are more subject to switching – in this case, a 50-percentage-point increase in the originality rate of an article leads to a 52.2% increase in the number of Facebook shares – than for the outlets that are in a less competitive environment (29.7% increase).

[Table 6 about here.]

Note however that these results should be interpreted with caution given the limits of the survey data we use to distinguish between "high-competition" and "low-competition" outlets. In particular, the survey data do not cover all the media outlets in France. Hence, we perform this estimation on nothing but a subset of the media outlets in our sample. We nonetheless believe they are of interest and help to highlight the mechanisms at play.

[77] The "low-competition" media outlets are TF1, BFM TV, France Television, *20 minutes*, Mediapart, *Le Monde*, Europe1, RMC, RTL, LCI, *Le Figaro*, France24, France Culture, France Info, France Inter, Metro, and Rue89. The "high-competition" media outlets are *Les Echos*, *Direct Matin*, *Courrier International*, I-TELE, *Liberation*, Slate, *La Croix*, *Marianne*, *L'Express*, *Le Point*, *Le Nouvel Obs*, and Atlantico.

Extent to which the media are copied The second dimension of heterogeneity we consider is the extent to which the media outlets are copied by their competitors. To do so, we rely on the results of Section 3.3 where we have computed, for each of the media outlets in our sample, the average share of their content that has been copied in 2013. Using the median, we split our sample into two groups and estimate equation (1) with an interaction term between the different explanatory variables and a "highly copied" indicator variable equal to one for the media outlets whose share of the content that has been copied is above the median (25.5%), and to zero otherwise.[78]

Table 7 presents the results. Whether we consider the number of Facebook shares, the number of Tweets or the predicted number of views at the article level, we find that, while the originality rate always has a positive and statistically significant effect on the audience received by the articles, this effect is lower for highly-copied media outlets. In other words, these results seem to indicate that the returns to originality are lower for the media outlets that suffer more from copying (a 50-percentage-point increase in the copy rate leads to a 40.5% increase in the number of views) than for the media outlets that suffer less from copying (49% increase in the number of views). This finding is also consistent with our simple theoretical framework.

[Table 7 about here.]

Furthermore, in the online Appendix Table F.12, we estimate equation (1) adding to the vector $\mathbf{Z}'_{\mathbf{aedn}}$ an additional characteristic of the article a published by media n on date d and included in the event e, namely the share of its content that has been copied by other media outlets. Evidently, this characteristic is hard to interpret given that not only the articles published first in an event – and the most original ones – tend to be the most copied, but also because the most copied articles may be the ones that are of higher "quality". The only proxy we have here for the "quality" of an article is its originality (originality rate or original content), and we probably miss out on other important dimensions. However, it is interesting to note that the share of the article that is copied is negatively correlated with our different measures of the article's audience once we control for the other characteristics of the article (the effect is not statistically significant for the number of predicted readers, though). E.g. a 50-percentage-point increase in the share of the article that is copied leads to a 3.9% decrease in the number of times it is shared on Facebook.

[78] The "highly copied" media outlets are (ranked by alphabetical order): *L'Alsace*, Arrêt sur images, Arte, BFM TV, *Le Bien Public*, *Capital*, *Centre Presse Aveyron*, *Challenges*, *La Charente Libre*, *Corse Matin*, *Le Courrier de L'Ouest*, *Le Dauphiné Libéré*, *La Dépêche du Midi*, *Les Dernières Nouvelles d'Alsace*, *Les Echos*, France Info, France Inter, France Télévision, France24, The Huffington Post, *Le JDD*, *Le Journal de Saone et Loire*, *L'Est Républicain*, LCI, *Le Midi Libre*, *Le Monde*, *La Montagne*, *Le Nouvel Obs*, *Ouest France*, *Le Parisien*, *Le Point*, *Presse Océan*, *Le Progrès*, RMC, RTL, *Le Républicain Lorrain*, *La République des Pyrénées*, *La République du Centre*, Rue89, TF1, *Le Télégramme*, and *Vosges Matin*.

Topic of the event Finally, do the characteristics of the events, and in particular their topic, affect the originality premium? In Table 8, we estimate equation (1) separately for the different events depending on their topic (politics, economy, sport, etc.). We find that the returns to originality – as measured by the effect of an increase in the originality rate on the number of times an article is shared on Facebook – are higher for "Crime, law and justice" events (a 50-percentage-point increase in the originality rate of an article leads to a 49.2% increase in the number of Facebook shares) as well as for "Politics" events (45.5% increase), and lower for events about "Economy, business and finance" (31.6%) and for "Sport" events (27.1%). Moreover, the difference in the magnitude of the effects is statistically significant. In other words, topics that generate less attention on social media, such as sport and economy[79], also seem to have lower returns to originality. In light of our very simple theoretical framework, this heterogeneity in the returns to originality can be interpreted in terms of the "easiness to find scoops" for a given investment in newsgathering. One can indeed assume that finding a "Sport event" scoop (e.g. reporting the results of a soccer game) may be "easier" than finding a "Politics event" scoop (e.g. reporting a political scandal). Assuming that readers are more willing to switch to the website of the original news producer when they acknowledge the "rareness" of the scoop found, we show that the returns to originality in terms of audience are lower for relatively more easy-to-find events. This prediction is consistent with the findings of Table 8.

We obtain similar results if we investigate heterogeneity in the returns to originality depending on the "general interest" of the events, as proxied by the total number of shares received by all the articles in an event. We generate a "High general interest" indicator variable equal to 1 for the events whose total number of shares received is higher than the median (201), and to 0 otherwise. We find that while originality always matters, the returns to originality are higher for the events with greater general interest, and that the difference is statistically significant (online Appendix Table F.13). Furthermore, this effect holds even within topics, i.e. if we perform the estimation separately for the different events depending on their topic. We present these results in the online Appendix Table F.14. While for "low general interest" events about "Crime, law and justice", a 50-percentage-point increase in the originality rate of an article leads to a 23.4% increase in the number of Facebook shares, it leads to a 54.5% increase for "high general interest" events about the same topic.

[Table 8 about here.]

[79] We show in Section 2.7 that "Sport" and "Economy, business and finance" events tend to generate fewer shares on Facebook than "Crime, law and justice" events (see also online Appendix Figure G.8).

5.2.5 Discussion

Ultimately, how can one rationalize the positive relationship between originality and news consumption as proxied by the number of shares on social media? As we have seen in Section 4, our preferred explanation is that consumers favor originality, and that the quality of the copy is lower than that of the original. The evidence we present in this section is consistent with the predictions of our simple theoretical framework on copying and returns to originality. It is also consistent with the fact that, as highlighted by Boczkowski and Mitchelstein (2013), consumption choices are *"often made at the story level"* (p.9). Hence, consumers willing to learn about a news event may decide to read the most original piece because they value its originality, or simply because this is the first article published within an event and so the first they have a chance to see. This may be partly due to the way search engines work. E.g., while the exact algorithm behind Google News is not public, it is well known that Google uses "freshness" and original content as a ranking signal.[80]

Note moreover that, while until now we have only considered consumers' switching behavior at the short-run level (using article-level estimations and media-outlet-event date variations), it is also possible that longer-run reputation effects allow original producers to recoup an even larger share of the audience. In the online Appendix Table F.15, we estimate the correlation between the average daily number of unique visitors (we compute this average over the year 2013) and the average content produced. We find that audience is positively correlated (with a statistically significant relationship) with the quantity of content classified in events, with the originality of the content produced, and with the number of breaking news. There is no statistically significant correlation between the quantity of content not classified in events and the number of unique visitors, however. In the online Appendix Table F.16, we perform a similar estimation but using the daily-level variations in audience and controlling for media and date fixed effects (the unit of observation is a media outlet-date and standard errors are clustered at the media outlet level). We find that the only characteristic of the content produced on a daily basis by a media outlet that has a statistically significant impact on the daily variations in its audience is the originality rate. The magnitude of the effect is small, however: a 50-percentage-point increase in the originality rate of the content published by a media outlet on a given date is associated with a 2.5% increase in its number of daily visitors. These results should be interpreted carefully though, given that these daily-level variations in the production of information and in the audience share of each media outlet allow us to estimate only correlations, not to identify causal effects.

[80] See e.g. "Google News: the secret sauce", published by Frederic Filloux in *The Guardian* on Monday 25 February 2013, and "An inside look at Google's news-ranking algorithm", by Jaikumar Vijayan, *Computerworld*, February 21, 2013. In 2011, Google published online a newsletter to webmasters, "More guidance on building high-quality sites". According to this newsletter, webmasters should ask themselves *"does the article provide original content or information, original reporting, original research, or original analysis?"*.

5.3 An audience-weighted measure of the importance of original content

Finally, we compute the audience-weighted share of original content in the dataset defined as:

$$\frac{\sum_a \text{original content}_a * \text{number of views}_a}{\sum_a \text{original content}_a * \text{number of views}_a + \sum_a \text{non-original content}_a * \text{number of views}_a}$$

where a index the articles. We do so by using our different measures of the number of views.

Figure 15 presents the results. First, for the sake of comparison, we compute the share of original content in the dataset. This share is equal to 32.5%.[81] Regardless of the methodology we use to compute article-level number of views, we find that the audience-weighted share of original content is higher than the actual share of original content in the dataset.

The audience-weighted share of original content varies from 45.4% when we use the naive approach (attributing to all the articles published by a media outlet on a given date the same number of views) to 61.4% when we allocate the number of views as a function of the number of shares on Facebook. It is important to highlight that the magnitude of our effect only slightly varies depending on the different methodologies: e.g. the audience-weighted share of original content is equal to 55.9% when we attribute the number of views assuming a linear relationship with the number of Tweets, and to 55.7% when we rely on the parameters estimated from *Le Monde*'s data. In other words, the relative consumption of original content online is always higher than its relative production, and the magnitude of the effect is fairly similar for our different specifications.

[Figure 15 about here.]

5.4 The returns to originality

The key question this paper attempts to address is the following: what fraction of the returns to original news content production is appropriated by the original news producers? Although our data sources do not allow us to fully address this question, our results can be used to provide some orders of magnitude.

Our basic result is that only 33% of the online content is original. Every time an original piece of content is published on the Internet, it is actually published three times: once by the original producer, and twice by media outlets who simply copy-and-paste this original content. In the event that Internet audience was distributed randomly on the different websites and

[81] We only consider here the articles for which we have audience data, and in particular we drop the AFP and Reuters. If we were to consider all the articles, then the share of original content in the dataset is equal to 32.6%. The difference with the average originality rate of the articles – 36.3% – comes from the fact that articles are of different length.

on the original and copied version of the articles, this result would imply that the original producer captures only 33% of the audience and of the economic returns to original news production (which as a first approximation can be assumed to be proportional to audience, even if as we saw above, the objective function of the media certainly includes other dimensions than audience size alone), and that the copiers capture up to 67% of the returns.

However, as we have just shown, audience is not randomly distributed on the Internet. First, if we weight content by media-level daily audience shares (using the naive approach), we find that original content represents 45.4% of online news consumption. This may reflect the fact that media outlets with a larger fraction of original content tend to attract a higher audience, possibly because they have a stronger reputation and/or because on days when more original content is published there is also a higher audience. To further investigate this issue, we weight the content of each article by the average annual audience of the media outlet in which it was published (assuming all the articles published on the website of an outlet in 2013 received the same number of views). When we do so, we find that the original content represents 45.6% of the online news consumption, i.e. almost the same share as when we use the naive approach (weighting the content by media-level daily audience shares). This shows that the daily-level audience almost hardly varies on average with daily-level average originality. This result – which is also consistent with the very small magnitude of the impact of original copy on audience we find when we only consider daily-level variations in audience and control for media and date fixed effects (see Section 5.2.5 above and the associated Table F.16) – suggests that media-level reputation effects play an important role.

Most importantly, if we weight content by media-level audience shares and article-level Facebook shares or article-level number of Tweets, we find that the original content represents between 53.1% and 61.4% of online news consumption, depending on the approach chosen. I.e. within a given media outlet, the articles that get more views (as approximated by the number of shares on social media) are those with more original content. In effect, thanks to the combination of media-level reputation effects and consumers' preference for originality at the article level, the audience share of original content jumps from 32.5% to between 53.1 and 61.4%.[82]

As a robustness check on this estimation of the returns to originality, we perform the same analysis but after having dropped all the content copied from the news agencies (given that the business model of news agencies is a different one). More precisely, we define the *total content* of an article as its content minus the content reproduced from the news agencies, and

[82] We obtain a similar result in terms of magnitude if we compute the originality rate excluding internal copy, i.e. a media outlet copying content from an article it has itself previously published in the event (online Appendix Figure G.22). Considering internal copy as original content, the share of original content in our dataset is 36.3% (when as before we focus on the media outlets for which we have audience data). If we weight this content by media-level audience shares and article-level Facebook shares or article-level number of Tweets, this share jumps to 50 to 65% depending on the specification.

the *original content* of an article as its content minus the content reproduced from the news agencies and the content reproduced from other media outlets (excluding itself). Doing so, we find that on average documents are 1,311 characters long. We also observe that 69.3% of the online content is original; higher originality is not surprising given that we have shown that media outlets mainly rely on content copied from the news agencies. What about the relative consumption of original content? We find that the audience-weighted share of original content is equal to 79.7% when we use the naive approach, and to between 81.9 and 84.4% when we allocate the number of views as a function of the number of shares on social media (online Appendix Figure G.23). Hence, despite a lower reliance on copying, media-level reputation and consumers' preference for originality still lead to a consumption of original content that is higher than its relative production.

We should stress that our computations might underestimate the extent of copying. This might arise first because our plagiarism detection algorithm is not perfect – it captures only exact verbatim copying but not rewording – and also because the copied segments of a given article might be the most "valuable" and original segments (something we cannot fully measure). Moreover, we might also underestimate the magnitude of the reputation effects. I.e. Internet viewers might well find ways to detect original articles (and discard copying-and-pasting) other than social media shares, e.g. via their own appraisal, friends, privately accessible social networks or other devices. Our estimates of the extent to which producers are able to capture the returns to original news production should be viewed as provisional and imperfect, and should be improved in the future. But at least they show that reputation mechanisms and the demand side of the market for online news need to be taken into account when studying the impact of copying on the incentives for news production.

6 Robustness checks and discussion

In this Section, we perform a number of robustness checks. First, we relax the "10 documents condition" previously used to define a media event. We investigate how this affects the evidence regarding the propagation of online news, as well as our findings regarding original content and news consumption. Second, we show that our results are robust to detecting the news events using alternative embedding techniques. Finally, we discuss the external validity of our main findings.

6.1 Relaxing the "10 documents condition"

Not surprisingly, the total number of media events identified by the algorithm strongly increases when we relax the 10 documents condition. We obtain a total number of 113,959 news events. Out of the 2,552,442 articles in our dataset, 1,203,521, i.e. 47.2%, are classified in

the events thus defined. Classified documents represent 53.5% of the total content produced in 2013. If the documents classified in the events thus defined are indeed on average smaller than when the 10 documents condition is imposed – 2,334 characters long compared to 2,467 – they are nonetheless longer than the documents that are not classified (1,810 characters on average) (online Appendix Table H.1). Unclassified documents mostly come from local daily newspapers which account for nearly 75.9% of the unclassified documents (bearing in mind that those newspapers represent 55.7% of the documents in our dataset) (online Appendix Figure H.1).

When we relax the 10 documents condition, the events are much shorter (they last less than 19 hours on average, compared to 41 when the condition is imposed) and comprise on average a lower number of documents (online Appendix Table H.2). As before, they are mainly about "politics" and "economy, business and finance" (online Appendix Figure H.2). If the "crime, law and justice" category is still in third place, it is associated with less than 14% of the events, and "sport" does nearly as well as "crime".[83]

Regarding the 1,203,521 documents classified in the events, their originality rate is equal on average to 42.6% (online Appendix Table H.3). This rate is nearly 6 percentage points higher than when we impose the condition that events should contain at least 10 documents. Figure 16 plots the distribution of the originality rate: as before, it appears clearly that this distribution is bimodal, with one peak for the articles with less than 1% of original content (around 14% of the documents) and another peak for the 100%-original articles. The latter, with nearly 30% of the documents, is higher than when we impose the condition. Note however that even when this condition is relaxed, nearly 50% of the articles classified have less than 20% originality. In other words, our finding regarding the importance of copying online is robust to this alternative definition of media events.

[Figure 16 about here.]

If we now turn to the ratio of original content in the dataset over the total content, it is equal to nearly 39%. Are our findings on consumers' taste for original content robust to this alternative definition of events? We follow exactly the same empirical strategy as before, using the different approaches defined above to compute article-level number of views and estimate the audience-weighted share of original content. Figure 17 shows the results. As before, it appears clearly that the audience-weighted share of original content (which varies between 57.1 and 66% depending on the specification) is much higher than the production of original content.

[83] Note however that, as we highlighted it above, we rely on the metadata associated with the AFP dispatches included in the events to define the topic of each event. Yet, when we relax the minimum number of 10 documents per event condition, the share of events including at least one AFP dispatch decreases from 95% to 63.4%. Hence, the statistics presented in the online Appendix Figure H.2 are computed for only 63% of the events. For the remaining 37%, we have no information on their topic.

[Figure 17 about here.]

Finally, we re-estimate equation (1) using the dataset where we relax the "10 documents condition" to define a media event. Table 9 presents the main results, with the effects on the number of Facebook shares reported in Columns (1) to (3), on the number of Tweets in Columns (4) to (6), and on the predicted number of views in Columns (7) to (9). As before, the dependent variable is in log and all the estimations include media outlets, date, and event fixed effects. While both the number of observations and the number of events differ compared to the estimations presented in Table 5, the coefficients we obtain for each of the explanatory variables of interest are of the same order of magnitude. For example, we find that an increase of $1,000$ in the number of original characters leads to a 21.2% increase in the number of shares on Facebook, a 10.6% increase in the number of shares on Twitter, and a 22.4% increase in the predicted number of views.

[Table 9 about here.]

Hence, the main findings of this paper do not depend on the threshold we impose regarding the number of articles to define an event. While changing this threshold by construction affects the number of events and the share of articles in the dataset that are classified in events, our main results on the one hand regarding the importance of copying online, and on the other hand regarding consumers' taste for originality and the role played by reputation mechanisms, are robust to this alternative approach.

As an additional robustness check, we exclude all the very large events from our dataset.[84] Such events can indeed be considered as "garbage" clusters, a concern raised by Allan et al. (2005). Online Appendix Table F.17 provides the results of the estimation of equation (1) when we do so. Reassuringly, while reducing by construction the number of observations, removing the big clusters does not impact the magnitude nor the statistical significance of our estimates.

6.2 Alternative event detection algorithms

The event detection algorithm – that we have developed to identify the media events – is a key element of this article. The algorithm is composed of two main parts: the clustering algorithm and the semantic features for the text representation. In the past few years, the Natural Language Processing (NPL) research field has made great progress in several tasks by using new text representation schemes that better model the language and thus the semantic. These language models – among which Word2Vec (Mikolov et al., 2013), Doc2Vec (Le and

[84] Specifically, we drop all the events with more than 269 documents, i.e. the 99th percentile of the distribution. See online Appendix Table F.3.

Mikolov, 2014), Glove (Pennington et al., 2014), etc. – are all learned with neural networks on very large corpus of texts. These new text representations have been used to replace the standard TF-IDF scheme (that we use in the event detection algorithm described in Section 2.2) in several NLP tasks (sentiment analysis, question answering, textual similarity, etc.), and brought significant improvements in terms of the performances of the algorithm. However, these new representations do not improve the performance of the Topic Detection and Tracking task (for media event detection using news articles).

In Mazoyer et al. (2019), we indeed explore the potential benefits of the new word embedding models for the Topic Detection and Tracking task.[85] In particular, we test the accuracy of the approach we use in this article against Word2Vec and Doc2Vec-based methods, using the dataset of media events we have created manually from our 2013 French corpus. First, we find that Doc2vec has much lower performances than the TF-IDF scheme; we thus decide to discard this approach. Second, we show that the best Word2Vec document representation is obtained by using the TF-IDF weighting of word vectors instead of a simple mean.[86] Third, we show that even the TF-IDF-weighted Word2Vec method does not perform better than the simple TF-IDF representation. Consequently, in our core specification, we use the TF-IDF scheme.

In this section, as an additional robustness check, we nonetheless investigate whether our main findings are robust to using the TF-IDF-weighted Word2Vec approach. We briefly describe the results here; the detailed results are available in the online Appendix Section I. When we use this alternative news text representation, we obtain a total number of $21,783$ media events. $793,106$ articles, i.e. 31% of the documents, are classified in the events.[87] As before, documents classified in events are on average longer than unclassified documents (online Appendix Table I.1), and unclassified documents mainly come from local daily newspapers (online Appendix Figure I.1).

[85] To the extent of our knowledge, this has never been done before. The purpose of these word embedding schemes is to better model the latent semantic meaning of words by taking into account the context in which they are used in a large corpus of texts. Succinctly, the models are based on the co-occurrences of words in a given window (typically 15 or 20 words), and are learned with neural networks. The result is a dense vector of fixed dimension (typically 300) representing each word. Different strategies are then used to combine these word embeddings in order to obtain each document representation. A max or mean strategy is generally used with the Word2Vec and the Glove representations. Doc2Vec is an extension of the Word2vec framework that learns the document representations jointly with the word embeddings. The cosine similarity measure (which is the semantic similarity measure we use in this article) is then used to compute the distance between documents.

[86] This result – that is presented in detail in Mazoyer et al. (2019) – was predictable. Indeed, the TF-IDF representation has the advantage of emphasizing the rare and particularly distinctive words of a document. This will typically be the case of names of people or places. Conversely, the TF-IDF scheme is unable to handle synonyms since each word is considered in a unique way. With Word2Vec, each vector representing a word richly encodes its context of use. Thus, lexical fields and synonyms are taken into account. But a simple average of these representations does not allow to accentuate the weight of the words specific to a document in the corpus. It is therefore logical that the weighting by TF-IDF of a Word2Vec embedding produces better performances than the use of the mean.

[87] Classified documents represent 39.9% of the total content produced in 2013.

Regarding the length of the events, we find that the events thus defined tend to last longer on average (46.6 hours compared to 41 hours in our preferred specification). Similarly, the average number of documents included in the events is higher. While the median is unchanged, this difference comes from the fact that this alternative news text representation leads to a higher number of "garbage" clusters (online Appendix Table I.2; this is an additional argument in favor of using the TF-IDF rather than the Word2Vec embedding). There is no difference regarding the topic of the events compared to the classification when using the TF-IDF scheme (online Appendix Figure I.2).

The average originality rate of the articles classified in events is 35.6% (online Appendix Table I.3, compared to 36.5% in our preferred specification). Online Appendix Figure I.3 plots the distribution of the originality rate; compared to Figure 5a, it is nearly unchanged. The distribution is bimodal, with one peak for the articles with less than 1% of original content and another peak for the 100%-original articles, and 56.3% of the documents have less than 20% originality (compared to 55.2% in our preferred specification). The ratio of original content in the dataset over the total content is equal to 34.1%.

Finally, we examine whether the use of the Word2Vec representation affects our main findings regarding consumers' taste for original content. Reassuringly, it does not. Figure I.4 shows the audience-weighted share of original content. Regardless of the methodology we use to compute the article-level number of views, we find as before that the audience-weighted share of original content is higher than the actual share of original content in the dataset. Compared to the estimates we obtain when using the TF-IDF scheme, the share of original content is around two percentage-point higher with this news text representation. Lastly, we re-estimate equation (1) using this new set of media events. Table I.4 presents the main results, with the effects on the number of Facebook shares reported in Columns (1) to (3), the effect on the number of Tweets reported in Columns (4) to (6), and in Columns (7) to (9) the predicted number of views. Whether we consider the sign, the magnitude or the statistical significance of the estimates, we find results similar to the ones presented in Table 5. In other words, the main findings of this paper do not depend on the embedding method used to identify the media events.

6.3 External validity

The results presented in this paper are based on French data for the year 2013. Hence, one final question is whether we should expect the patterns we have uncovered in the case of 2013 France to be repeated in other contexts. First, should these patterns hold in other countries? And second, should they still hold nowadays? There are good reasons to think this could be the case.

First, while the French media market certainly presents specific features, but no more

than any other given market, it is by and large very similar to other Western media markets, whether we consider Internet penetration (87%, like Italy and Spain and only slightly below Belgium – 88% – and Germany – 90%), the use of social media for news (36%, compared to 31% for Germany and 39% for the U.K.), or the proportion of the population who paid for online news (11%, like in Spain, slightly above Germany or Canada – 8% – but below Italy – 12%) (Reuters Institute, 2018). In France, like in other Western European media markets, many publishers offer online news for free (see below), and largely rely on advertising. Moreover France, like the U.S., has an international news agency, the AFP, which is the third leading agency in the world after Reuters and Associated Press. From this point of view, the French market is more similar to the U.S. market than the Spanish, Italian, or German markets. Therefore, overall, we believe the patterns we uncover regarding the propagation of online information, the importance of copying and the valuation of originality using French data would hold in other contexts.

Obviously, we are also well aware of the fact that the digital news market has evolved since 2013. In particular, pay models are becoming an important part of the business of digital news, while they were just in their infancy in 2013. On the one hand, however, in many markets there are still many publishers who offer online news for free.[88] And on the other hand, digital advertising revenues are still the main source of revenue for the media online. In 2018, only 16% of the consumers paid for online news content in the United States[89], 12% in Italy, 11% in France, and 7% in the United Kingdom (Reuters Institute, 2018). Hence, even if new paywall systems have developed since 2013 and may further develop in the future, it is important to highlight that digital advertising remains a critical source of revenue. Furthermore, the growing importance of the paywall systems, while modifying media outlets' sources of revenues, should not significantly modify the impact of copying on newsgathering incentives.[90]

Lastly, it should be noted that even though the news market has changed online in recent years, the French media market has been far from upset since 2013. First, according to Reuters, there was no change between 2013 and 2018 in the use of the Internet as a source of news (68%) (Reuters Institute, 2013, 2018). Second, the French media landscape has remained pretty stable, in particular if we consider the main media outlets in terms of audience. E.g. the main media outlets in terms of audience in 2013 are still the main ones in 2018 (with *20*

[88] *"Despite the shift towards reader payment models, it is worth remembering that the majority of online news consumption still happens through free websites, largely supported by advertising"* (Reuters Institute, 2018).

[89] This relatively high share in the United States in 2018 comes from the so-called "Trump Bump".

[90] The main effect could have been through a decrease in readers' mobility across media outlets. But we actually observe an increase in consumers' switching. Using similar Reuters' data to the one we use in Section 4.2 but for the year 2018, we indeed find that the average number of media outlets consumed by consumers who consume at least one news media increased from 2.35 in 2013 to 2.83 in 2018. Furthermore, the introduction of paywalls may also raise the audience-driven incentives to invest in newsgathering, since it implies that the audience would be positively correlated both with the advertising and the subscription revenues.

Minutes, *Le Monde*, and *Le Figaro* attracting the largest share of the audience).

Overall, we thus believe that the results presented in this paper have implications for other Western countries and still hold nowadays.

7 Conclusion

This paper documents the extent of copying online and estimates the returns to originality in online news production. It builds a unique dataset combining the online production of information of the French news media during the year 2013 with micro audience data, and develops a number of algorithms which could be of future use to other researchers studying media content. We investigate the speed of news dissemination and distinguish between original information production and copy-and-paste. We find that less than 33% of the online news content is original.

This scale of copying online might help rationalize the observed drop in media companies' employment of journalists in recent years, raising growing concern about the industry's ability to produce high-quality information (see e.g. Angelucci and Cagé, 2019). In the event that online audience was distributed randomly and revenues were proportional to audience, our results would imply that the original news producers capture only one third of the economic returns to the original news content they provide. We investigate whether the scale of copying online negatively affects media's newsgathering incentives, and discuss a number of mechanisms.

We show that long-term reputation mechanisms and the short-run behavior of Internet viewers – in particular their preference for original content at the article level – make it possible to mitigate a significant part of the plagiarism problem. We indeed find that original content represents up to 61% of online news consumption, i.e. much more than its share of online news production.

Of course, greater intellectual property protection could also play a role in reducing copyright violation and raising the incentives for original news production, and we certainly do not mean to downplay the extent of this problem. In 2010, the Federal Trade Commission (FTC) in the United States issued a discussion paper outlining the enactment of "Federal Hot News Legislation" as a proposal aimed at reinventing journalism and addressing newspapers' revenue problems. But now that digital information is very easy to copy and distribute, copyright laws may become almost impossible to enforce. Furthermore, our results suggest that in order to effectively address these issues, it is important to study preference for originality, reputation effects and how viewers react to the newsgathering investment strategies of media outlets.

Finally, we think that our results – as well as the algorithms we developed for this study

– may help to improve our future understanding of "where people get their news", combining consumption and production data. Prat (2018) and Kennedy and Prat (2019) have documented news consumption across platforms; a complementary strategy to estimate media power would be to weight the influence of media companies by their supply of original news and how much other companies rely on that news. More research is still needed, but we hope this paper will inform the debate on concentration in media power.

References

Algan, Yann, Yochai Benkler, Jérôme Hergueux, and Mayo Fuster-Morell, "Cooperation in a Peer Production Economy: Experimental Evidence from Wikipedia," Working Paper 2016.

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.

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

Anderson, Simon P., "Advertising on the Internet," in Martin Peitz and Joel Waldfogel, eds., *The Oxford Handbook of the Digital Economy*, Oxford University Press, 2012.

Anderson, Simon P and John McLaren, "Media Mergers And Media Bias With Rational Consumers," *Journal of the European Economic Association*, 2012, *10* (4), 831–859.

Angelucci, Charles and Julia Cagé, "Newspapers in Times of Low Advertising Revenues," *American Economic Journal: Microeconomics*, 2019, *Forthcomin.*

Athey, Susan and Markus Mobius, "The Impact of News Aggregators on Internet News Consumption: The Case of Localization," Technical Report 2012.

_ , Emilio Calvano, and Joshua Gans, "The Impact of the Internet on Advertising Markets for News Media," Working Paper 19419, National Bureau of Economic Research 2013.

Bae, Sang Hoo and Jay Pil Choi, "A model of piracy," *Information Economics and Policy*, 2006, *18* (3), 303–320.

Balan, David J., Patrick DeGrava, and Abraham L. Wickelgren, "Ideological Persuasion in the Media," Working Paper 2014.

Besley, Timothy and Andrea Prat, "Handcuffs for the Grabbing Hand? Media Capture and Government Accountability," *American Economic Review*, 2006, *96* (3), 720–736.

Biasi, Barbara and Petra Moser, "Effects of Copyrights on Science: Evidence from the WWII Book Replication Program," Working Paper 2015.

Boczkowski, Pablo J., *News at Work: Imitation in an Age of Information Abundance*, University of Chicago Press, 2010.

_ and Eugenia Mitchelstein, *The News Gap: When the Information Preferences of the Media and the Public Diverge* The News Gap, MIT Press, 2013.

Cagé, Julia, "Media Competition, Information Provision and Political Participation: Evidence from French Local Newspapers and Elections, 1944-2014," CEPR Discussion Papers 12198, C.E.P.R. Discussion Papers 2017.

_ and Lucie Gadenne, "Tax Revenues and the Fiscal Cost of Trade Liberalization, 1792-2006," PSE Working Papers halshs-00705354, HAL 2015.

Calzada, Joan and Ricard Gil, "What Do News Aggregators Do? Evidence from Google News in Spain and Germany," Working Paper 2016.

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

Cornia, Alessio, Annika Sehl, Felix Simon, and Rasmus Kleis Nielsen, "Pay Models in European News," Factsheet 2017.

Deloitte, "The impact of web traffic on revenues of traditional newspaper publishers. A study for France, Germany, Spain, and the UK," Technical Report 2016.

Digital Strategy Group of the European Broadcasting Union, "Media with a purpose. Public Service Broadcasting in the digital age," Technical Report 2002.

Duggan, John and Cesar Martinelli, "A Spatial Theory of Media Slant and Voter Choice," *The Review of Economic Studies*, 2011, *78* (2), 640–666.

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

Fink, Katherine and Michael Schudson, "The rise of contextual journalism, 1950s-2000s," *Journalism*, 2014, *15* (1), 3–20.

Franceschelli, Ignacio, "When the Ink is Gone: The Transition from Print to Online Editions," Technical Report, Northwestern University 2011.

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*, 2018.

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

_ , "Valuing New Goods in a Model with Complementarity: Online Newspapers," *American Economic Review*, jun 2007, *97* (3), 713–744.

_ **and Jesse M Shapiro**, "Competition and Truth in the Market for News," *Journal of Economic Perspectives*, 2008, *22* (2), 133–154.

_ **and** _ , "What Drives Media Slant? Evidence from US Daily Newspapers," *Econometrica*, 2010, *78* (1).

_ **,** _ **, and Daniel F Stone**, "Chapter 14 - Media Bias in the Marketplace: Theory," in Simon P Anderson, Joel Waldfogel, and David Strömberg, eds., *Handbook of Media Economics*, Vol. 1 of *Handbook of Media Economics*, North-Holland, 2015, pp. 623–645.

_ **, Nathan Petek, Jesse M Shapiro, and Michael Sinkinson**, "Do Newspapers Serve The State? Incumbent Party Influence On The US Press, 18691928," *Journal of the European Economic Association*, 2015, *13* (1), 29–61.

George, Lisa M, "The Internet and the Market for Daily Newspapers," *The B.E. Journal of Economic Analysis & Policy*, 2008, *8* (1), 1–33.

_ **and Christiaan Hogendorn**, "Aggregators, search and the economics of new media institutions," *Information Economics and Policy*, 2012, *24* (1), 40–51.

_ **and** _ , "Local News Online: Aggregators, Geo-Targeting and the Market for Local News," Working Paper 2013.

_ **and Joel Waldfogel**, "The New York Times and the Market for Local Newspapers," *American Economic Review*, mar 2006, *96* (1), 435–447.

Giorcelli, Michela and Petra Moser, "Copyright and Creativity: Evidence from Italian Operas,"

Working Paper 2015.

Greenstein, Shane and Feng Zhu, "Collective Intelligence and Neutral Point of View: The Case of Wikipedia," Working Paper 18167, National Bureau of Economic Research 2012.

_ , Yuan Gu, and Feng Zhu, "Ideological Segregation among Online Collaborators: Evidence from Wikipedians," Working Paper 22744, National Bureau of Economic Research 2016.

Hamilton, James T., *All the News That's Fit to Shell: How the Market Transforms Information Into News*, Princeton University Press, 2004.

_ , *Democracy's Detectives: The Economics of Investigative Journalism*, Harvard University Press, 2016.

Haveman, Heather A., *Magazines and the Making of America: Modernization, Community, and Print Culture, 1741-1860* Princeton Studies in Cultural Sociology, Princeton University Press, 2015.

_ **and Daniel N. Kluttz**, "Property in Print: Copyright Law and the American Magazine Industry," Working Paper 2014.

Kennedy, Patrick J and Andrea Prat, "Where Do People Get Their News?," *Economic Policy*, 2019.

Le, Quoc and Tomas Mikolov, "Distributed Representations of Sentences and Documents," in "Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32" ICML'14 JMLR.org 2014, pp. 1188–1196.

Li, Xing, Megan MacGarvie, and Petra Moser, "Dead poets' property - How Does Copyright Influence Price?," *The RAND Journal of Economics*, 2018, *49* (1), 181–205.

MacGarvie, Megan and Petra Moser, "Copyright and the Profitability of Authorship: Evidence from Payments to Writers in the Romantic Period," in "Economic Analysis of the Digital Economy" NBER Chapters, National Bureau of Economic Research, Inc, 2014, pp. 357–379.

Mazoyer, Béatrice, Nicolas Hervé, Marc Evrard, Julia Cagé, and Céline Hudelot, "Word Embeddings for Topic Detection and Tracking," Technical Report, INA Working Paper 2019.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," in "Advances in neural information processing systems" 2013, pp. 3111–3119.

Nagaraj, Abhishek, "Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia," *Management Science*, 2018, *64* (7), 3091–3107.

Nicholls, Tom, Nabeelah Shabbir, and Rasmus Kleis Nielsen, "Digital-Born News Media in Europe," techreport, Reuters Institute for the Study of Journalism 2016.

_ , _ , Lucas Graves, and Rasmus Kleis Nielsen, "Coming of Age: Developments in Digital-Born News Media in Europe," techreport, Reuteurs Institute for the Study of Journalism 2018.

Noam, Eli, *Who Owns the World's Media?: Media Concentration and Ownership around the World*, New York: Oxford University Press, 2016.

OberholzerGee, Felix and Koleman Strumpf, "The Effect of File Sharing on Record Sales: An Empirical Analysis," *Journal of Political Economy*, 2007, *115* (1), pp. 1–42.

Open Society Foundations, "Mapping Digital Media: France," Technical Report 2013.

Peitz, Martin and Markus Reisinger, "Chapter 10 - The Economics of Internet Media," in Joel Waldfogel Simon P. Anderson and David Strömberg, eds., *Handbook of Media Economics*, Vol. 1 of *Handbook of Media Economics*, North-Holland, 2016, pp. 445–530.

_ **and Patrick Waelbroeck**, "Piracy of digital products: A critical review of the theoretical literature," *Information Economics and Policy*, 2006, *18* (4), 449–476.

_ **and** _ , "Why the music industry may gain from free downloading The role of sampling," *International Journal of Industrial Organization*, 2006, *24* (5), 907–913.

Pennington, Jeffrey, Richard Socher, and Christopher D Manning, "Glove: Global Vectors for Word Representation.," in "EMNLP," Vol. 14 2014, pp. 1532–1543.

Pew Research Center, "State of the News Media Report 2016," Report 2016.

Picone, Ike, Cédric Courtois, and Steve Paulussen, "When News is Everywhere," *Journalism Practice*, 2015, *9* (1), 35–49.

Prat, Andrea, "Media Power," *Journal of Political Economy*, 2018, *126* (4), 1747–1783.

Puglisi, Riccardo and James M Snyder, "Newspaper Coverage of Political Scandals," *The Journal of Politics*, 2011, *73* (3), 931–950.

Reimers, Imke, "Copyright and Generic Entry in Book Publishing," *American Economic Journal: Microeconomics*, 2019, *Forthcomin.*

Reuters Institute, "Digital News Report 2013," Annual Report 2013.

_ , "Digital News Report 2017," Annual Report 2017.

_ , "Digital News Report 2018," Annual Report 2018.

Rob, Rafael and Joel Waldfogel, "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students," *Journal of Law and Economics*, 2006, *49* (1), pp. 29–62.

Rosenstiel, Tom, Marion Just, Todd L. Belt, Atiba Pertilla, Walter Dean, and Dante Chinni, *We Interrupt This Newscast: How to Improve Local News and Win Ratings, Too*, Cambridge University Press, 2007.

Salami, Abdallah and Robert Seamans, "The Effect of the Internet on Newspaper Readability," Working Papers 14-13, NET Institute 2014.

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

Schudson, Michael, *Discovering the News: A Social History of American Newspapers*, Basic Books, 1981.

_ , *The Rise of the Right to Know: Politics and the Culture of Transparency, 1945-1975*, Harvard University Press, 2015.

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.

Stein, Benno, "Principles of Hash-based Text Retrieval," in Charles Clarke, Norbert Fuhr, Noriko Kando, Wessel Kraaij, and Arjen P de Vries, eds., *30th International ACM Conference on Research and Development in Information Retrieval (SIGIR 07)*, ACM 2007, pp. 527–534.

Swart, Joëlle, Chris Peters, and Marcel Broersma, "Navigating cross-media news use," *Journalism Studies*, 2017, *18* (11), 1343–1362.

Tuchman, Gaye, *Making News*, Free Press, 1980.

Waldfogel, Joel, "Copyright Protection, Technological Change, and the Quality of New Products:

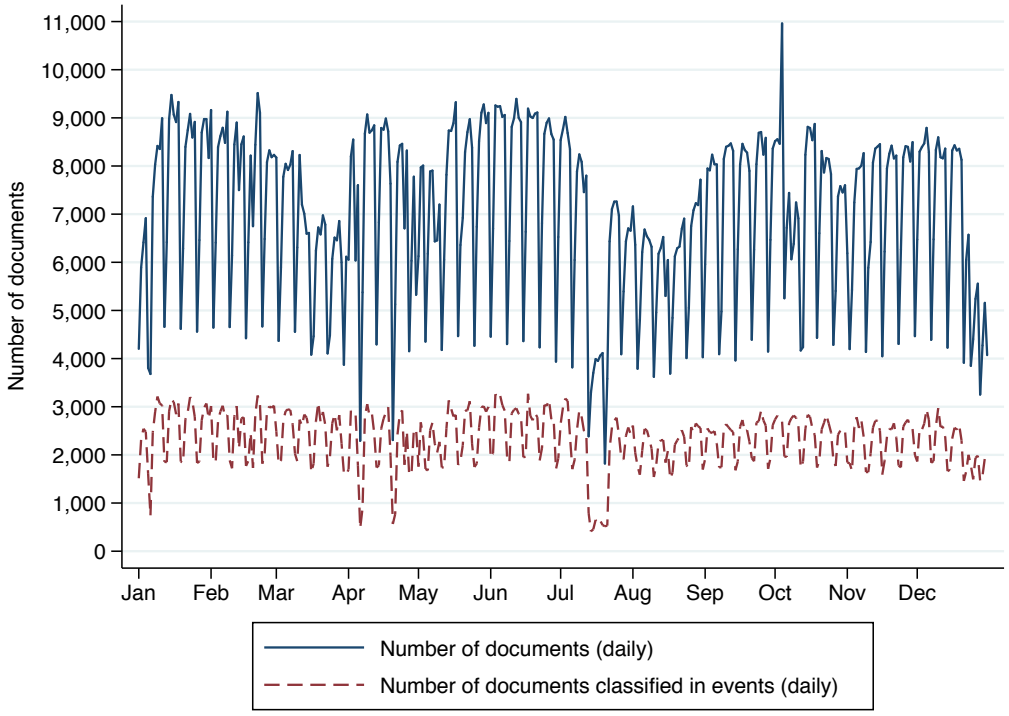
Evidence from Recorded Music since Napster," *The Journal of Law & Economics*, 2012, *55* (4), 715–740.

_ , *Digitization and the Quality of New Media Products: The Case of Music*, University of Chicago Press,

_ , "How Digitization Has Created a Golden Age of Music, Movies, Books, and Television," *Journal of Economic Perspectives*, 2017, *31* (3), 195–214.

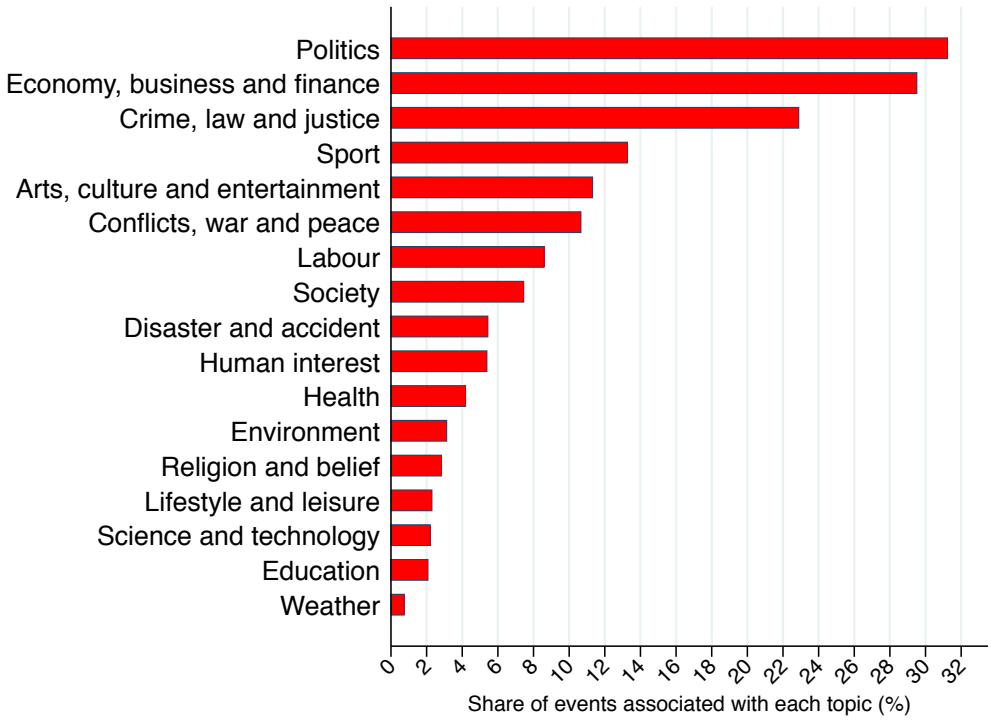
Yuan, Elaine, "News Consumption Across Multiple Media Platforms," *Information, Communication & Society*, 2011, *14* (7), 998–1016.

Zhu, Yi and Kenneth C. Wilbur, "Hybrid Advertising Auctions," *Marketing Science*, 2011, *30* (2), 249–273.



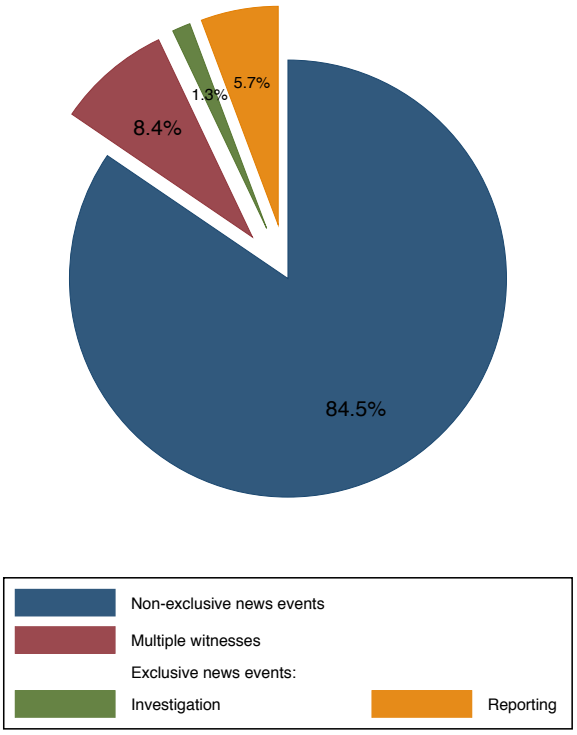
Notes: The figure plots the total daily number of documents included in our dataset. The solid blue line shows the total number of documents. The red dashed line shows the number of documents that are classified in news events. News events are defined in details in the text.

Figure 1: Daily distribution of the number of documents and of the number of documents classified in events in the dataset



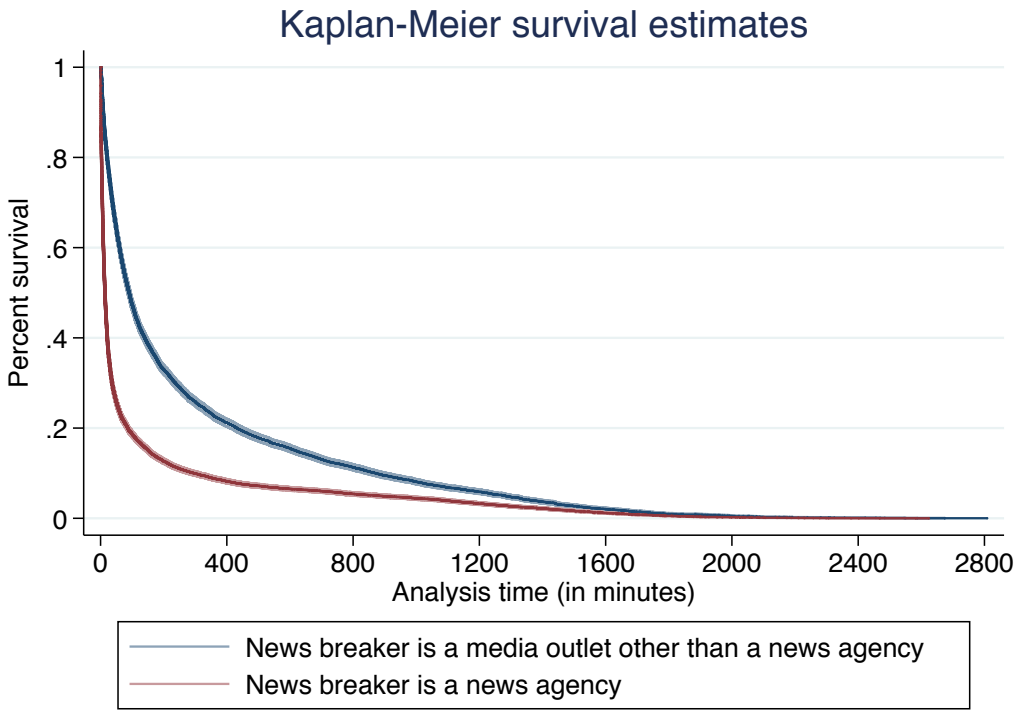
Notes: The figure shows the share of events associated with each media topic. The topics correspond to the IPTC media topics described in the text and defined in the online Appendix. Because some events are associated with more than one topic, the sum of the shares is higher than 100%.

Figure 2: Share of events associated with each media topic



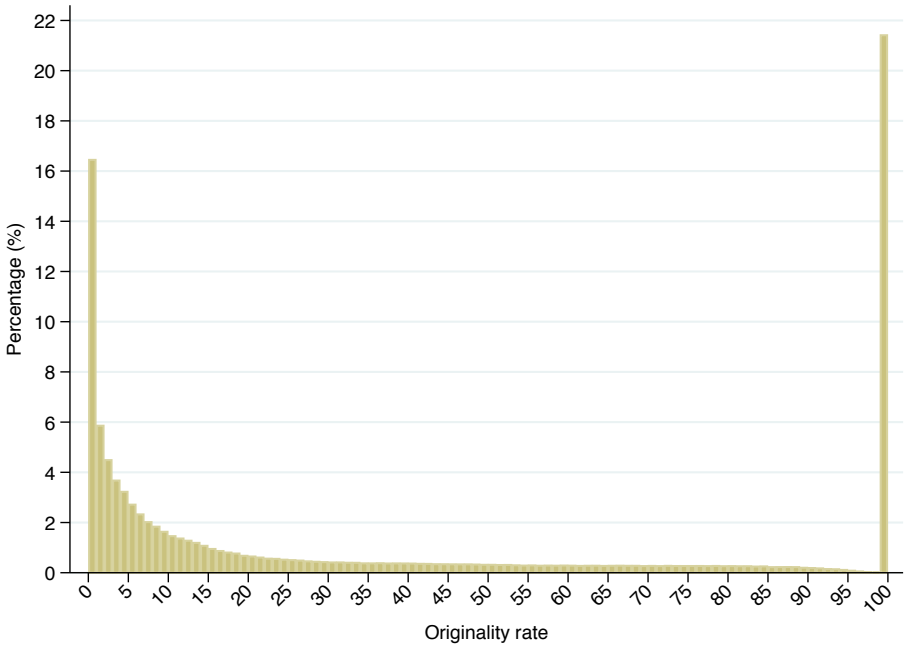
Notes: The figure plots the share of the news events classified in each category depending on their nature: exclusive news events, non-exclusive news events, and short news items with multiple witnesses. Exclusive news events can be either investigative stories or (non-investigative) reporting stories. The classification of the news events and the definition of the categories are described in details in the text.

Figure 3: Share of the events depending on the information issuer

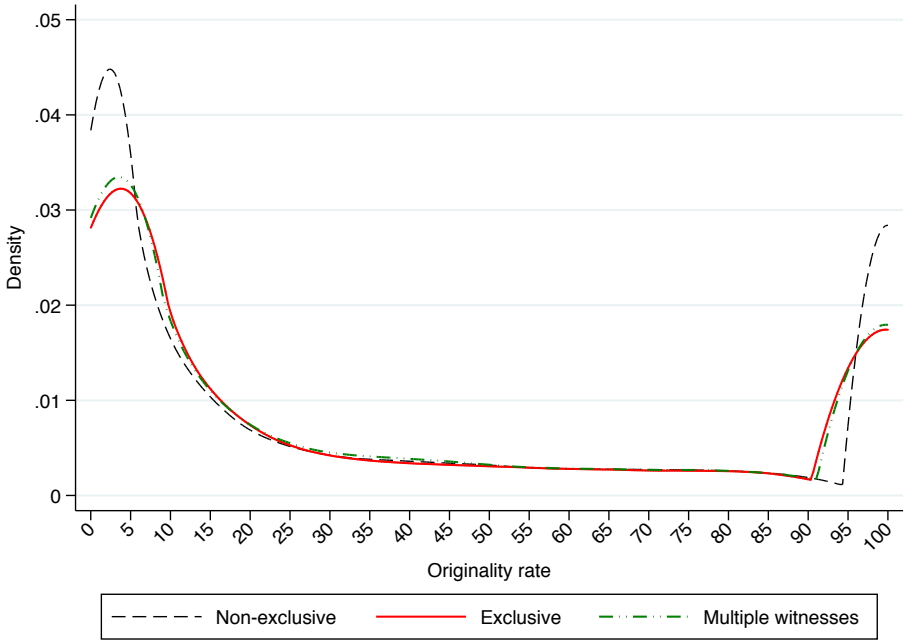


Notes: The figure plots the Kaplan-Meier survivor functions when the news breaker is a news agency (the AFP or Reuters) (red line) and when the breaker is a media outlet other than a news agency (blue line). The confidence level for the pointwise confidence bands is 95%.

Figure 4: Average reaction time depending on whether the news breaker is a news agency: Kaplan-Meier survival estimates



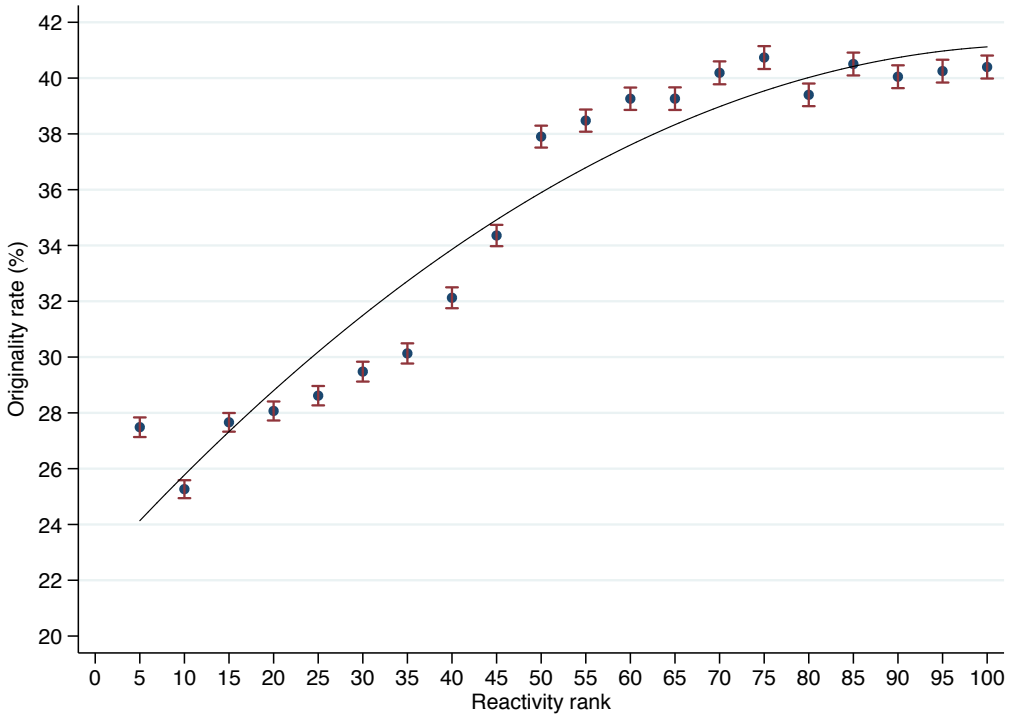
(a) Distribution (all documents classified in events)



(b) Kernel density estimate

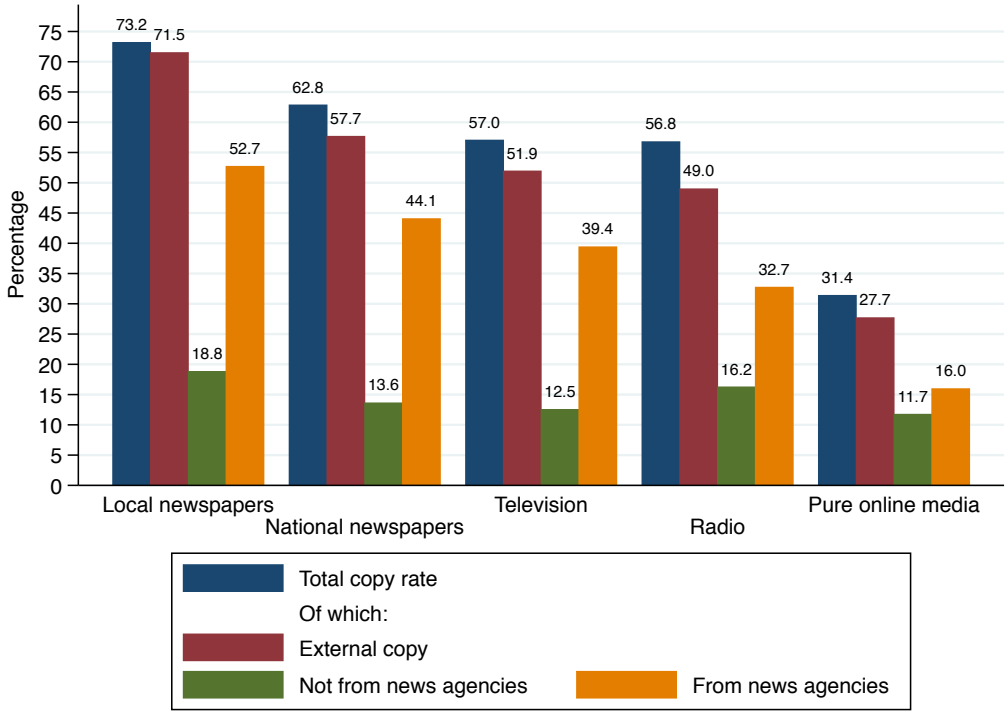
Notes: The upper figure plots the distribution of the originality rate (with bins equal to one percent). The bottom figure plots the Kernel density estimates depending on the nature of the news event. News events are either exclusive news events, non-exclusive news events, or short news items with multiple witnesses. The classification of the news events and the definition of the categories are described in details in the text.

Figure 5: Originality rate



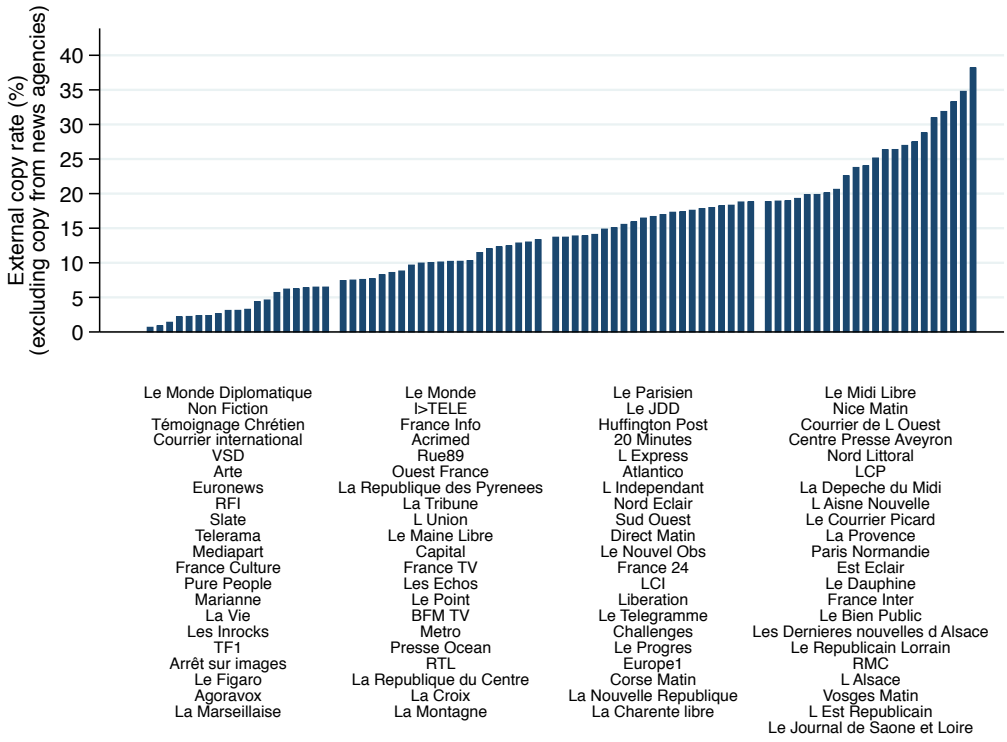
Notes: The figure plots the average originality rate of the articles depending on the reactivity rank (error bars in red represent the 95% confidence interval).

Figure 6: Correlation between originality and reaction time: average originality rate depending on the reactivity rank



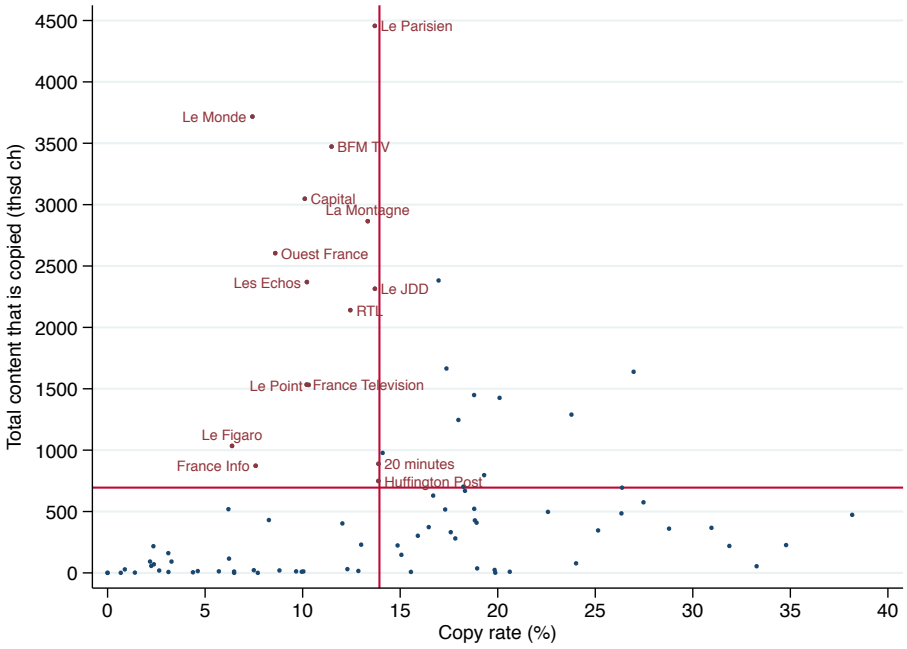
Notes: The figure plots the average copy rate of the articles depending on the type of media outlet that publishes the articles: local newspapers, national newspapers, television, radio, and pure online media. The blue bars represent total copy rate; the red bar the external copy rate (i.e. copy from articles published by another media outlet); the green bars the external copy rate from media outlets other than the news agencies (the AFP and Reuters); and the orange bars the copy rate from the AFP and Reuters.

Figure 7: Average copy rate depending on the media type

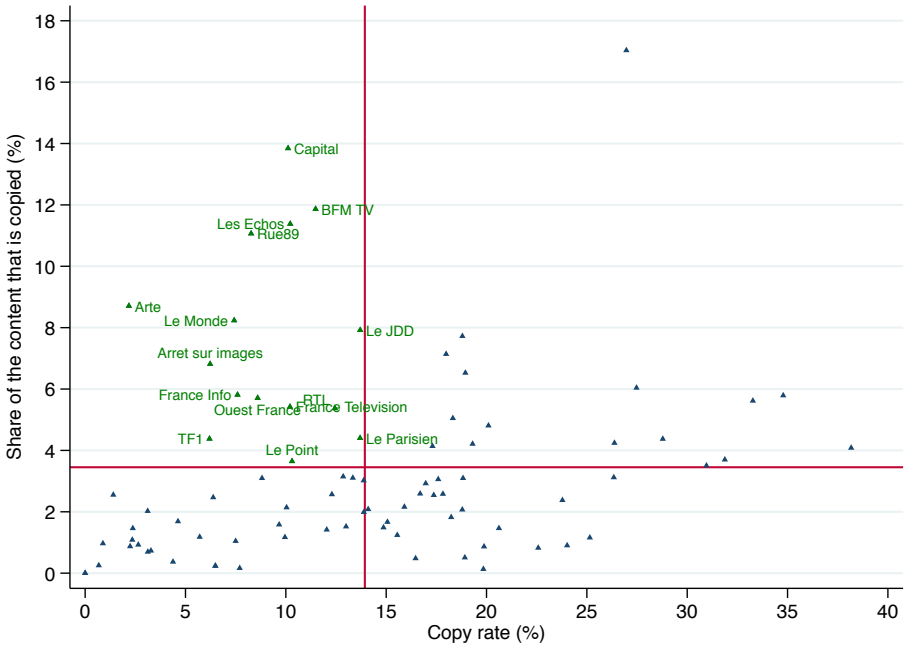


Notes: The figure plots the average external copy rate excluding content copied from the news agencies (the AFP and Reuters) of the articles published by the media outlets in our sample.

Figure 8: Average external copy rate, excluding content copied from the news agencies: media-level analysis



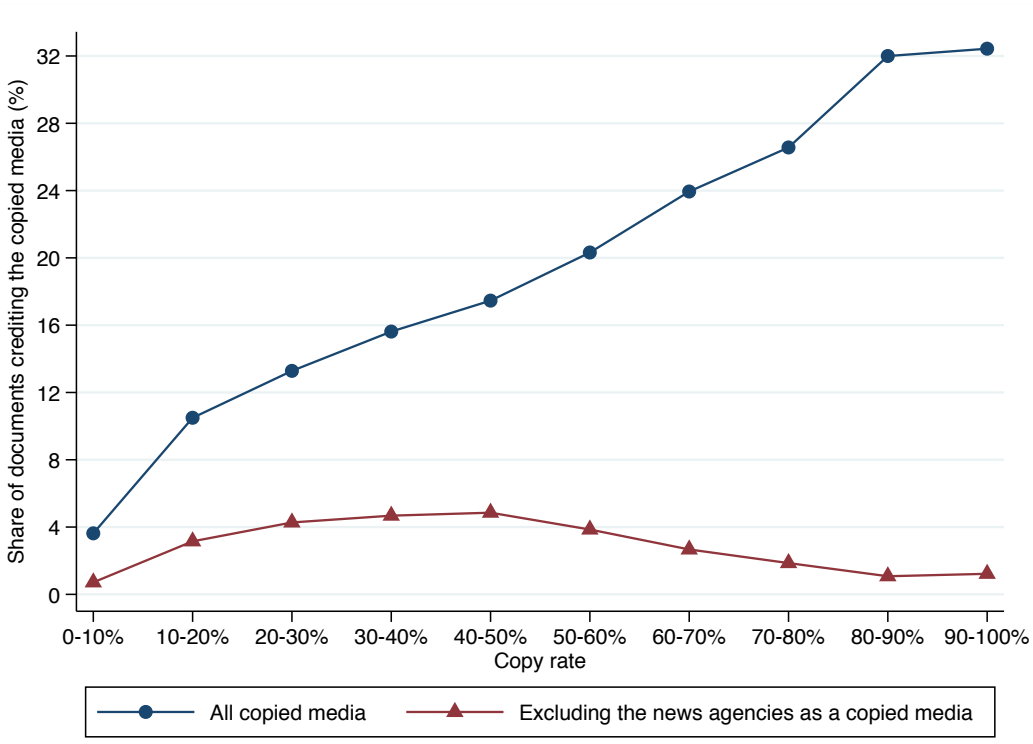
(a) Total content copied by at least one other media outlet



(b) Share of the content copied by at least one other media outlet

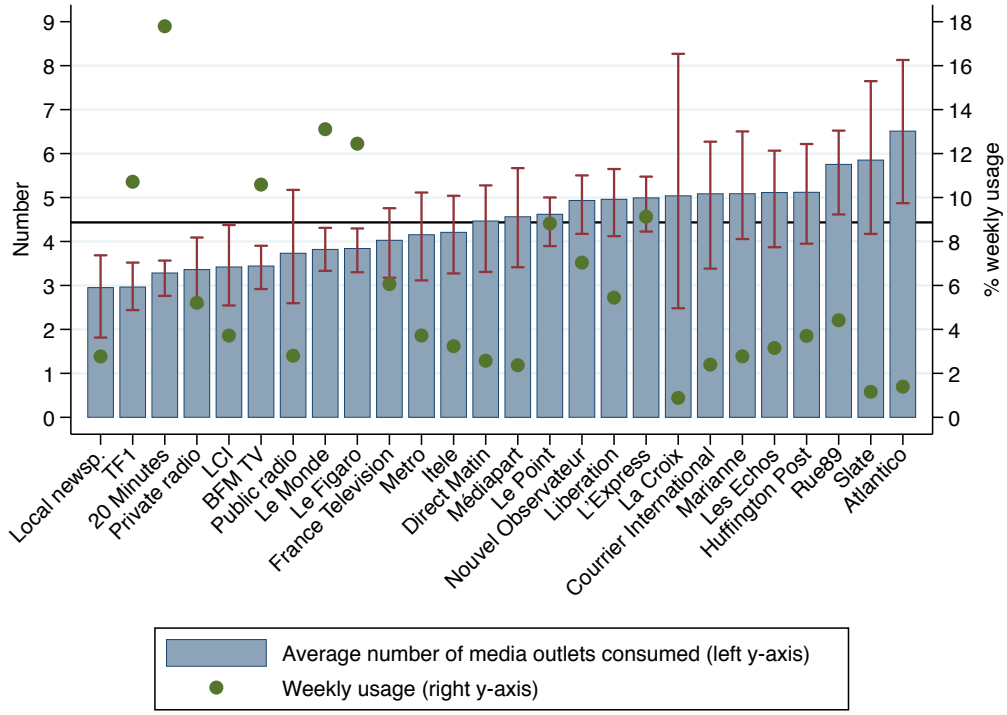
Notes: The upper Figure 9a plots the external copy rate on the x-axis and the total content that is produced by each media outlet in our sample and copied by at least one other media outlet on the y-axis. The x-axis is unchanged in the bottom Figure 9b, but we report on the y-axis the share of the content that is produced by each media outlet and copied by at least one other media outlet. The copy rate is the external copy rate excluding content reproduced from the news agencies (the AFP and Reuters).

Figure 9: Probability of been copied: media-level analysis



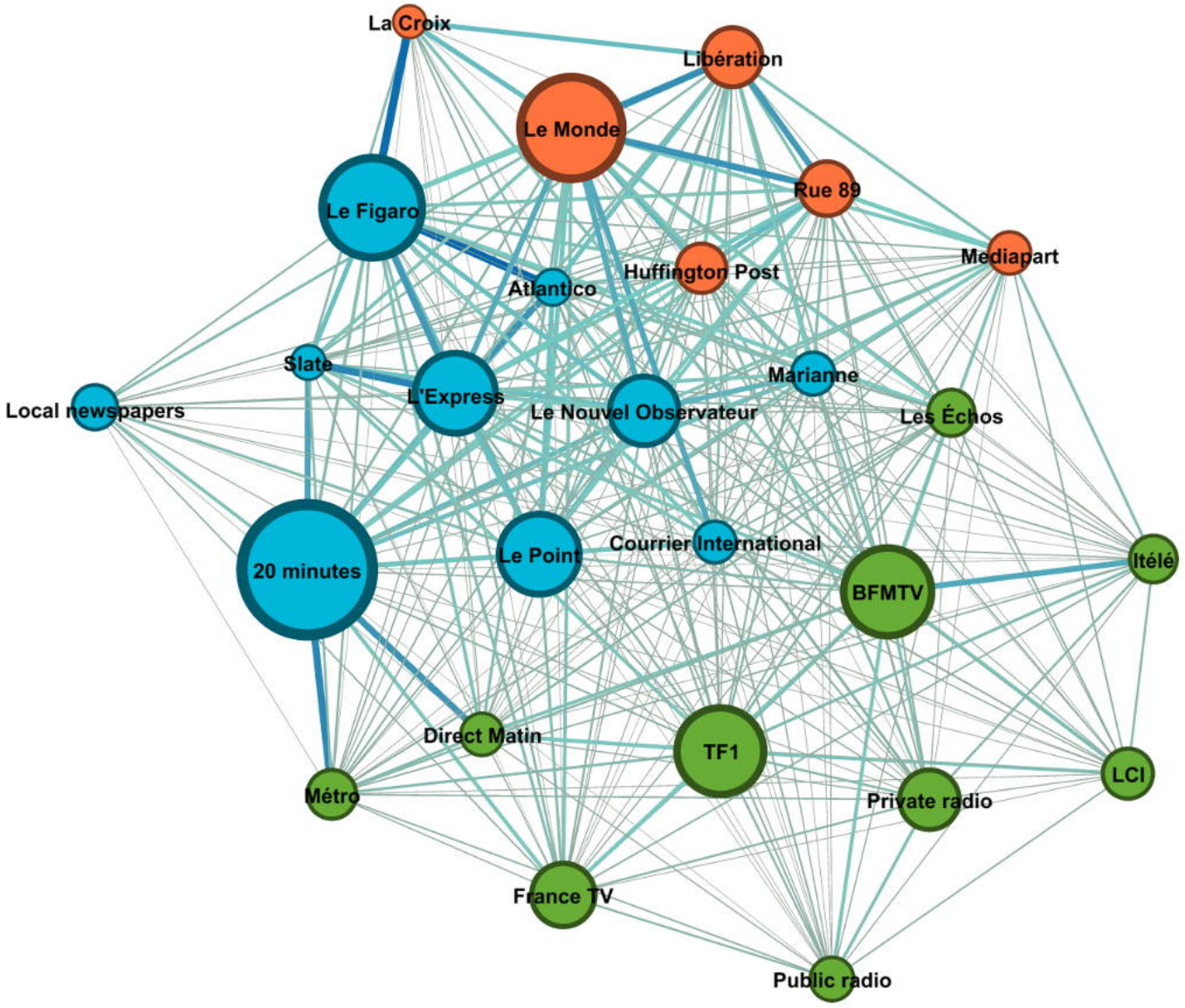
Notes: The figure plots the share of the documents crediting the copied media as a function of the copy rate. We define 10 different intervals for the copy rate: below 10%, between 10 and 20%,..., between 90 and 100%. The share is computed for each of these intervals.

Figure 10: Share of documents crediting the copied media as a function of the copy rate



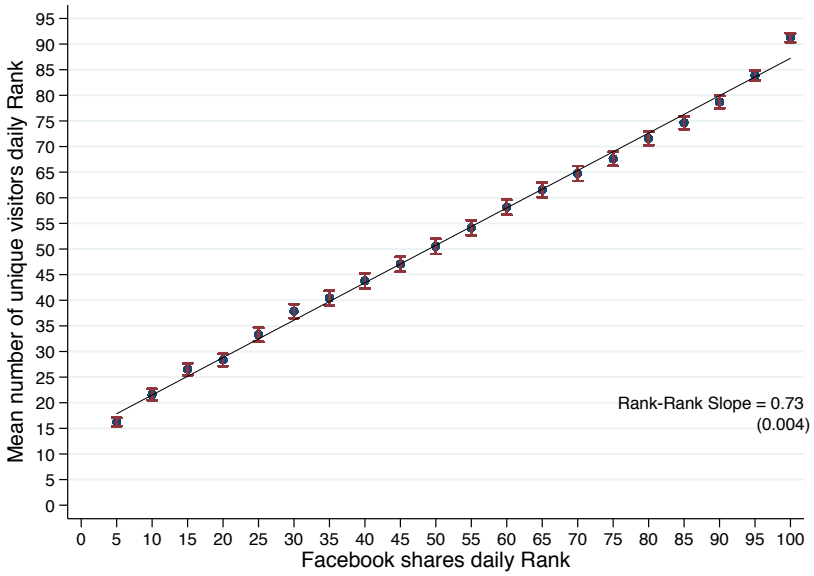
Notes: The Figure reports the average number of media outlets consumed by users who have accessed news via a given media outlet (blue bars, left y-axis). Error bars are ±1.96 standard errors. E.g., on average, users who have accessed news via the website of TF1 have consumed news online from 3 different media outlets. It also reports the weekly online media consumption for each of the outlet (green dots, right y-axis). E.g., 10.7% of the survey respondents have used the website of TF1 to access news in the last week. Data are from the *2013 Digital News Report* (Reuters Institute, 2013). The sample includes 1,016 individuals for France for the year 2013.

Figure 11: Average number of media outlets consumed by users who have accessed news via a given media outlet

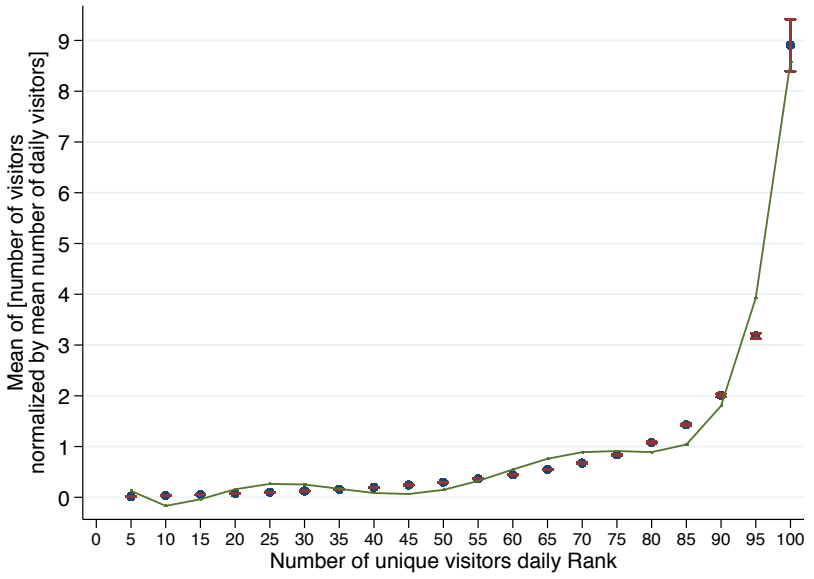


Notes: The Figure illustrates the proximity across French media outlets using the 2013 patterns of online readership. The data are from the 2013 Digital News Report (Reuters Institute, 2013), and the observations used to draw the graph are at the surveyed individual level. The size of the circles represents the audience of the media outlets (the larger the circle, the higher the audience), and the size of the rows between two websites the probability that a respondent accessing a website also accesses the other one (the thicker the row, the higher this probability). We use Gephi, a network analysis and visualization software package, to draw this graph.

Figure 12: Proximity across outlets using survey data on patterns of online readership



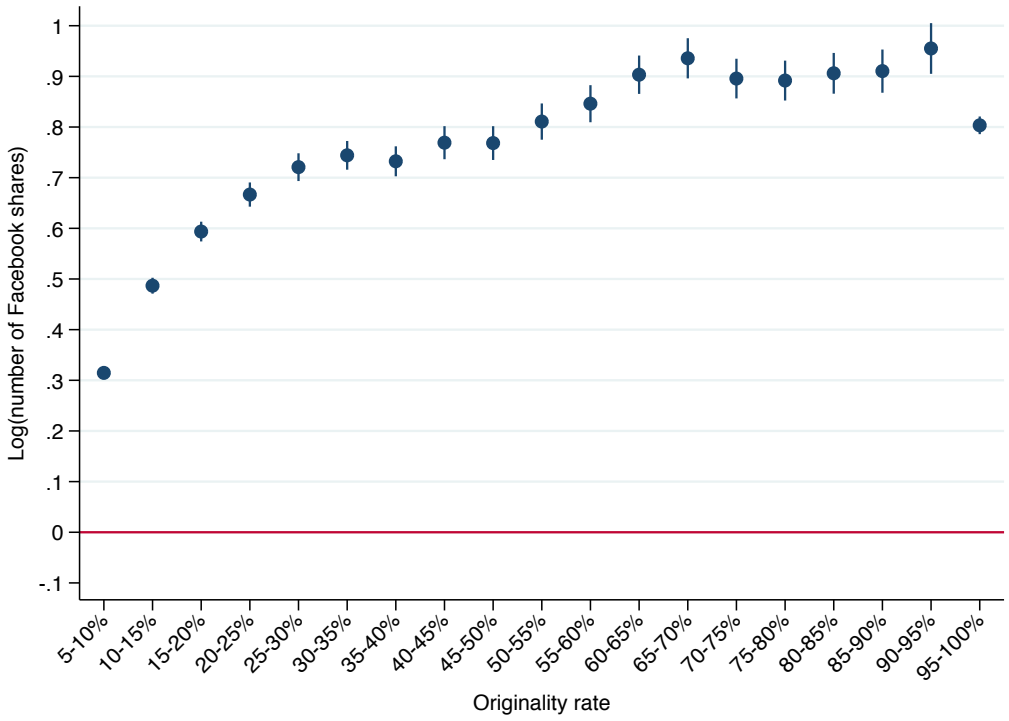
(a) **Association between number of Unique visitors' and Facebook shares' Percentile Ranks**



(b) **Association between number of Unique visitors' Percentile Rank and Number of Unique visitors (as a multiple of the average number of daily visitors)**

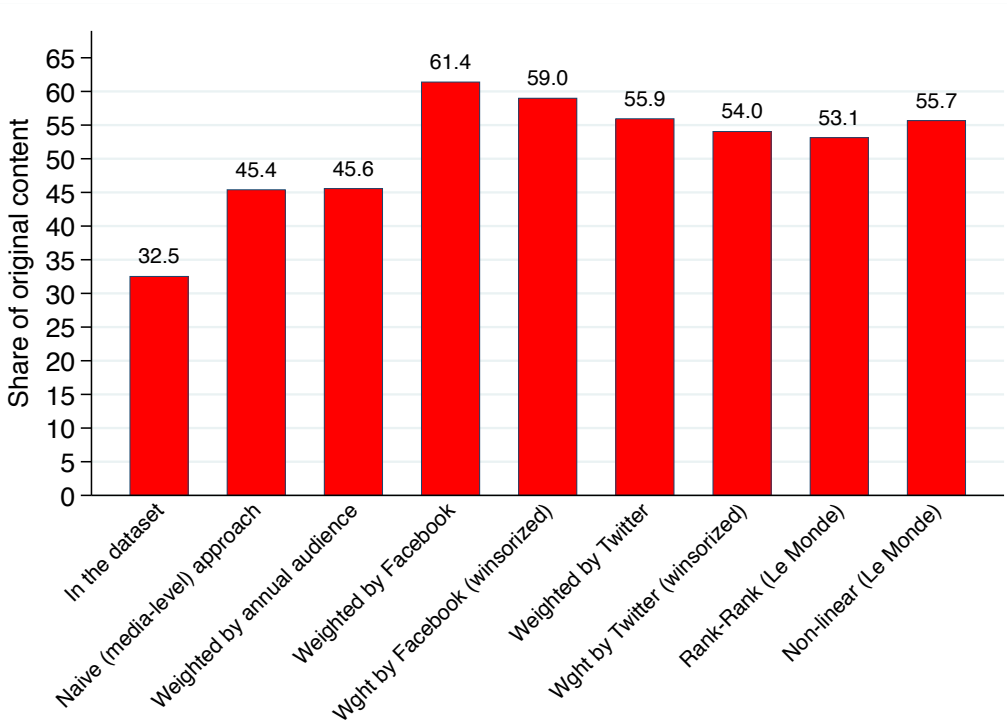
Notes: The Figure investigates the relationship between the number of unique visitors and the number of Facebook shares, using article-level information from the national daily newspaper *Le Monde*. The data includes all the articles published by *Le Monde* between April and August 2017 (17, 314 articles). In the upper Figure 13a, we plot the relationship between the articles' Facebook shares' percentile rank and the average value of the visitors percentile rank (error bars in red represent the 95% confidence interval). The slope of this relationship is equal to 0.73. In the bottom Figure 13b, we plot the relationship between the rank in the number of visitors distribution and the average number of visitors as a multiple of the mean number of daily visitors (error bars in red represent the 95% confidence interval). The green line is the predicted value of the average number of visitors when this relationship is approximated by a polynomial of degree six.

Figure 13: Relationship between the number of Unique visitors and the number Facebook shares, using article-level information from the national daily newspaper *Le Monde*, April-August 2017



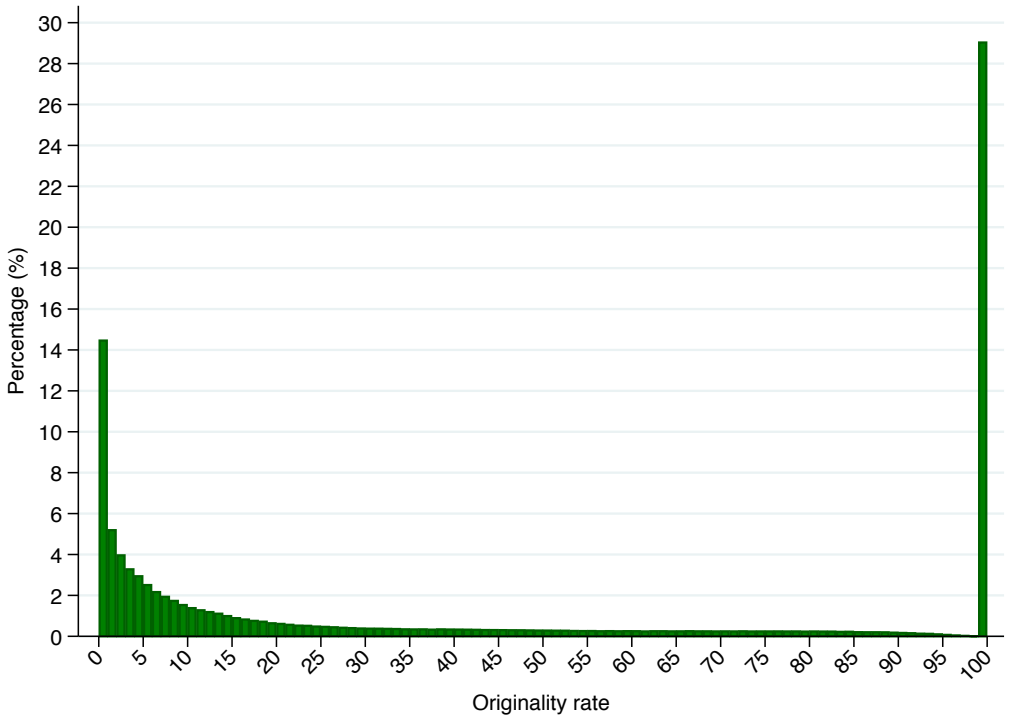
Notes: Figure shows coefficients from a regression of log of the number of times an article is shared on Facebook on twenty categorical variables depending on the originality rate of the articles (articles with an originality rate lower than 5% are the omitted category). Models include media, day and event fixed effects. Error bars are ±1.96 standard errors. Standard errors are clustered by event. The unit of observation is an article.

Figure 14: Facebook shares and originality rate



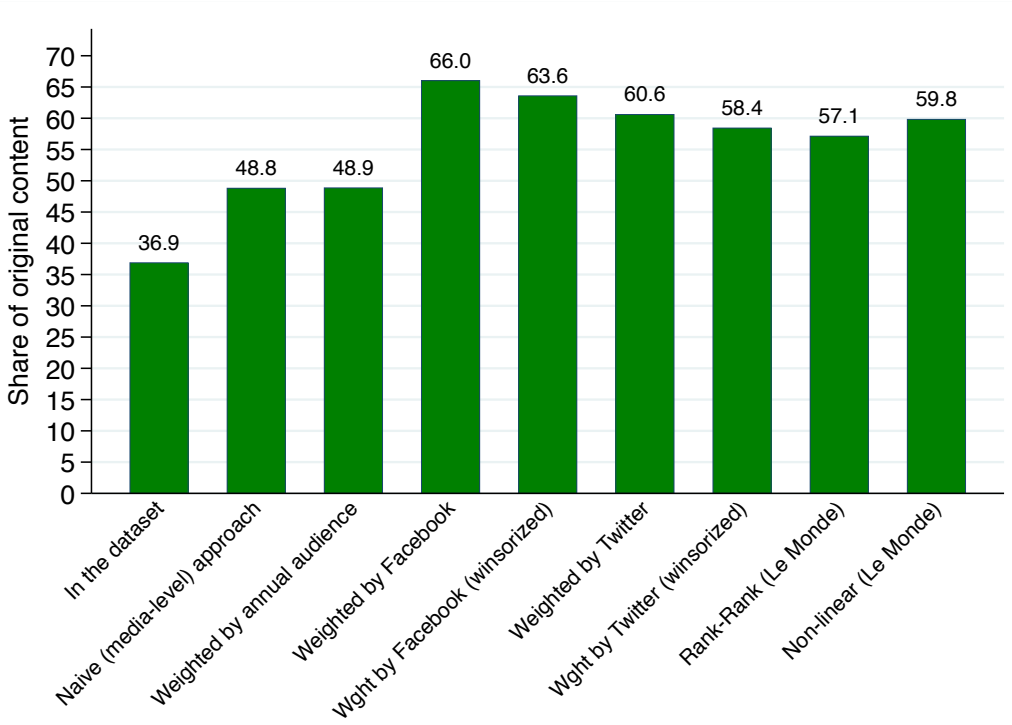
Notes: The Figure reports the audience-weighted share of original content we obtain using our different approaches to compute article-level number of views. The first bar ("In the dataset") simply reports the share of original content in the dataset (with no weight). The second bar ("Naive (media-level) approach") reports the share of original content we obtain when we attribute the same number of views to all the articles published by a media outlet on a given date. The third bar ("Weighted by annual audience") reports the share of original content we obtain when we weight the content of each article by the average annual audience of the media outlet in which it was published. To compute the fourth bar ("Weighted by Facebook"), we attribute number of views to articles assuming a linear relationship between the number of Facebook shares and the number of article views. The fifth bar ("Wght by Facebook (winsorized)") relies on the same methodology but with the winsorized version of the Facebook shares variable. The sixth ("Weighted by Twitter") and seventh ("Wght by Twitter (winsorized)") bars are computed similarly than the fourth and fifth bars, except that we use the number of shares on Twitter rather than on Facebook. To compute the number of views at the article level, the eighth bar ("Rank-Rank (Le Monde)") relies on the parameters obtained from the analysis of the joint distribution of the number of Facebook shares and the number of visitors using *Le Monde*'s data (April-August 2017). Finally, the ninth bar ("Non-linear (Le Monde)") also uses *Le Monde*'s data but relies on the parameters obtained when regressing the share of the total number of visits represented by each article on its share of the total number of Facebook shares (using a polynomial of degree six). The different methodologies used are described in details in the text.

Figure 15: The audience-weighted share of original content



Notes: The figure plots the distribution of the originality rate (with bins equal to one percent). Events are defined without imposing a minimum number of documents per event.

Figure 16: Originality rate: Relaxing the "10 documents condition"



Notes: The Figure reports the audience-weighted share of original content we obtain using our different approaches to compute article-level number of views. Compared to Figure 15, events here are defined without imposing a minimum number of documents per event. The first bar ("In the dataset") simply reports the share of original content in the dataset (with no weight). The second bar ("Naive (media-level) approach") reports the share of original content we obtain when we attribute the same number of view to all the articles published by a media outlet on a given date. The third bar ("Weighted by annual audience") reports the share of original content we obtain when we weight the content of each article by the average annual audience of the media outlet in which it was published. To compute the fourth bar ("Weighted by Facebook"), we attribute number of views to articles assuming a linear relationship between the number of Facebook shares and the number of article views. The fifth bar ("Wght by Facebook (winsorized)") relies on the same methodology but with the winsorized version of the Facebook shares variable. The sixth ("Weighted by Twitter") and seventh ("Wght by Twitter (winsorized)") bars are computed similarly than the fourth and fifth bars, except that we use the number of shares on Twitter rather than on Facebook. To compute the number of views at the article level, the eighth bar ("Rank-Rank (Le Monde)") relies on the parameters obtained from the analysis of the joint distribution of the number of Facebook shares and the number of visitors using *Le Monde*'s data (April-August 2017). Finally, the ninth bar ("Non-linear (Le Monde)") also uses *Le Monde*'s data but relies on the parameters obtained when regressing the share of the total number of visits represented by each article on its share of the total number of Facebook shares (using a polynomial of degree six). The different methodologies used are described in details in the text.

Figure 17: The audience-weighted share of original content, Relaxing the "10 documents condition"

Table 1: Summary statistics: Articles (classified in events)

	Mean	Median	sd	Min	Max
Content					
Length (number of characters)	2,467	2,192	1,577	100	98,340
Original content (number of characters)	805	253	1,287	1	53,424
Non-original content (number of characters)	1,661	1,326	1,539	0	48,374
Originality (%)	36.5	14.5	39.8	0	100
Reactivity in hours	41.7	19.1	65.2	0	6,257
Audience					
Number of shares on Facebook	64	0	956	0	240,450
Number of shares on Facebook (winsorized)	37	0	136	0	1,017
Number of shares on Twitter	9	0	42	0	11,908
Number of shares on Twitter (winsorized)	7	0	19	0	126
Obs	851,864				

Notes: The table gives summary statistics. Year is 2013. Variables are values for the articles classified in events. The observations are at the article level. The "Number of shares on Facebook (winsorized)" variable is the version of the Facebook variable winsorized at the 99th percentile. Similarly, the "Number of shares on Twitter (winsorized)" variable is the version of the Twitter variable winsorized at the 99th percentile. Variables are described in more details in the text.

Table 2: Summary statistics: Media outlets

	Mean	Median	sd	Min	Max
Online audience (daily)					
Number of unique visitors	248,529	107,856	384,001	3,689	2,031,580
Number of visits	340,506	156,735	543,690	4,650	2,945,172
Number of pages views	1,617,616	647,576	2,956,979	12,203	15,203,845
Audience share	1.66	0.72	2.57	0.02	13.65
Facebook (annual)					
Total number of shares	1,137,580	309,176	2,190,098	1,066	13,459,510
Twitter (annual)					
Total number of direct tweets	138,648	27,188	343,000	0	2,464,651
Total number of indirect tweets	3,627	577	8,792	0	58,507
Content (nb of characters) **(annual)**					
Total content not classified	32,255,744	14,999,537	114,887,872	419,234	1,065,079,616
Total content classified	19,708,659	11,580,943	23,729,089	1,114	101,246,288
Total original content	6,381,766	3,787,462	7,395,088	1,114	31,799,058
Total non-original content	13,326,893	6,860,454	19,705,976	0	76,923,528
Number of breaking news	115	54	174	0	1,011
Observations	85				

Notes: The table gives summary statistics. Year is 2013. Variables are values for media outlets (excepting the AFP and Reuters). The observations are at the media outlet/day level for the online audience statistics (first four rows) at the media outlet/year level for the total number of Facebook shares and the content data.

Table 3: Reaction time

(a) **Depending on the offline format of the news breaker**

	Mean	sd	Median	Min	Max	Obs
Reaction time (in minutes)	169	358	22	0	2,809	25,215
If news breaker is						
Print media	247	400	73	0	2,809	7,201
Television	231	391	57	0	2,098	1,135
Radio	248	398	76	0	2,191	964
Pure online media	394	473	190	0	2,164	510
News agency	116	314	11	0	2,624	15,405

(b) **Depending on the nature of the news event**

				Differences		
	(1)	(2)	(3)	(4)	(5)	(6)
	Non-exclusive	Exclusive	Mult wit	Non-exc vs. Exc	Non-exc vs. Mult	Exc vs. Mult
	mean/sd	mean/sd	mean/sd	b/t	b/t	b/t
	170.5	183.8	151.6	13.3	-19.0**	32.2**
	(361.4)	(372.6)	(331.8)	(1.3)	(-2.0)	(2.5)
Obs	16,473	1,286	1,595	17,759	18,068	2,881

Notes: * p<0.10, ** p<0.05, *** p<0.01. The tables give summary statistics for the reaction time (in minutes). The upper table 3a presents the results for all the events in our sample, as well as depending on the offline format of the news breaker. The bottom table 3b provides the reaction time depending on the nature of the news event. Column 1 presents the results for non-exclusive news events. Column 2 presents the results for exclusive news events. Column 3 presents the results for short news items with multiple witnesses. In columns 4 to 6, we perform a t-test on the equality of means respectively for non-exclusive versus exclusive news events, non-exclusive versus short news items with multiple witnesses, and exclusive versus short news items with multiple witnesses (robust standard errors are in parentheses).

Table 4: Summary statistics: Copy

	All copy			External copy	Excl. copy from agencies
	(1)	(2)	(3)	(4)	(5)
	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd
Originality rate	35.1				
	(39.0)				
Originality rate wghtd by nb of views (Facebook)		58.0			
		(38.2)			
Copying media					
Nb docs copied			4.1	3.9	2.3
			(5.0)	(4.7)	(3.3)
External copy rate				61.0	15.9
				(39.3)	(25.0)
External copy rate conditional on copying				78.7	25.7
				(24.5)	(27.6)
Copied media					
Nb copying docs			3.9	3.3	
			(9.1)	(8.2)	
% of the doc that is copied				3.9	
				(12.5)	
% of the doc that is copied conditional on being copied				9.4	
				(17.9)	
If copied media bk news					
% of the doc that is copied				24.1	
				(33.9)	
% of the doc that is copied conditional on being copied				53.5	
				(31.3)	

Notes: The table gives summary statistics. Year is 2013. Variables are values for documents. We consider all the documents classified in events, with the exception of the documents published by the AFP and Reuters. In columns (1) to (3), both internal and external verbatim copying are taken into account. In column (4), we focus on external copy only. In column (5), we focus on external copy and exclude the content copied from the news agencies (the AFP and Reuters). "bk news" stands for breaking news. The different variables are described in detail in the text.

Table 5: Article-level analysis: Number of Facebook shares, of Tweets and of predicted readers (log-linear estimation)

	Facebook shares			Number of tweets			Number of predicted readers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Publication rank	-0.0005***	-0.0005***		-0.0002***	-0.0002***		-0.0005***	-0.0005***	
	(0.0001)	(0.0001)		(0.0000)	(0.0000)		(0.0001)	(0.0001)	
Reaction time (in hours)	-0.0023***	-0.0023***		-0.0021***	-0.0021***		-0.0045***	-0.0045***	
	(0.0006)	(0.0006)		(0.0003)	(0.0003)		(0.0007)	(0.0007)	
Originality rate (%)	0.0068***			0.0032***			0.0074***		
	(0.0001)			(0.0001)			(0.0001)		
Length (thsd ch)	0.0855***		0.0808***	0.0593***		0.0573***	0.0796***		0.0768***
	(0.0025)		(0.0025)	(0.0014)		(0.0014)	(0.0030)		(0.0030)
Original content (thsd ch)		0.1989***			0.1083***			0.2083***	
		(0.0033)			(0.0018)			(0.0038)	
Non-original content (thsd ch)		-0.0296***			0.0087***			-0.0439***	
		(0.0025)			(0.0014)			(0.0032)	
News breaker			0.7694***			0.4569***			0.8129***
			(0.0202)			(0.0134)			(0.0232)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.50	0.50	0.49	0.60	0.60	0.60	0.50	0.50	0.49
Adjusted R-sq	0.48	0.48	0.47	0.58	0.58	0.58	0.47	0.47	0.46
Observations	664,650	664,650	664,650	656,129	656,129	656,129	509,378	509,378	509,378
Clusters (event)	25,200	25,200	25,200	25,200	25,200	25,200	25,109	25,109	25,109

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is shared on Facebook in Columns (1) to (3), the log of the number of times an article is tweeted, retweeted or liked in Columns (4) to (6), and the log of the number of views per article in Columns (7) to (9). The number of views per article is computed by combining media-level information on the daily number of page views with article-level information on the number of times an article is shared on Facebook (as detailed in Section 5.3). Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters".

Table 6: Article-level analysis: Number of Facebook shares, of Tweets, and of article views (log-linear estimation), Heterogeneity of the effects depending on the competitiveness of the environment

	Facebook shares		Tweets		Number of views	
	(1)	(2)	(3)	(4)	(5)	(6)
Publication rank	-0.0005***	-0.0004***	-0.0002***	-0.0002***	-0.0004***	-0.0003***
	(0.0001)	(0.0001)	(0.0000)	(0.0001)	(0.0001)	(0.0001)
Publication rank * High competition		-0.0001***		0.0001***		-0.0001*
		(0.0000)		(0.0000)		(0.0001)
Reaction time	-0.0135***	-0.0140***	-0.0025***	-0.0025***	-0.0223***	-0.0228***
	(0.0008)	(0.0008)	(0.0005)	(0.0005)	(0.0011)	(0.0011)
Reaction time * High competition		0.0002		0.0001		0.0002
		(0.0002)		(0.0001)		(0.0003)
Originality rate	0.0063***	0.0052***	0.0024***	0.0022***	0.0072***	0.0064***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0002)
Originality rate * High competition		0.0032***		0.0006***		0.0023***
		(0.0002)		(0.0001)		(0.0003)
Length	0.1001***	0.1041***	0.0736***	0.0751***	0.0956***	0.0827***
	(0.0039)	(0.0057)	(0.0021)	(0.0029)	(0.0055)	(0.0075)
Length * High competition		-0.0079		-0.0040		0.0307***
		(0.0069)		(0.0033)		(0.0095)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.39	0.39	0.54	0.54	0.26	0.26
Observations	318,196	318,196	310,512	310,512	213,718	213,718
Clusters (event)	24,691	24,691	24,691	24,691	23,846	23,846

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is shared on Facebook in Columns (1) and (2), and the log of the number of times an article is shared on Twitter in Columns 3 and 4. Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. "High competition" is an indicator variable equal to 1 for the media outlets that have been highly copied in 2013, i.e. whose share of the content that has been copied is higher than the median, and to 0 otherwise (see the text for more details). All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters".

Table 7: Article-level analysis: Number of Facebook shares, of Tweets, and of article views (log-linear estimation), Heterogeneity of the effects depending on whether the media outlet is copied

	Facebook shares		Tweets		Number of views	
	(1)	(2)	(3)	(4)	(5)	(6)
Publication rank	-0.0005***	-0.0005***	-0.0002***	-0.0002***	-0.0005***	-0.0005***
	(0.0001)	(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0001)
Publication rank * Highly copied		-0.0000		0.0000		-0.0001*
		(0.0000)		(0.0000)		(0.0001)
Reaction time	-0.0023***	-0.0019***	-0.0021***	-0.0018***	-0.0045***	-0.0044***
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0007)	(0.0007)
Reaction time * Highly copied		-0.0003*		-0.0004***		0.0001
		(0.0002)		(0.0001)		(0.0002)
Originality rate	0.0068***	0.0083***	0.0032***	0.0052***	0.0074***	0.0087***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Originality rate * Highly copied		-0.0022***		-0.0031***		-0.0019***
		(0.0002)		(0.0001)		(0.0002)
Length	0.0855***	0.0644***	0.0593***	0.0505***	0.0796***	0.0680***
	(0.0025)	(0.0030)	(0.0014)	(0.0017)	(0.0030)	(0.0041)
Length * Highly copied		0.0405***		0.0175***		0.0212***
		(0.0038)		(0.0022)		(0.0049)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.40	0.41	0.54	0.54	0.42	0.42
Observations	664,650	664,627	656,129	656,106	509,378	509,378
Clusters (event)	25,200	25,200	25,200	25,200	25,109	25,109

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is shared on Facebook in Columns 1 and 2 an the log of the number of times an article is shared on Twitter in Columns 3 and 4. Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. "Highly copied" is an indicator variable equal to 1 for the media outlets that have been highly copied in 2013, i.e. whose share of the content that has been copied is higher than the median, and to 0 otherwise (see the text for more details). All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters".

Table 8: Article-level analysis: Number of Facebook shares (log-linear estimation), Heterogeneity of the effects depending on the topic of the events

	All	Crime	Politics	Economy	Arts	Sport	Disaster	Other
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Publication rank	-0.0005***	-0.0008***	-0.0003***	-0.0006***	-0.0025***	-0.0010***	-0.0027***	-0.0010***
	(0.0001)	(0.0001)	(0.0001)	(0.0002)	(0.0003)	(0.0003)	(0.0004)	(0.0002)
Reaction time	-0.0023***	-0.0064***	-0.0030***	-0.0072***	-0.0073***	0.0088***	-0.0077***	-0.0015
	(0.0006)	(0.0013)	(0.0010)	(0.0012)	(0.0017)	(0.0012)	(0.0023)	(0.0013)
Originality rate (%)	0.0068***	0.0080***	0.0075***	0.0055***	0.0067***	0.0048***	0.0066***	0.0068***
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0004)	(0.0002)
Length	0.0855***	0.0982***	0.0757***	0.0978***	0.1012***	0.0316***	0.1331***	0.0928***
	(0.0025)	(0.0065)	(0.0058)	(0.0048)	(0.0063)	(0.0054)	(0.0081)	(0.0052)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.50	0.48	0.53	0.48	0.52	0.50	0.49	0.53
Adjusted R-sq	0.48	0.47	0.52	0.46	0.49	0.46	0.47	0.51
Observations	664,650	153,658	121,031	121,054	63,795	56,381	35,038	113,693
Clusters (event)	25,200	5,132	3,422	5,093	2,854	2,940	1,065	4,694

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is shared on Facebook. In columns (1), all the events in our sample are included in the estimation. Column (2) reports the estimates for the "Crime, law and justice" events, Column (3) for the "Politics" events, Column (4) for the "Economy, business and finance" events, Column (5) for the "Arts, culture and entertainment" events, Column (6) for the "Sport" events, Column (7) for the "Disaster and accident" events, and Column (8) for the events classified in all the other IPTC categories. The topics correspond to the IPTC media topics described in the article and defined in the online Appendix Section C.5. Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters".

Table 9: Article-level analysis: Number of Facebook & of Twitter shares (log-linear estimation), Relaxing the "10 documents condition"

	Facebook shares			Number of tweets			Number of predicted readers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Publication rank	-0.0004***	-0.0004***		-0.0001***	-0.0001***		-0.0005***	-0.0004***	
	(0.0001)	(0.0001)		(0.0000)	(0.0000)		(0.0001)	(0.0001)	
Reaction time (in hours)	-0.0011**	-0.0011**		-0.0025***	-0.0025***		-0.0031***	-0.0031***	
	(0.0005)	(0.0005)		(0.0003)	(0.0003)		(0.0007)	(0.0006)	
Originality rate (%)	0.0067***			0.0030***			0.0073***		
	(0.0001)			(0.0000)			(0.0001)		
Length (thsd ch)	0.0897***		0.0866***	0.0585***		0.0572***	0.0828***		0.0816***
	(0.0025)		(0.0026)	(0.0014)		(0.0014)	(0.0030)		(0.0031)
Original content (thsd ch)		0.1927***			0.1005***			0.2024***	
		(0.0036)			(0.0020)			(0.0038)	
Non-original content (thsd ch)		-0.0358***			0.0062***			-0.0514***	
		(0.0026)			(0.0014)			(0.0033)	
News breaker			0.4540***			0.2513***			0.4778***
			(0.0097)			(0.0062)			(0.0120)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.54	0.54	0.53	0.63	0.63	0.62	0.54	0.54	0.53
Adjusted R-sq	0.48	0.48	0.47	0.57	0.57	0.57	0.46	0.46	0.45
Observations	920,173	920,173	920,173	909,439	909,439	909,439	703,729	703,729	703,729
Clusters (event)	112,804	112,804	112,804	112,799	112,799	112,799	104,404	104,404	104,404

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is shared on Facebook in Columns (1) to (3), the log of the number of times an article is shared on Twitter in Columns (4) to (6), and the log of the predicted number of readers in Columns (7) to (9). Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters".