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

Julia Cage, Nicolas Hervé, Marie-Luce Viaud

DISCUSSION PAPER SERIES

DP12066

THE PRODUCTION OF INFORMATION IN AN ONLINE WORLD: IS COPY RIGHT?

Julia Cagé, Nicolas Hervé and Marie-Luce Viaud

INDUSTRIAL ORGANIZATION and PUBLIC ECONOMICS

ISSN 0265-8003

THE PRODUCTION OF INFORMATION IN AN ONLINE WORLD: IS COPY RIGHT?

Julia Cagé, Nicolas Hervé and Marie-Luce Viaud

Discussion Paper DP12066
Published 29 May 2017
Submitted 18 December 2017

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

THE PRODUCTION OF INFORMATION IN AN ONLINE WORLD: IS COPY RIGHT?

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 (newspaper, television, radio, pure online media, and a news agency) during the year 2013 with new micro audience data. We develop a topic detection algorithm that identifies each news event, we trace the timeline of each story and study news propagation. We unravel new evidence on online news production. First, we show that one quarter of the news stories are reproduced online in less than 4 minutes. Second, we find that only 32:6% of the online content is original. Third, we show that reputation effects partly counterbalance the negative impact of plagiarism on newsgathering incentives. By using media-level daily audience and article-level social media statistics (Facebook and Twitter shares), we find that original content represents between 54 and 62% of online news consumption. Reputation mechanisms actually appear to solve about 30 to 40% of the copyright violation problem.

JEL Classification: L11, L15, L82, L86

Keywords: internet, Information spreading, Copyright, Investigative journalism, Facebook, reputation

Julia Cagé - julia.cage@sciencespo.fr
Sciences Po Paris and CEPR

Nicolas Hervé - nherve@ina.fr
Institut National de l'Audiovisuel

Marie-Luce Viaud - mlviaud@ina.fr
Institut National de l'Audiovisuel

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

December 2017

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 (newspaper, television, radio, pure online media, and a news agency) during the year 2013 with new micro audience data. We develop a topic detection algorithm that identifies each news event, we trace the timeline of each story and study news propagation. We unravel new evidence on online news production. First, we show that one quarter of the news stories are reproduced online in less than 4 minutes. Second, we find that only 32.6% of the online content is original. Third, we show that reputation effects partly counterbalance the negative impact of plagiarism on newsgathering incentives. By using media-level daily audience and article-level social media statistics (Facebook and Twitter shares), we find that original content represents between 54 and 62% of online news consumption. Reputation mechanisms actually appear to solve about 30 to 40% of the copyright violation problem.

Keywords: Internet, Information spreading, Copyright, Investigative journalism, social media, Facebook, Twitter, Reputation

JEL No: L11, L15, L82, L86

*We gratefully acknowledge the many helpful comments and suggestions from Yasmine Bekkouche, Filipe Campante, Lucien Castex, Etienne Fize, Matthew Gentzkow, Sergei Guriev, Emeric Henry, Ali Hortacsu, Elise Huillery, Laurent Joyeux, 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. 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, and Sciences Po's Scientific advisory board (SAB). 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 responsability 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]

This paper documents the extent of copying online and estimates 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. In this paper, we attempt to open up this black box by using new micro data and relying on a machine-learning approach. We examine the main French news media – including newspapers, television channels, radio stations, pure online media and the French news agency Agence France Presse (AFP) – and track every piece of content these outlets produced online in 2013. Our dataset contains 2.5 million documents. To the extent of our knowledge, it is the very first time that such a transmedia approach has been adopted, covering the integrality of the content produced by media online, whatever their offline format.[3]

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 out the story, and then analyze the propagation of the story, second-by-second. We investigate the speed of news dissemination and the length

[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] As we will see, on the Internet, there is a tendency of different media to converge, and it becomes increasingly difficult to classify a media outlet as belonging purely to one of the traditional media formats. The general structure of our dataset is illustrated by the online Appendix Figure F.1, where we plot the total original content and number of journalists for each media outlet. One can see the general positive relationship between these two variables, and the special role played by the AFP. In a companion paper (Cagé et al., 2017), we further investigate the structure of the production function for online news. In the present paper, we leave aside the input side and focus upon the pattern of copying between media outlets.

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. Because some stories are not the product of original reporting – e.g. in the case of a government press release giving rise to a number of articles, the first outlet covering the story cannot be considered as a news breaker providing exclusive news – we also code manually all the news stories in our sample and isolate the stories that are the result of a piece of original reporting.

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).

We show that on average news is delivered to readers of different media outlets 172 minutes after having been published first on the website of the news breaker, but in less than 4 minutes (224 seconds) in 25% of the cases. The reaction time is the shortest when the news breaker is the news agency, and the longest when it is a pure online media, most likely because of the need for verification.

High reactivity comes with verbatim copying. We find that only 32.6% of the online content is original. In effect, 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. Obviously in practice, we often observe large numbers 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[4] and that on average, conditional on being copied, 21% of the content of a document is copied. But in terms of numbers of original characters copied, this is equivalent

[4]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.

to a situation where each piece of original content is published three times.[5] Moreover, despite the substantiality of copying, media outlets hardly name the sources they copy: once we exclude copy from the news agency, we show that only 3.5% of the documents mention 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 negatively affect media outlets' newsgathering incentives. In case 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, e.g. via online advertising revenues). However, we show that reputation mechanisms and the behavior of Internet viewers allow to mitigate a significant part of this copyright violation problem.[6]

First, 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 39% increase in the number of times it is shared on Facebook. If we rather consider the number of times the article is shared on Twitter, we find that a thousand increase in the number of original characters leads to a 11% increase in the number of Tweets.

Second, 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 Faceboo 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. Depending on the specification we use, we find that the original content represents between 54 and 62% of online news consumption, i.e. much more than its relative production. Reputation mechanisms actually appear to solve about 30 to 40% of the copyright violation problem.

[5] Furthermore, 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.

[6] Through a small misuse of language and for the sake of simplicity, we call here "copyright violation" the external verbatim copying we capture in the data. In a previous version of this article, we documented the extent to which verbatim copying falls outside the bounds of copyright law. Online Appendix Section A provides a quick overview of copyright laws in the context of information production by news media. National copyright laws act within the framework imposed by international agreements (Ginsburg, 2016). They rely on two main principles. First, copyright excludes ideas; it protects only the form of expression in which the ideas are communicated. The form of expression is what we capture in this paper by identifying and quantifying verbatim copying. Second, to violate the exclusive right of reproduction, copying should be "substantial". In this article, we quantify substantiality quantitatively, both from the point of view of the copying and of the copied media outlet.

Our results are robust to the use of different functional forms, as well as different thresholds (e.g. regarding the minimum amount of articles to define an event) for the algorithms we implement to detect media events.

Of course, greater intellectual property protection could also play a role in solving the copyright violation problem and 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) has been 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, independently of 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. Third and most importantly, we quantify the importance of plagiarism online and 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., 2015). 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] To the exception of Chiou and Tucker (2015), there is little evidence on copying and intellectual property regarding online news media. We contribute to this literature by providing new

[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.

[10] Recent work has also investigated the effect of digitization projects like Google Books (Reimers, 2015; Nagaraj, 2016). 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).

empirical evidence on the extent of copying online and estimating the returns to originality. Our paper is a unique attempt at understanding 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, and quantifies verbatim copying without acknowledgement. In Section 4, 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 5 performs a number of robustness checks. Finally, Section 6 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 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). The list of the media outlets included in our dataset is provided in the online Appendix Section B.1.

These media outlets are by far the main French news media both during our period of interest (2013) and still today. In October 2017, the first ten general information websites – that are all part of our dataset[12] – have received around $630,000$ visits, i.e. nearly four times more visits than the following ten websites that are also all part of our sample, except for one.[13] Only missing here are those local daily newspapers that had no websites at the time, and some very small digital news media that could not be considered as important information providers in 2013.[14]

[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.

[12] In decreasing order of importance: *Le Figaro*, *Le Monde*, BFM TV, *20 Minutes*, France Télévision, *Le Parisien – Aujourd'hui en France*, *Ouest France*, the Huffington Post, *Le Nouvel Observateur*, and *L'Express*.

[13] In descending order: LCI, *Le Point*, *Les Echos*, Europe 1, *Le Dauphiné Libéré*, *Libération*, *Sud Ouest*, *La Dépêche du Midi*, *La Voix du Nord*, and RFI. The local daily newspaper *La Voix du Nord* – and its $13,000$ visits per month in 2017, i.e. one tenth of the total number of visits received by *Le Figaro* – is the only top-20 general information news website missing from our sample.

[14] 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.

We choose this "transmedia" approach because, on the Internet, there is a tendency for different media to converge (see e.g. Peitz and Reisinger, 2016). Users interested in news tend to balance and compare multiple sources, regardless of the offline format of the media. One cannot infer the offline format of a media by visiting a website, as illustrated in the online Appendix Figure F.2. On the web, media all offer texts, videos and photos. We include the AFP even though it does not deliver news straight to individual consumers[15] because it is a key provider of original information in the online world. 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.[16]

Using their RSS feeds, we track every piece of content news media produced online in 2013. This content data is from the OTMedia research projet conducted by the INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French radio and television audiovisual archives). 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 dispatches directly from the agency. Merging these datasets, we obtain the universe of all the articles published online by French news media in 2013. These articles contain text and often photos, as well as videos. Our focus here is on text.[17]

Our dataset contains $2,493,360$ documents for the year 2013; around $6,800$ 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.[18] 72.6% of the documents are from the websites of the print media; 4.6% from radio; 6.5% from television; 13.1% from the AFP and the remaining documents from the pure online media (online Appendix Figure F.3a). On average, these documents are $2,062$ characters long.[19] Table E.1 in the online Appendix provides summary statistics for the entire sample, as well as by media format (print media, television, radio, pure online media and news agency).

[15] News agencies are based on a Business-to-Business model (they sell news to other media outlets), not on a Business-to-Consumer model. We provide more details on the specifics of the AFP in our companion paper, Cagé et al. (2017).

[16] 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 (2015); Calzada and Gil (2016).

[17] 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.

[18] 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.

[19] Online Appendix Figure F.4 plots the distribution of the length of the articles. For the reader to have in mind an order of magnitude, opinion pieces by Paul Krugman in the *New York Times* are around 4,000 characters long.

[Figure 1 about here.]

In the rest of this section, we review the algorithms we develop to study the production and propagation of information online. In Section D, 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 refered 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.[20] A semantic vector represents the relative importance of each word of the document compared to the full dataset. A standard scheme is TF-IDF.[21] As in most of natural language processing methods, we first pre-process our documents by removing very common words (called stop words) and applying a stemming algorithm so as to keep only the stem of the words. We also apply a multiplicative factor of five to the words of the title as they are supposed to describe well the event, resulting in an overweigth in the global vector describing the document.

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. As these vectors lie in a very high dimensional space, it is well known that the angle between the vector is a good measure for assessing similarity. We thus use the cosine similarity measure (Salton et al., 1975).

[20] 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 until now in the economic literature, e.g. Franceschelli (2011) considering only proper nouns.

[21] 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 more formally the TF-IDF weight in the online Appendix B.2.

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.[22]

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 5, we relax this condition and investigate the extent to which it affects our main results.

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.[23]

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 the one of humans.[24] 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).[25]

Note also that we find that the main parameter of our implementation, the distance threshold on semantic similarity, is the same for this English test corpus and our corpus of French news articles. This is very reassuring as to the quality of our algorithm. In particular, it ensures that the news events that are detected by our algorithm are as close as possible to what a human would be able to do.

[22]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.

[23]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 very 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 whose aim is different. Hence, the GDELT's dataset cannot be used to assess our results or the quality of our event detection algorithm.

[24]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 B.2.

[25]We provide details of the statistical measures of the performance of the algorithm in the online Appendix Section B.2.

As an additional robustness check, we compare our events to those obtained by the Europe Media Monitor (EMM) NewsExplorer.[26] 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 24,502 news events. Events can last more than one day; on average, they last 41 hours (we provide more details below on the length of the events; 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). The average number of documents per event is 34 and, on average, 15 media outlets refer to an event (online Appendix Table E.3). There are 178 events per day on average, with 67 new events beginning every day. These events are roughly equally distributed during the year. Figure 2 plots the total number of events per day, as well as the number of new events.

[Figure 2 about here.]

Out of the 2,493,360 documents in the dataset, 829,578 (33%) are classified in events (for a daily plot of this ratio, see Figure F.5 in the online Appendix). The remaining 67% of the documents are not classified in events. Note however that the classified documents represent, in terms of characters, 40.1% of the total content produced in 2013. Classified documents are indeed longer on average (online Appendix Table E.1). Note moreover that relaxing the "10 documents condition" to define an event increases the share of articles classified in events (we then classify 48% of the articles in events, and these articles represent more than 54% of the total content), but does not affect our main results (Section 5).

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 57% of the documents in our dataset, they account for 68% of the unclassified documents). By construction, these are documents that are not followed by other documents on the same issue, and many of them deal with extremely local issues. Local newspapers indeed cover very local "events" that are not covered either by other local outlets

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

(whose market differs) nor by national outlets.[27] While more than 50% of the documents published by national newspapers, radio, TV and the AFP are classified in events, less than 27% of those published by local newspapers are (see online Appendix Figure F.6 for a plot of these ratios).

A manual inspection of unclassified documents published in local newspapers shows that a large proportion of them deals with extremely local issues (and relatively "uninteresting" issues, at least from the viewpoint of our study), including the food menu served in local school canteens, the obituary section or the local weather. We feel that leaving them out of our event analysis is not a major problem.

Furthermore, these unclassified documents are much less popular in terms of social media audience. Online Appendix Table E.2 presents statistical differences between the articles classified and not classified in events: on average, documents that are not classified have approximately twice as less Facebook and Twitter shares than those that are classified (Facebook and Twitter shares data is described in details in Section 2.7 below). These differences are statistically significant at the one-percent level and not driven by outliers.

Other unclassified documents either correspond to one-off reports or to what Schudson (2015) calls "contextual reporting".[28] These articles tend to be relatively long. E.g. on November 20th, 2013, the national daily newspaper *Le Monde* published a 14,968 character-long article revisiting the Rey-Maupin affair (*"Retour sur l'affaire Rey-Maupin : les tueurs de la Nation"*). Florence Rey and Audry Maupin were involved in a shoot-out in Paris in October 1994 following a high speed car chase, causing the deaths of five people. Obviously, in 2013, i.e. 9 years later, this affair is no longer a news event; but lengthy articles revisiting affairs are a good illustration of what contextual reporting is. While on average unclassified documents are smaller than classified documents, the variance of the distribution of their size is higher because they tend to be either very short articles covering local events or much longer leading articles.

Finally, unclassified documents also include editorial and opinion pieces. This last category of unclassified document can actually matter a lot for public debate and information (and in some media outlets attract a lot of visitors). However, we feel that these documents would require a specific analysis. 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). However, a significant fraction of op-eds address

[27] Remember that the algorithm defining our events relies on the assumption that for a news story to exist, at least two different media outlets should cover it.

[28] *"The journalist's work is less to record the views of key actors in political events and more to analyze and explain them with a voice of his or her own."* 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.

issues that are not related to a specific event.

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 829,578 articles classified in our 24,502 events.[29] Table 1 provides summary statistics on these articles.[30] In Cagé et al. (2017), when estimating the production function of news, we take into account both the classified and unclassified documents.

[Table 1 about here.]

Length of the events On average, events last 41 hours. Figure 3 plots the distribution of this length. Around 10.5% of the events in our sample last less than six hours. These short events are mainly minor news items (e.g. on February 27th, 2013 the death of a woman in Paris poisoned by her leaking boiler, an event which was first covered by the AFP at 6:49pm and last mentioned by the free newspaper *Metro* at 8:02pm).

Some longer events, lasting over multiple days, are also minor news items, often first covered by a local newspaper before generating buzz at the national level. E.g. at 6:05am on January 1st, 2013, *Le Dauphiné Libéré* wrote the story of robbers who returned the loot they had stolen from a jewelry store along with a box of chocolates. Only two other media outlets (*Le Parisien* and *Direct Matin*) covered the story on January 1st, while it got increasing coverage on January 2nd after an article published on the website of the radio RTL. Finally, this event was last mentioned on January 3rd on France Télévision's website.

In some cases, the length of the media coverage is due to the nature of the event, e.g. when there is a revelation followed by a refutation. On January 30th, 2013, at 10:12am, *Le Monde* published an article claiming that Coca-Cola had removed its advertising from France Télévision after the airing of a "critical" documentary. This event was covered by 11 media outlets on January 30th and 31st before France Télévision's refutation denying any advertising boycott by Coca-Cola (the denial was reported by the free newspaper *20 Minutes* on February 1st).

Hence we cannot infer the nature of an event directly from its length. While the length of the coverage sometimes reflects the actual length of the event, it may also simply stem from editorial choices of media outlets. In Section 2.5, we resort to manual coding to improve our understanding of the specific nature of the different events in our sample.

[Figure 3 about here.]

[29] In Section 5, the number of classified articles increases to 1,192,598 when relaxing the "10 documents condition" to define an event.

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

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.[31] 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. For 95% of the events it covers, the AFP also provides information on sub-categories (e.g. "crime" is a sub-category of "Crime, law and justice", and "agriculture" a sub-category of "Economy, business and finance"). An event can be associated with more than one top-level media topic (8,428 events in our dataset are). E.g. when on January 1, 2013, the US Senate passed a compromise bill to eliminate the fiscal cliff, this event was classified by the AFP both in the "Economy, business and finance" category (with a sub-category "macroeconomy") and in the "Politics" category (with a sub-category "government").

Figure 4 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 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 4 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 out 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

[31] 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 B.5).

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 tracks efficiently identical portions of text between documents.[32] 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 code manually all the news stories in our sample to isolate the stories that are the results of a piece of original reporting by (at least) one outlet.[33] 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 C, 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 are stories for which the news originates from *"the revelation of new facts"* that someone wants to keep secret by the media outlet

[32]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 B.3. We focus on exact (verbatim) copying only.

[33]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 disagree to begin with has then been discussed at length by the authors.

(Hamilton, 2016)[34]. These stories involve substantial in-depth reporting, whereby media outlets are playing watchdog. (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. In other words, what distinguishes both types of stories is the amount of in-depth expository reporting involved. E.g. the NSA spying scandal revealed by *Le Monde* on October 21, 2013 and described in online Appendix Section D is an example of an investigative story. On the contrary, we classify as a (non-investigative) reporting story a series of articles on Chinese couples divorcing to avoid property tax or an event dealing with a Japanese man's vacations on Syrian front lines.

Finally we define short news items with multiple witnesses as our third category. These are news events for which there are multiple witnesses: e.g. a public protest; a murder in public space; a terrorist attack; a plane crash.[35] 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 around 84% of the events, as illustrated in Figure 5. Short news items with multiple witnesses account for 9.2% of the events, and exclusive news events for 6.6%. The average size of the documents included in the events varies with the nature of the news events, however. In the online Appendix Table E.4, we show that on average documents classified in exclusive news events tend to be longer (around $3,000$ characters) than documents classified in non-exclusive news events ($2,500$) or in news events with multiple witnesses ($2,400$), and that this difference is statistically significant.

Finally, only 1.4% 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. In a content analysis of over 33,000 stories aired between 1998 and 2001 on 154 local television stations, Rosenstiel et al. (2007) show that station-initiated investigations accounted for 0.62% of all political stories and 1.10% of nonpolitical stories.[36]

[Figure 5 about here.]

[34] *"Investigative reporting involves original work, about substantive issues, that someone wants to keep secret."*
[35] We include weather-related events in this category (but they only represent .6% of the events).
[36] 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, take over,...) This algorithm is described in details in the online Appendix Section B.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 3.3.

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[37] –, we have information on the number of unique visitors, the number of visits and the number of page views.[38] 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. Data sources are described in details in the online Appendix.

[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) which is the tool developed by Facebook to let external applications retrieve information about the number of shares of each article (using the URL of the articles). We provide more details on how we use the Facebook Graph API in the online Appendix Section B.7.

[37] The AFP being based on a Business-to-Business model, it does not deliver news to individual consumers on its website.

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

We obtain information on this variable for all the documents in our sample, to the exception of the articles published by the AFP that are not available online to the general audience. On average, articles are shared 65 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 968 and the maximum is $240,450$ while the 99th percentile is "only" $1,025$). 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 top-coded version of the Facebook shares variable where we code the values in the 99th percentile at the bottom value of the percentile.

If the data on the number of Facebook shares is useful for us to have 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 B.8.

For each article, we have eight different measures of the number of times it its "shared" on Twitter, as illustrated in the online Appendix Figure B.3: (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, as shown in the online Appendix Table B.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.

Note to finish that there is a positive relationship between the number of shares on Facebook and the number of shares on Twitter thus defined, as illustrated in the online Appendix Figure F.7.

The algorithms described in this Section are illustrated in the online Appendix Section D with the example of a news event.

3 Empirical analysis

3.1 The speed of news dissemination

In this Section, we study the speed of news dissemination online.[39] We construct the typical timeline of a news story. More precisely, we investigate how fast news is delivered to readers of different media outlets after being published first on the website of the news breaker.[40]

Studying the speed of news dissemination is of interest because the commercial value of a news 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 172 minutes for some information published 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 224 seconds in 25% of the cases and less than 5 seconds in 10% of the cases.

Table 3 reports the average reaction time depending on the offline format of the news breaker. If the news agency (AFP) is the first media outlet to publish some information (which is the case for half of the events), then the reaction time is shorter. When the AFP is the news breaker, we find that the second media outlet covers it after 117 minutes on average, but after only 10 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 AFP; 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.

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.

[Table 3 about here.]

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

[40] 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, and then we know the exact publication time of all the articles related to this event until the last media outlet reports on it.

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 non-exclusive news events and the longest for exclusive news events, and that the differences are statistically significant.[41] Finally, in the online Appendix Section H.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 we just discussed 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 out 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%.[42] In Figure 6a, 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, to the exception of the documents which are entirely original, the articles published within events consist mainly of verbatim copying: 55.2% 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 6b 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 (and reassuring as to our manual coding of the events): non-exclusive news events are 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 6 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 F.8a illustrates how

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

[42] 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.

the average originality rate varies depending on the document length: short articles tend to rely less on copy than longer ones. Regarding the offline format of the source, we also find that pure online media tend to be on average more original than other media outlets (online Appendix Figure F.8b). Note however that pure online media only account for 3% of the documents in our dataset.[43]

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. We find that, on average, documents include content from 3.9 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 3 documents in the event, 2 when we exclude the AFP.[44]

We find that out of the $829,578$ documents classified in events, $607,367$ (73.2%) present at least some external copy. On average, documents include content from 3.3 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 50.8% (72.6% conditional on copying).

Excluding the AFP When considering the returns to originality, one needs to distinguish between content copied from the AFP (the news agency) and content copied from other media outlets. All the media outlets that are clients of the AFP 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. We show that on average, documents include content from 1.9 documents published by competing media outlets other than the

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

[44] The AFP publication strategy is characteristic of the work of news agencies which consists in publishing first short dispatches and then by supplementing them with more details during the day.

AFP. If we exclude content copied from the AFP, we find that the average external copy rate is 16.3% (28% conditional on copying).

Share of the original story that is copied Finally, we compute the share of each document which is copied. On average, we find that each document is copied by 3.9 documents published later in the event, 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 at least reproduced by one external media outlet, we find that on average the share of a document that is copied is equal to 9.6%.

The majority of the documents are not copied, however. If we restrict our analysis to copied documents, we obtain that the share of a document that is copied by at least one external media outlet is 21.2% on average. If, as before, we exclude documents published by the AFP, we find that this share is equal to 10.4%.

This share varies strongly depending on the publication rank of the document. Online Appendix Figure F.9 plots the average share of a document that is copied by at least one external media outlet depending on the publication rank of the document. We find that for breaking news documents (documents that are first published within the event), this share is above 60%, 25% when we exclude documents published by the AFP. It then decreases to nearly 25% (12%) for the second document and converge rapidly to around 5%.

Finally, we investigate whether the share of the original story that is copied varies depending on the nature of the news events. We show that the share of the breaking news document that is copied is higher for exclusive news events, at 67.4%.

[Table 4 about here.]

3.3 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 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.

We perform this analysis at the event level. For each document presenting at least some external verbatim copying, we investigate whether it refers to the media outlet(s) it copies as the source of the information. Table 5 summarizes the main results. We find that documents mention (at least one of) the competing media outlet(s) they copy as the source of the information in 32.5% of the cases. Moreover, if we exclude verbatim copying from documents published by the AFP, the probability of crediting decreases to 3.5%.

When a media outlet reproduces content from multiple documents, it may choose only to refer to the competitor whose document it copies the most. We study the extent to which this is the case, and find that documents refer to media outlets whose document they copy the most in 26.3% of cases. However, if we exclude all cases where the most copied media is the AFP, we show that this probability drops to 3.8%.

In Figure 7, we plot the share of the documents crediting the copied media depending on the copy rate. We do so both including and excluding the AFP as a copied media. We find that not only do media outlets hardly name the media they copy, but that their propensity to do so scarcely increases with the extent of copying. Once the AFP is excluded as a copied media, the share of crediting documents is always below 5%. If we also consider documents copied from the AFP, we show that this share increases from 4% to 28% with the importance of copying. Why do media outlets tend to credit the AFP more than the other outlets? Most probably because they are not competing directly with the AFP.

[Figure 7 about here.]

Rather than referring to the outlets they copy as the source of the information, media outlets may choose simply to credit the breaking news outlet. We find that they do so in only 25.1% of the cases, 8.8% when the breaking news outlet is not the AFP. Finally, we show that the media refer more to the breaking news outlet when the news event is exclusive: 27.3% of the outlets refer to the breaking news outlet as the source of the information in this case.

[Table 5 about here.]

4 Online audience and the returns to originality

Do original news producers nonetheless benefit from their investment in newsgathering? Reputation can provide one way to understand why media invest in information production (Gentzkow and Shapiro, 2008). In this section, we investigate the relationship between the production of original information and the audience of the websites. We provide tentative estimates of the returns to originality.

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

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 that we are well-aware off. We first attempt to overcome it by using alternative article-level information

that we collect from two different social media, namely Facebook and Twitter. We will then use an additional dataset to relate article-level Facebook and Twitter shares and article-level numbers of views.

4.1.1 Number of Facebook shares

First, we use article-level data to investigate how the number of times an article is shared on Facebook varies with its originality and reactivity.[45] 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):

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

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.[46]

$\mathbf{Z}'_{\mathbf{aedn}}$ 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 then is 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 thousand); (vi) the original content (also in number of thousand characters); and (vii) the non-original content. 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).

Table 6 presents the results. Regarding the originality, we find that a thousand increase in

[45] We have also investigated how online audience varies with the number of breaking news using daily audience data, and estimating a model with outlet and day fixed effects (the observations were at the media outlet-day level). We found that a one-percent increase in the production of original content leads to a .018% increase in the number of unique visitors, and that breaking at least one news story during the day increases the number of unique visitors by 1.4%. Results are available upon demand.

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

the number of original characters leads to a 22% increase in the number of Facebook shares. If we rather 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 39% increase in the number of Facebook shares. If we now turn to the reactivity, we find that both the publication rank and the reaction time matter. The effect is economically small, however: an increase by one in the publication rank leads to a 0.05% decrease in the number of shares on Facebook. Each event is covered on average by 34 articles: moving from being the breaking news article to being the last article covering an event leads to a 1.7% percentage point decrease in the number of times the article is shared on Facebook. Finally, 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 9.4%.

[Table 6 about here.]

In the online Appendix Table E.5, we present the results of the estimations when we use the number of times an article is viewed (using the Facebook approach detailed below) rather than the number of Facebook shares as a dependent variable. The signs of the coefficients are consistent with those we obtain in Table 6. In terms of magnitude, a thousand increase in the number of original characters leads to a 23% increase in the number of times this article is viewed.

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%;...; between 95% and 100%). We then estimate equation (1) using as independent variables these dummies rather than the continuous originality rate measure. Figure 8 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 8 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 E.6 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 0.62 to 0.66% 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. However, when we use the percentile instead of the absolute publication rank, the negative effect of the reaction time disappears. Hence, all the effect of the reaction time seems to be captured by the publication rank.

Finally, as an alternative strategy to deal with the skewness of the Facebook shares variable distribution, we top code the extreme values of the variable. More specifically, we attribute to all the values above the 99th percentile a value equal to the bottom threshold of the 99th percentile. We then perform a linear estimation. Online Appendix Table E.7 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 12 additional shares of the article on Facebook (the effect is statistically significant at the one-percent level).

4.1.2 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. Table 7 presents 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.

[Table 7 about here.]

The results we obtain are consistent with the findings of Table 6. On the one hand, social media audience increases with the number of original characters: a thousand increase in the number of original characters leads to a 11% increase in the number of Tweets. If we rather consider the originality rate, a 50 percentage point increase in the originality rate of an article leads to a 16% 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.[47]

No more than the number of shares of Facebook, the number of Tweets is a perfect measure of the audience of an article. However, the consistent findings we obtain by using

[47] Furthermore, when we use the number of Tweets as a dependent variable, the statistical significance of the reaction time variable remains when we control for the percentile rather than for the absolute publication rank of the article (online Appendix Table E.8).

both measures seem to reveal the fact that consumers favor original content and reactivity. In Section 4.2 below, we combine social media and audience statistics to build an audience-weighted measure of the importance of original content.

4.1.3 Discussion

How can one rationalize the positive relationship between originality and news consumption as proxied by the number of shares on social media? Our favorite explanation is that consumers may favor originality. Moreover, as we will see below, this partly happens through a long-term reputation effect: overall in 2013, media outlets with a larger fraction of original content tend to receive more audience. But other channels may be at play. First, it is important to keep in mind that, as highlighted for example by Boczkowski and Mitchelstein (2013), consumption choices are *"often made at the story level"* (p.9). Hence, independently of the reputation of a given media, consumers willing to learn about a news event may decide to read the most original piece, for example simply because this is the first article published within an event (and so the first they are given to see).

However, the originality of an article matters even when we control for its publication rank, meaning that, for a given publication rank (or reaction time), consumers do value originality *per se*. 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.[48]

4.2 Social media statistics and number of views

4.2.1 Evidence from *Le Monde*'s data

In this section, we document the relationship between the number of times an article is viewed and the number of times it is shared on Facebook. This is of particular importance for us given that our approach in the next section uses this relationship to compute number of views per article statistics.[49] 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 and August 2017, as well as the URL of the articles. We use the URL to compute as before the

[48] 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, one of the questions webmasters should ask themselves is the following: *"does the article provide original content or information, original reporting, original research, or original analysis?"*.

[49] 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 empirically this relationship.

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 (but the distribution of the number of shares is right-skewed; the median is 161).

Figure 9 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 9a). 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. 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 9b shows the results. We then 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 9 about here.]

4.2.2 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 it both by using the raw number of Facebook shares (of Tweets) and the top-coded 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, 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) Finally, we use the estimated parameters from *Le Monde*'s article-level data (described in Section 4.2.1 just above) 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 4.2.1). 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 Share Visits$_{adn}$ is the share of the total views received by media n on date d represented by article a, and 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.

4.3 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 10 presents the results. First, for the sake of comparison, we compute the share of original content in the dataset. This share is equal to 33.5%.[50] 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 46% 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 62% 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 56.6% when we attribute the number of views assuming a linear relationship with the number of Tweets and to 56.4% 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 10 about here.]

4.4 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

[50] We only consider here the articles for which we have audience data, and in particular we drop the AFP. 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).

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 case Internet audience was distributed randomly on the different websites and 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, e.g. via online advertising revenues, on which we unfortunately do not have direct data[51]), 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 audience shares (using the naive approach), we find that original content represents 46% of online news consumption. This reflects the fact that media outlets with a larger fraction of original content tend to receive more audience.

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 53.8% and 62% 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 reputation effects, the audience share of original content jumps from 33.5% to between 53.8 and 62%, which also means that about 30 (20 points out of 67) to 42% (28 points out of 67) of the copyright violation problem is effectively addressed by reputation mechanisms and viewer behavior.[52]

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 AFP (given the business model of the AFP is a different one – we also drop all the AFP dispatches). More precisely, we define the *total content* of an article as its content minus the content reproduced from the AFP, and the *original content* of an article as its content minus the content reproduced from

[51] Media outlets' advertising revenues can be split in two parts: their offline revenues and their online revenues. While we were able to collect information on aggregate advertising revenues received by each media outlets in 2013, to the extent of our knowledge, there exist no data allowing us to isolate the share of advertising from online audience.

[52] 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 F.10). Considering internal copy as original content, the share of original content in our dataset is 37.4% (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 57.1 to 65.5% depending on the specification (50.5% when we use the naive approach.) In other words, the share of the copyright violation problem which is solved by reputation mechanisms and viewer behavior is still equal to about 31 to 45% (20 to 28 points out of 62) depending on the specifications.

the AFP and the content reproduced from other media outlets (excluding itself). Doing so, we find that on average documents are 1,355 characters long (compared to 2,062 characters when content copied from the AFP is included). We also obtain that 69% of the online content is original; higher originality is not surprising given that we have showed that media outlets mainly rely on content copied from the AFP. What about the relative consumption of original content? We obtain that the audience-weighted share of original content is equal to 78.8% when we use the naive approach, and to between 81 and 83.7% when we allocate the number of views as a function of the number of shares on social media (online Appendix F.11). Hence, despite the lower magnitude of the copyright violation problem – two third of the returns captured by the original information producer, one third by the copiers once copy from the AFP is excluded from the analysis –, the reputation mechanism still solves about 38 to 46% of this problem (11.7 to 14.4 points out of 31).

Of course, this conclusion requires that the media outlets realize this and allocate their effort and journalist time accordingly. In case media outlets only took into account media-level audience, and did not realize that more original articles generate more audience, then they would misallocate some of their resources (i.e. more time in copying-and-pasting and less time in investigation and original content production).

We should also 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 next because the copied segments of a given article might be the most "valuable" and original segments (something we cannot fully measure). On the other hand, 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 Facebook 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 copyright violation on the incentives for news production.

5 Robustness checks

In this Section, we perform a number of robustness checks. In particular, 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.

Not surprisingly, the total number of media events identified by the algorithm strongly increases when we relax this condition. We obtain a total number of 112,891 news events. Out of the 2,493,360 articles in our dataset, 1,192,598, i.e. 48%, are classified in the events thus defined. Classified documents represent 54.2% 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 conditions is imposed – 2,335 characters long compared to 2,484 – they are nonetheless longer than the documents that are not classified (1,812 characters on average) (online Appendix Table G.1). Unclassified documents mostly come from local daily newspapers which account for nearly 77% of the unclassified documents (while remember that those newspapers represent 57% of the documents in our dataset) (online Appendix Figure G.1).

When we relax the 10 documents condition, the events are much shorter (they last less than 21 hours on average, compared to 41.4 when the condition is imposed) and comprise on average a lower number of documents (online Appendix Table G.2). As before, they are mainly about "politics" and "economy, business and finance" (online Appendix Figure G.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".[53]

Regarding the 1,192,598 documents classified in the events, their originality rate is equal on average to 43.2% (online Appendix Table G.3). This rate is nearly 6 percentage points higher than when we impose that events should contain at least 10 documents. Figure 11 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 (nearly 15% of the documents) and another peak for the 100%-original articles. The later, with 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 11 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 12 shows the results. As

[53] 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 G.2 are computed for only 63% of the events. For the remaining 37%, we have no information on their topic.

before, it appears clearly that the audience-weighted share of original content (which varies between 58.3 and 67.1% depending on the specification) is much higher than the production of original content. Doing the same simple calculation as above, we obtain that 32 to 46% of the copyright violation problem is solved thanks to reputation effects.

[Figure 12 about here.]

Finally, we re-estimate equation (1) using the dataset where we relax the "10 documents condition" to define a media event. Table 8 presents the main results, with the effects on the number of Facebook shares reported in Columns 1 and 2, and the effect on the number of Tweets reported in Columns 3 and 4. 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 Tables 6 and 7, the coefficients we obtain for each of the explanatory variables of interest are of the same order of magnitude. For example, we find that a thousand increase in the number of original characters leads to a 21% increase in the number of shares on Facebook and a 10.6% increase in the number of shares on Twitter.

[Table 8 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.

6 Conclusion

This paper documents the extent of copying online and estimates the returns to originality in online news production. It uses a unique dataset covering the online production of information of the universe of French news media during the year 2013 and develops a number of algorithms which could be of use in the future 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, 2016). In case online audience was distributed randomly, our results would imply that the original news producers

capture only one third of the economic returns to the original news content they provide. However, we show that reputation mechanisms and the behavior of Internet viewers allow to mitigate a significant part of this plagiarism problem. We indeed find that original content represents up to 58% of online news consumption, and that reputation mechanisms actually appear to solve about 40% of these newsgathering incentives issues.

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.[54] Furthermore, our results suggest that in order to effectively address these issues, it is important to study 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 be of use in the future to improve our understanding of "where people get their news", combining consumption and production data. Prat (2017) and Kennedy and Prat (2017) 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 these news. More research is still needed, but we hope this paper will inform the debate on concentration in media power.

[54]Moreover, copyright is a second-best solution to intellectual property provision.

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.

Angelucci, Charles and Julia Cagé, "Newspapers in Times of Low Advertising Revenues," CEPR Discussion Paper 11414, C.E.P.R. Discussion Papers 2016.

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.

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

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

_ and E 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.

_ , **Nicolas Hervé, and Marie-Luce Viaud**, "Estimating the Production and Demand for Online News: Micro-Level Evidence from the Universe of French News Media," Working Paper 2017.

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," Working Paper 21404, National Bureau of Economic Research 2015.

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.

Gentzkow, Matthew, "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.

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.

Ginsburg, Jane C., "Overview of Copyright Law," in Rochelle Dreyfuss and Justine Pila, eds., *Oxford Handbook of Intellectual Property*, 2016.

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, J T, *Democracy's Detectives: The Economics of Investigative Journalism*, Harvard University Press, 2016.

Hamilton, James, *All the News That's Fit to Shell: How the Market Transforms Information Into News*, Princeton University Press, 2004.

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.

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

Kennedy, Patrick and Andrea Prat, "Where Do People Get Their News?," Working Paper 2017.

Li, Xing, Megan MacGarvie, and Petra Moser, "Dead Poet's Property - How Does Copyright Influence Price?," NBER Working Papers 21522, National Bureau of Economic Research, Inc 2015.

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.

Nagaraj, Abhishek, "Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project," Working Paper 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.

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.

Prat, Andrea, "Media Power," *Journal of Political Economy*, 2017.

Reimers, Imke, "Copyright and Generic Entry in Book Publishing," Working Paper 2015.

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, T, M Just, T Belt, A Pertilla, W Dean, and D 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, G, A Wong, and C S 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.

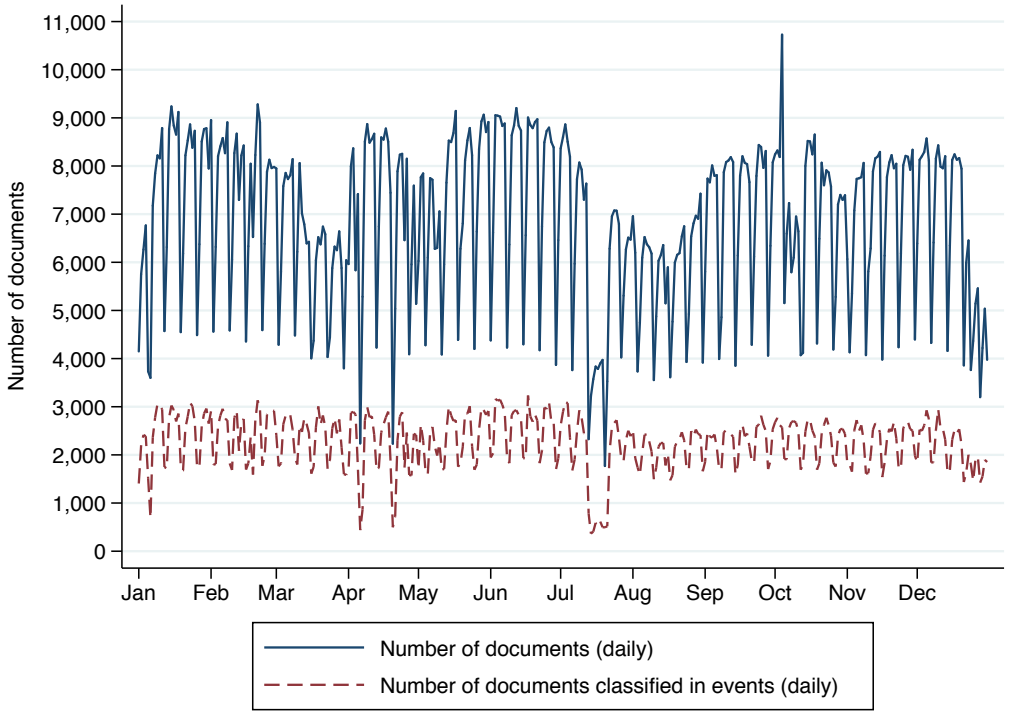
Sen, Ananya and Pinar Yildirim, "Clicks and Editorial Decisions: How Does Popularity Shape Online News Coverage?," Working Paper 2015.

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.

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

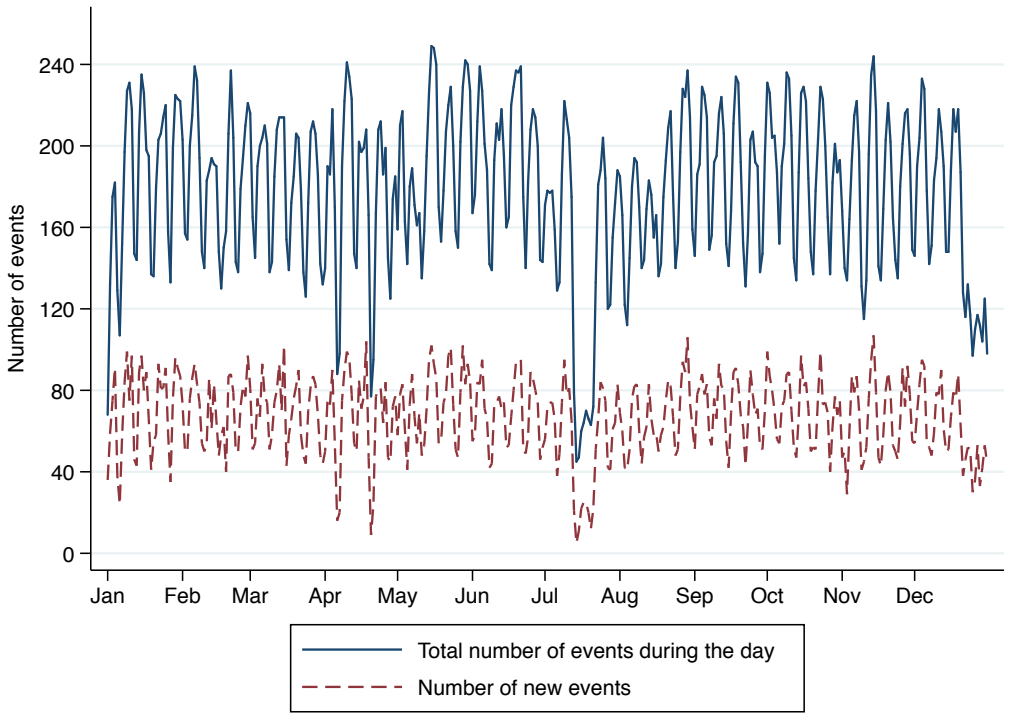
Waldfogel, Joel, "Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster," *The Journal of Law & Economics*, 2012, *55* (4), 715–740.

_ , "Digitization and the Quality of New Media Products: The Case of Music," in "Economic Analysis of the Digital Economy," University of Chicago Press, 2015, pp. 407–442.



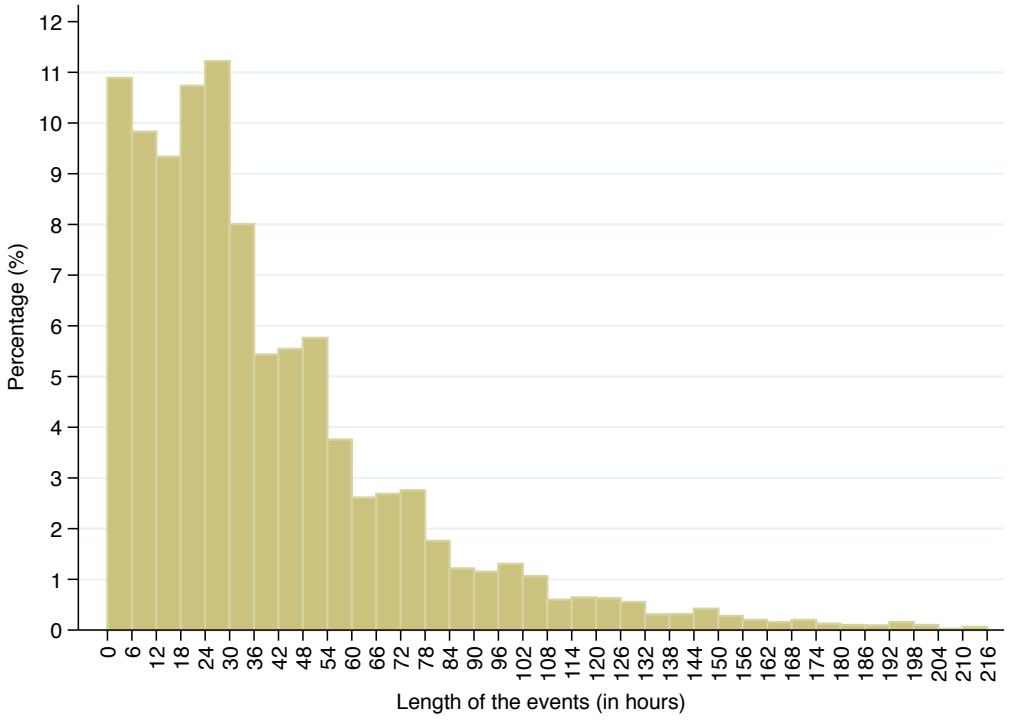
Notes: The figure plots the total daily number of documents included in our dataset. The solid blue line shows the total number of documents. The red dashed line shows the number of documents that are classified in news events. News events are defined in details in the text.

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



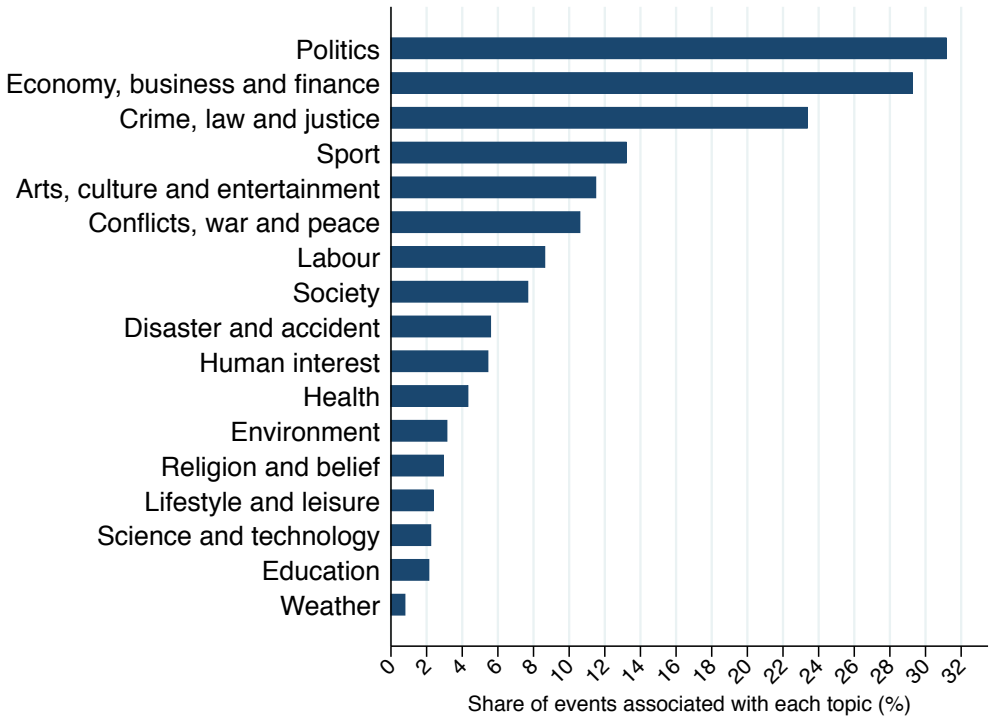
Notes: The figure plots the total number of news events taking place during the day (solid blue line) and the number of news events beginning on a given day (dashed red line). News events are defined in details in the text.

Figure 2: Daily distribution of the number of events



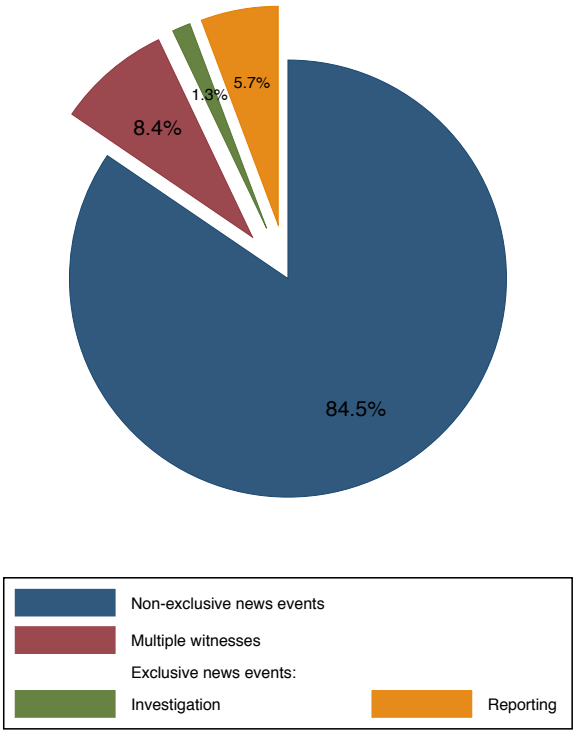
Notes: The figure plots the distribution of the length of the events in hours (with bins equal to six hours).

Figure 3: Distribution of the length of the events (in hours)



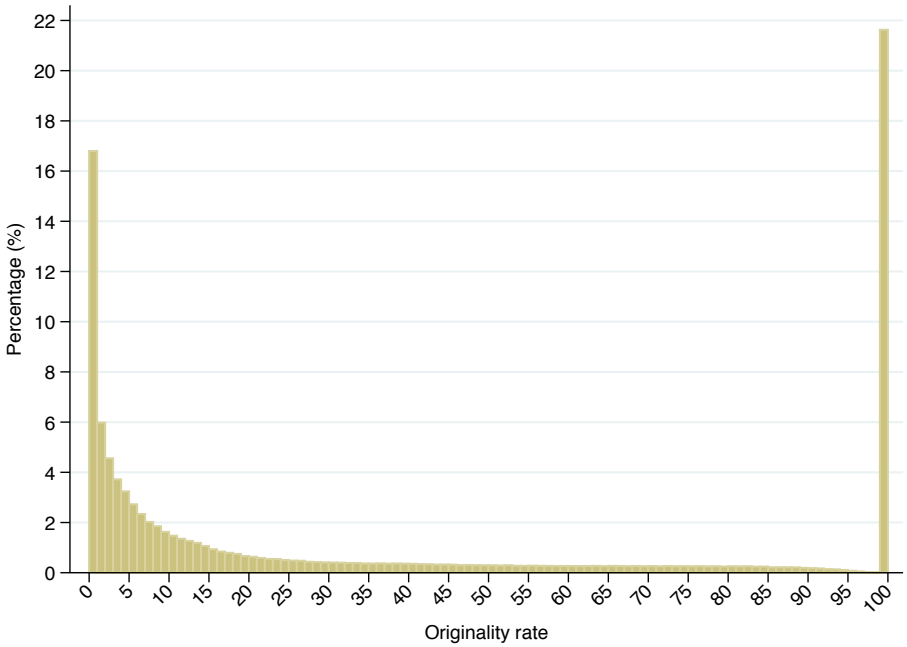
Notes: The figure shows the share of events associated with each media topic. The topics correspond to the IPTC media topics described in the text and defined in the online Appendix. Because some events are associated with more than one topic, the sum of the shares is higher than 100%.

Figure 4: Share of events associated with each media topic

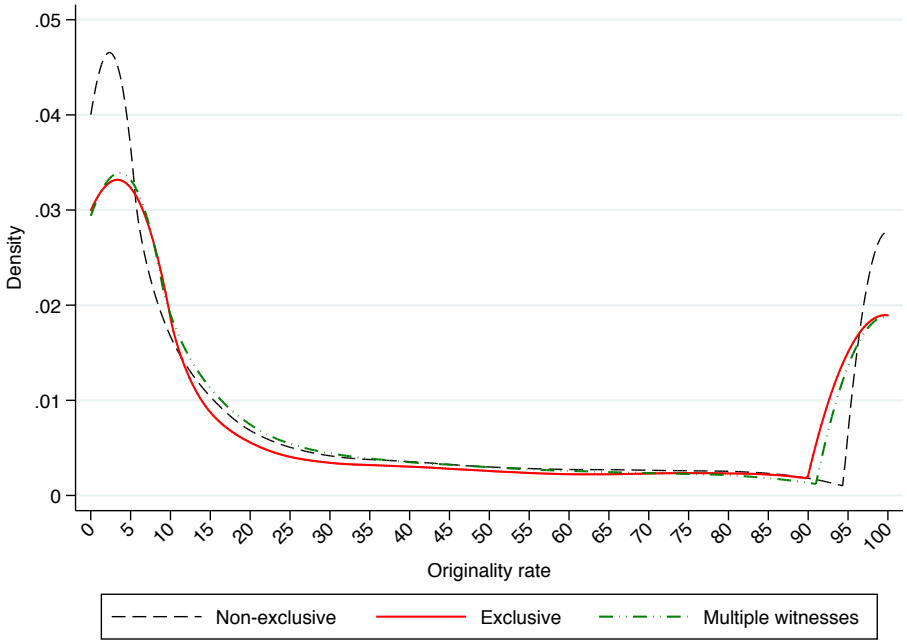


Notes: The figure plots the share of the news events classified in each category depending on their nature: exclusive news events, non-exclusive news events, and short news items with multiple witnesses. Exclusive news events can be either investigative stories or (non-investigative) reporting stories.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

Figure 5: Share of the events depending on the information issuer



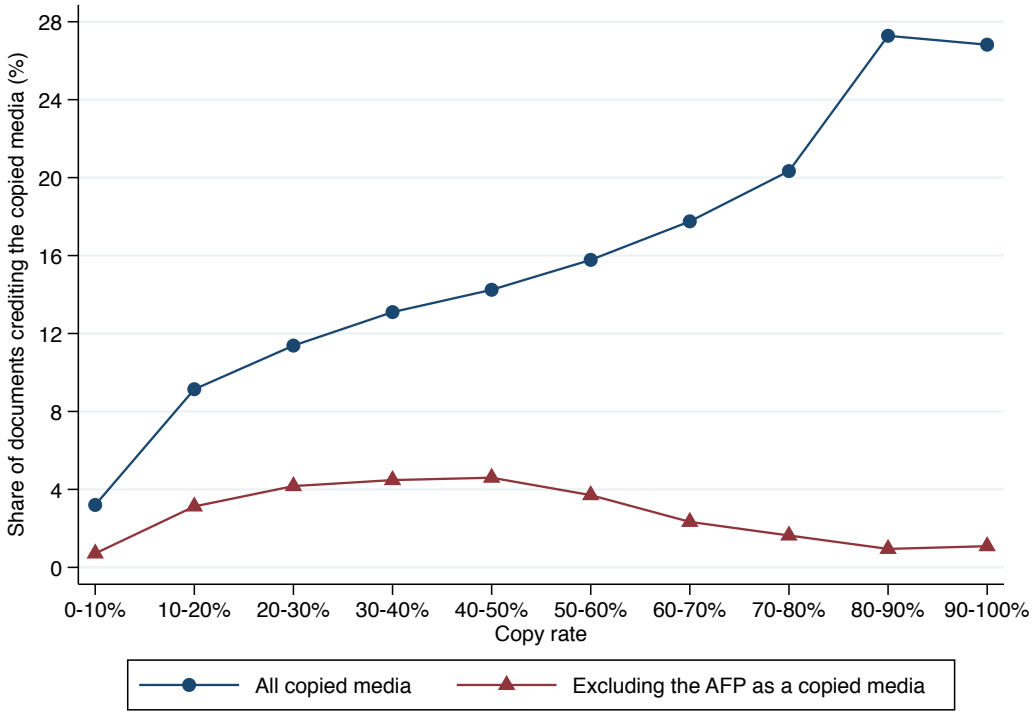
(a) Distribution (all documents classified in events)



(b) Kernel density estimate

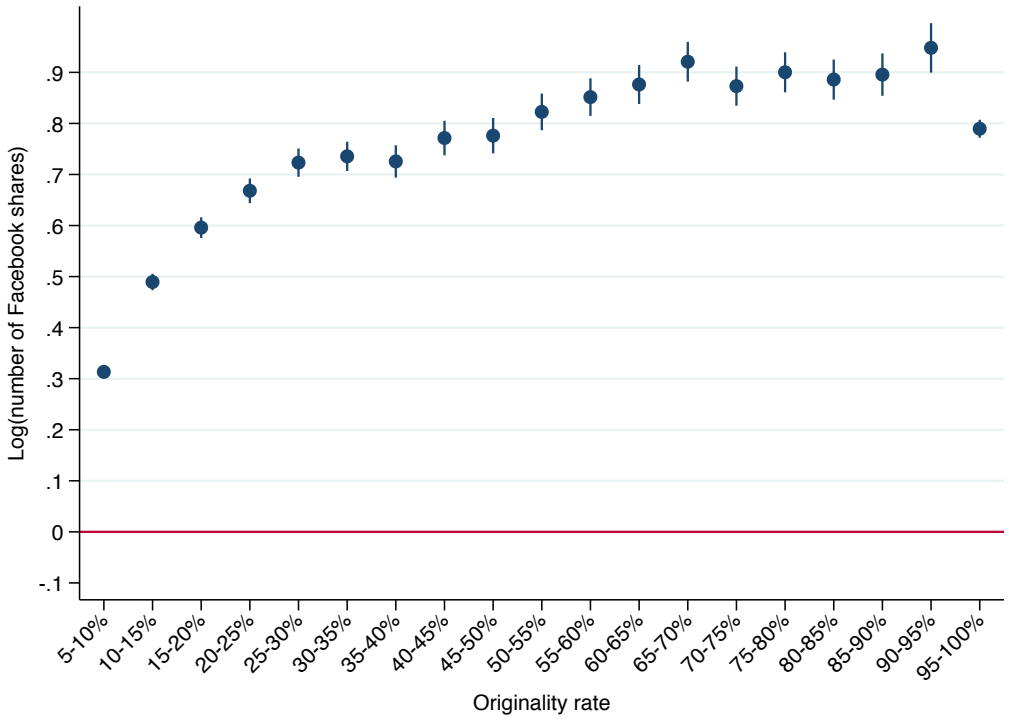
Notes: The upper figure plots the distribution of the originality rate (with bins equal to one percent). The bottom figure plots the Kernel density estimates depending on the nature of the news event. News events are either exclusive news events, non-exclusive news events, or short news items with multiple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

Figure 6: Originality rate



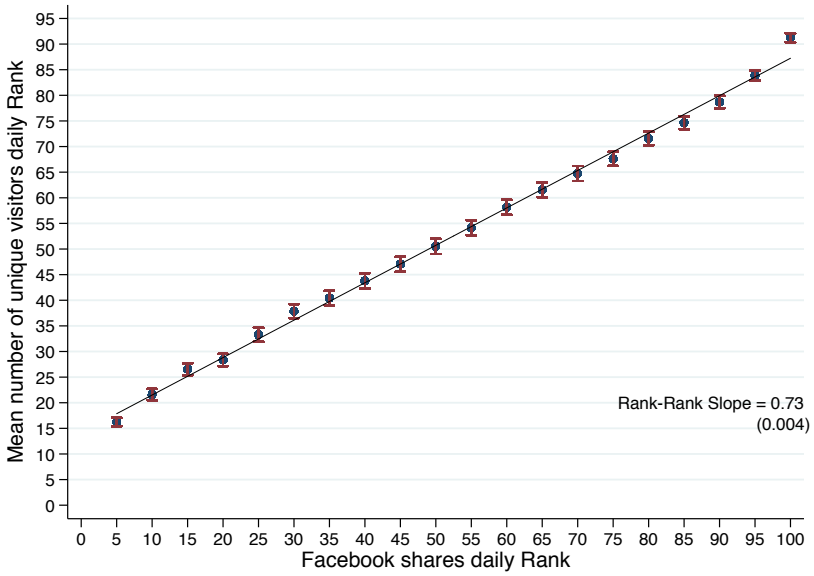
Notes: The figure plots the share of the documents crediting the copied media as a function of the copy rate. We define 10 different intervals for the copy rate: below 10%, between 10 and 20%,..., between 90 and 100%. The share is computed for each of these intervals.

Figure 7: Share of documents crediting the copied media as a function of the copy rate

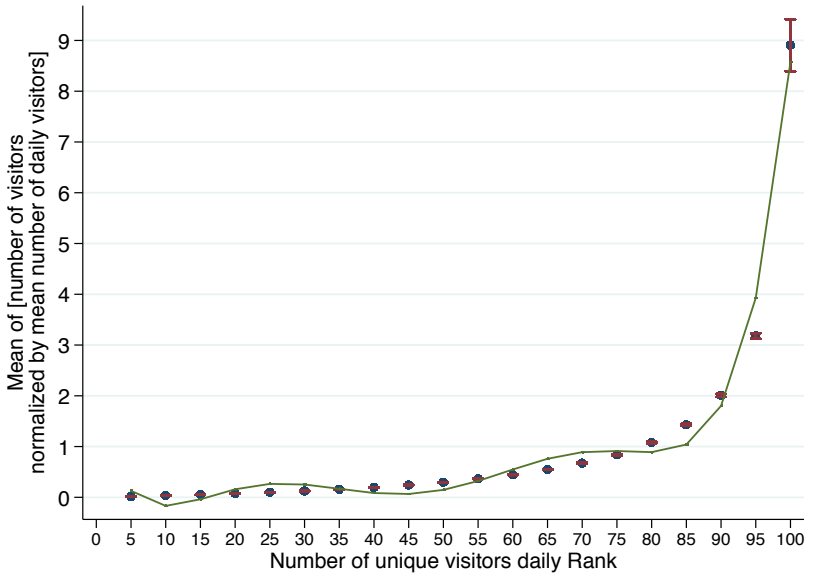


Notes: Figure shows coefficients from a regression of log of the number of times an article is shared on Facebook on twenty categorical variables depending on the originality rate of the articles (articles with an originality rate lower than 5% are the omitted category). Models include media, day and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

Figure 8: Facebook shares and originality rate



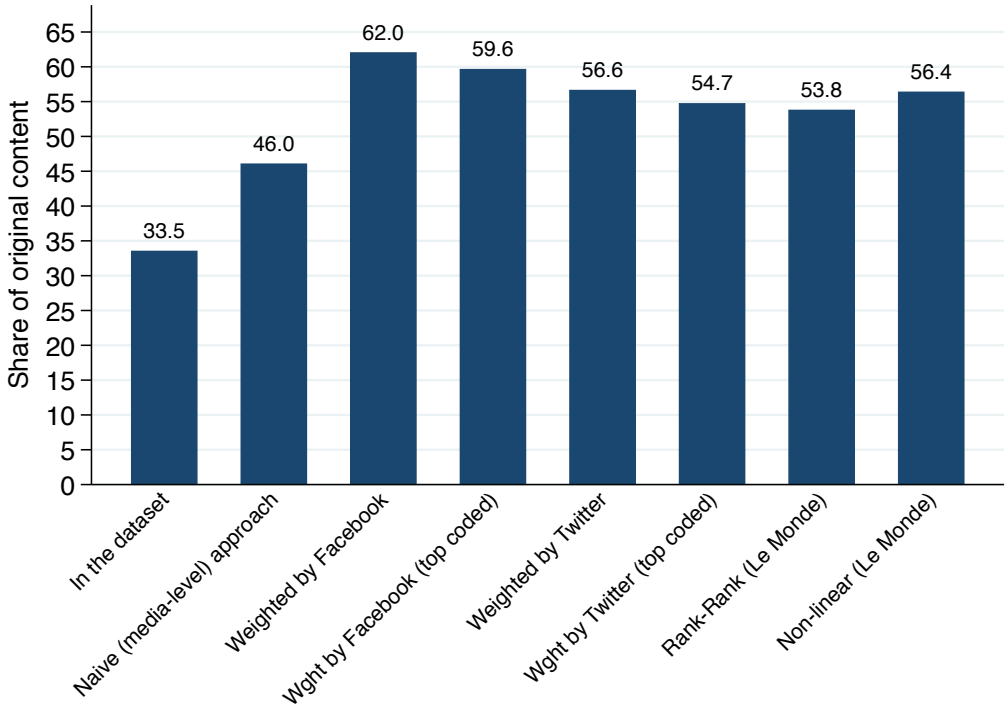
(a) **Association between number of Unique visitors' and Facebook shares' Percentile Ranks**



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

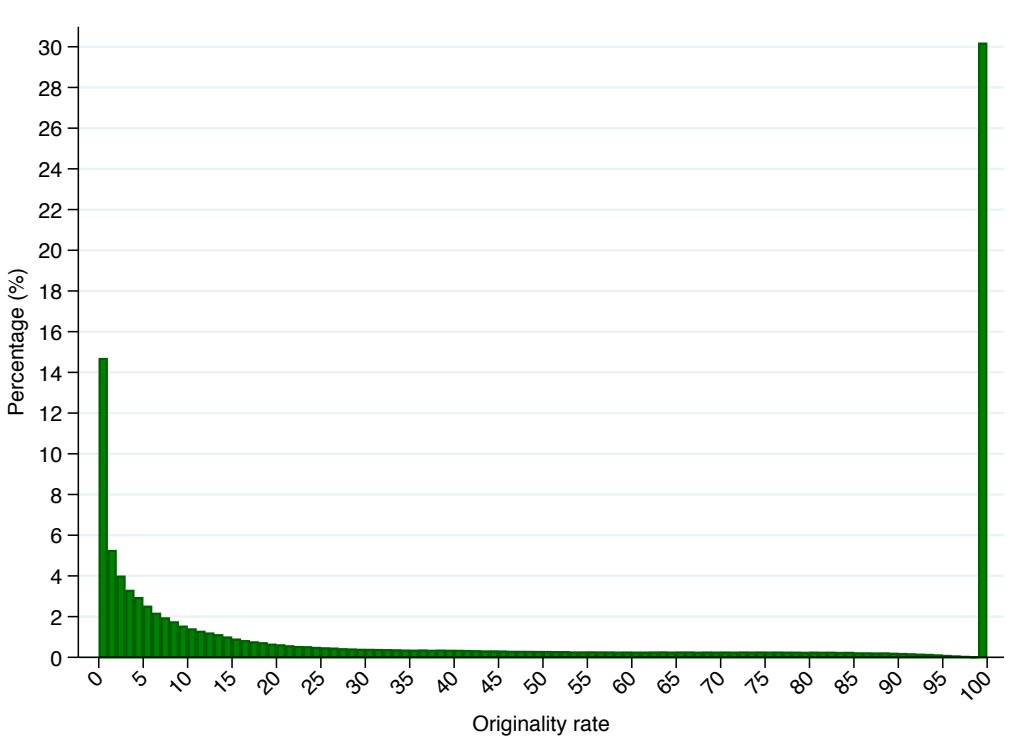
Notes: The Figure investigates the relationship between the number of unique visitors and the number of Facebook shares, using article-level information from the national daily newspaper *Le Monde*. The data includes all the articles published by *Le Monde* between April and August 2017 (17, 314 articles). In the upper Figure 9a, 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 9b, 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 9: 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



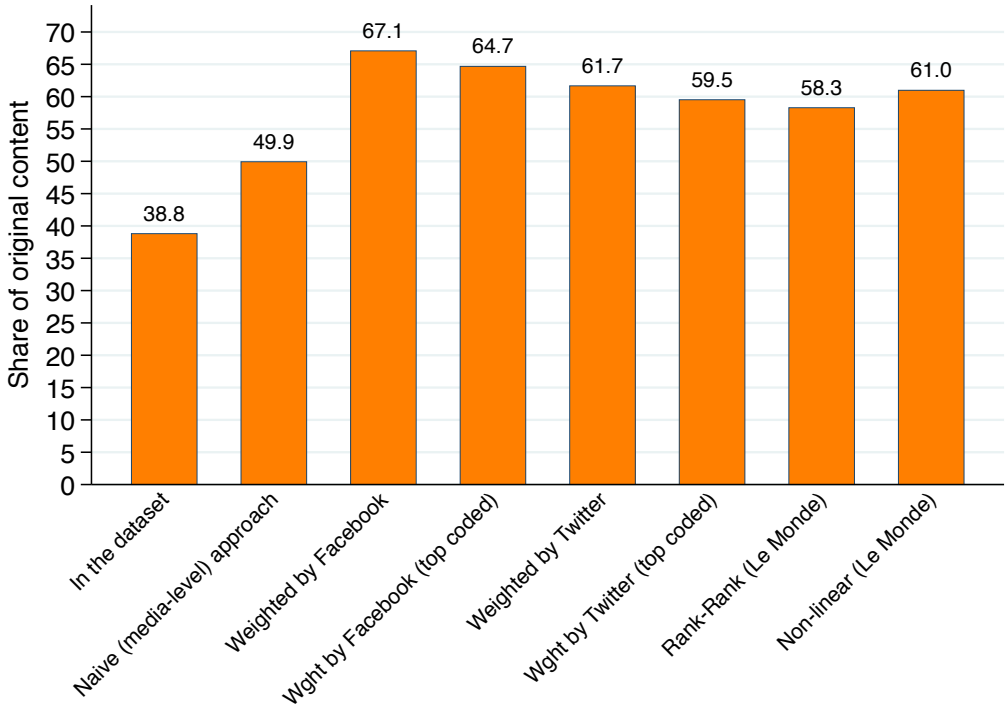
Notes: The Figure reports the audience-weighted share of original content we obtain using our different approaches to compute article-level number of views. The first bar ("In the dataset") simply reports the share of original content in the dataset (with no weight). The second bar ("Naive (media-level) approach") reports the share of original content we obtain when we attribute to all the articles published by a media outlet on a given date the same number of views. To compute the third 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 fourth bar ("Wght by Facebook (top coded)") relies on the same methodology but the Facebook shares variable is top coded. The fifth ("Weighted by Twitter") and sixth ("Wght by Twitter (top coded)") bars are computed similarly than the third and fourth 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 seventh 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 eighth 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 10: 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 11: 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 10, 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 to all the articles published by a media outlet on a given date the same number of views. To compute the third 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 fourth bar ("Wght by Facebook (top coded)") relies on the same methodology but the Facebook shares variable is top coded. The fifth ("Weighted by Twitter") and sixth ("Wght by Twitter (top coded)") bars are computed similarly than the third and fourth 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 seventh 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 eighth 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 12: 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)	2,484	2,213	1,587	100	98,340
Original content (number of characters)	811	243	1,306	1	53,424
Non-original content (number of characters)	1,673	1,341	1,548	0	63,738
Originality (%)	36.3	13.9	40.0	0	100
Reactivity in hours	42.5	19.4	68.6	0	6,258
Audience					
Number of shares on Facebook	65	0	968	0	240,450
Number of shares on Facebook (top coded)	37	0	137	0	1,025
Number of shares on Twitter	9	0	42	0	11,908
Number of shares on Twitter (top coded)	8	0	19	0	126
Obs	829,578				

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 (top coded)" variable is the number of shares on Facebook variable where we code the values in the 99th percentile at the bottom value of the percentile. Similarly, the "Number of shares on Twitter (top coded)" variable is the number of shares on Twitter variable where we code the values in the 99th percentile at the bottom value of the 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	247,349	107,856	382,574	3,689	2,031,580
Number of visits	339,097	156,735	542,470	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 direct tweets	3,627	577	8,792	0	58,507
Content (nb of characters) **(annual)**					
Total content not classified	32,202,326	14,746,971	114,804,997	425,989	1,064,234,304
Total content classified	19,762,077	11,709,107	23,725,294	1,114	101,030,320
Total original content	6,574,005	3,852,751	7,527,736	1,114	31,837,022
Total non-original content	13,188,071	6,715,187	19,675,283	0	76,895,000
Number of breaking news	125	54	185	0	1,052
Observations	85				

Notes: The table gives summary statistics. Year is 2013. Variables are values for media outlets (excepting the AFP). 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)	172	362	22	0	2,674	24,502
If news breaker is						
Print media	238	393	69	0	2,674	7,687
Television	226	397	51	0	2,222	1,210
Radio	228	391	66	0	2,402	1,040
Pure online media	406	499	183	0	2,164	518
News agency	117	317	10	0	2,624	13,887

(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
	168.9	197.2	163.8	28.3***	-5.0	33.3**
	(361.2)	(381.3)	(339.3)	(2.7)	(-0.5)	(2.5)
Obs	16,415	1,290	1,538	17,705	17,953	2,828

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			All external copy		Excluding content copied from AFP	
	(1)	(2)	(3)	(4) Including AFP	(5) Excluding AFP	(6) Including AFP	(7) Excluding AFP
	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd	mean/sd
Originality rate	36.3						
	(40.0)						
Originality rate wghtd by nb of views (Facebook)		54.6					
		(39.0)					
Copying media							
Nb docs copied			3.9	3.3	3.8	1.9	2.3
			(4.9)	(4.6)	(4.8)	(3.1)	(3.3)
External copy rate				50.8	59.9	16.3	17.0
				(41.7)	(39.9)	(26.2)	(26.3)
External copy rate conditional on copying				72.6	78.6	28.0	27.6
				(30.1)	(24.8)	(29.1)	(28.8)
Copied media							
Nb copying docs			3.9	3.3	3.3		
			(9.2)	(8.3)	(8.3)		
% of the doc that is copied				9.6	4.3		
				(23.5)	(13.8)		
% of the doc that is copied conditional on being copied				21.2	10.4		
				(31.3)	(19.8)		
If copied media bk news							
% of the doc that is copied				60.2	25.4		
				(43.5)	(34.7)		
% of the doc that is copied conditional on being copied				82.7	55.0		
				(27.3)	(31.3)		

Notes: The table gives summary statistics. Year is 2013. Variables are values for documents. In columns 1 to 3, both internal and external verbatim copying are taken into account. In columns 4 and 5, we focus on external copy only. Column 4 (respectively column 6) includes the documents published by the AFP, and columns 5 (respectively column 7) excludes them. In columns 6 and 7, we focus on external copy and exclude the content copied from the AFP. "bk news" stands for breaking news. The different variables are described in details in the text.

Table 5: Summary statistics: Citations

	All events		Non-exclusive		Exclusive		Multiple witnesses	
	(1) All external copy	(2) Excl. copy from AFP	(3) All external copy	(4) Excl. copy from AFP	(5) All external copy	(6) Excl. copy from AFP	(7) All external copy	(8) Excl. copy from AFP
	share	share	share	share	share	share	share	share
% of docs crediting (at least one of) their source	32.5	3.5						
% of docs crediting doc they copy the most	26.3	3.8						
% of docs crediting breaking news outlet	25.1	8.8	24.5	8.3	34.1	16.7	30.8	6.1

Notes: The table provides the share of the documents presenting at least some external verbatim copying (i) crediting (at least one of) the media outlet(s) whose document(s) they copy (their source); (ii) crediting the media outlet whose document they copy the most; (iii) crediting the breaking news media outlet. Columns 1 and 2 present the results for all the news events. Columns 3 and 4 present the results for non-exclusive news events. Columns 5 and 6 present the results for exclusive news events. Columns 7 and 8 presents the results for short news items with multiple witnesses. In the odd columns, we present the results when considering all external verbatim copying. In the even columns, we exclude copy of the AFP.

Table 6: Article-level analysis: Number of Facebook shares (log-linear estimation)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Publication rank	-0.0005***						-0.0005***	-0.0005***
	(0.0001)						(0.0001)	(0.0001)
Reaction time (in hours)		-0.0040***					-0.0022***	-0.0022***
		(0.0005)					(0.0006)	(0.0005)
Originality rate (%)			0.0065***				0.0066***	
			(0.0001)				(0.0001)	
Length (thsd ch)				0.0825***			0.0876***	
				(0.0026)			(0.0026)	
Original content (thsd ch)					0.2083***			0.1974***
					(0.0030)			(0.0031)
Non-original content (thsd ch)						-0.1029***		-0.0274***
						(0.0025)		(0.0024)
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	0.48	0.48	0.49	0.48	0.50	0.49	0.50	0.50
Adjusted R-sq	0.46	0.46	0.47	0.46	0.48	0.47	0.48	0.48
Observations	667,180	667,180	667,180	667,180	667,180	667,180	667,180	667,180
Clusters (event)	24,502	24,502	24,502	24,502	24,502	24,502	24,502	24,502

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. 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 7: Article-level analysis: Number of tweets (log-linear estimation)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Publication rank	-0.0002***						-0.0002***	-0.0002***
	(0.0000)						(0.0000)	(0.0000)
Reaction time (in hours)		-0.0029***					-0.0024***	-0.0024***
		(0.0003)					(0.0003)	(0.0003)
Originality rate (%)			0.0030***				0.0030***	
			(0.0001)				(0.0001)	
Length (thsd ch)				0.0575***			0.0597***	
				(0.0014)			(0.0014)	
Original content (thsd ch)					0.1024***			0.1061***
					(0.0016)			(0.0018)
Non-original content (thsd ch)						-0.0305***		0.0099***
						(0.0014)		(0.0014)
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	0.59	0.59	0.60	0.59	0.60	0.59	0.60	0.60
Adjusted R-sq	0.58	0.58	0.58	0.58	0.58	0.58	0.58	0.58
Observations	658,606	658,606	658,606	658,606	658,606	658,606	658,606	658,606
Clusters (event)	24,502	24,502	24,502	24,502	24,502	24,502	24,502	24,502

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependent variable is the log of the number of times an article is tweeted, retweeted or liked. 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 8: Article-level analysis: Number of Facebook & of Twitter shares (log-linear estimation), Relaxing the 10 documents condition

	Facebook shares		Tweets	
	(1)	(2)	(3)	(4)
Publication rank	-0.0005***	-0.0004***	-0.0001***	-0.0001***
	(0.0001)	(0.0001)	(0.0000)	(0.0000)
Reaction time (in hours)	-0.0006	-0.0005	-0.0026***	-0.0027***
	(0.0005)	(0.0005)	(0.0003)	(0.0003)
Originality rate (%)	0.0064***		0.0029***	
	(0.0001)		(0.0000)	
Length (thsd ch)	0.0922***		0.0591***	
	(0.0026)		(0.0013)	
Original content (thsd ch)		0.1910***		0.0985***
		(0.0034)		(0.0019)
Non-original content (thsd ch)		-0.0338***		0.0073***
		(0.0024)		(0.0014)
Media outlets FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes
R-sq	0.54	0.54	0.63	0.63
Adjusted R-sq	0.48	0.48	0.58	0.58
Observations	947,721	947,721	936,907	936,907
Clusters (event)	112,891	112,891	112,886	112,886

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. 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".