Review of Economic Studies (2019) **0**, 1–39 doi:10.1093/restud/rdz061

Advance access publication 11 December 2019

The Production of Information in an Online World

JULIA CAGÉ
Sciences Po Paris and CEPR
NICOLAS HERVÉ and MARIE-LUCE VIAUD
Institut National de l'Audiovisuel

First version received December 2017*; Editorial decision November* 2019*; Accepted November* 2019 (*Eds.*)

News production requires investment, and competitors' ability to appropriate a story may reduce a media's incentives to provide original content. Yet, there is little legal protection of intellectual property rights in online news production, which raises the issue of the extent of copying online and the incentives to provide original content. In this article, we build a unique dataset combining all the online content produced by French news media during the year 2013 with new micro audience data. We develop a topic detection algorithm that identifies each news event, trace the timeline of each story, and study news propagation. We provide new evidence on online news production. First, we document high reactivity of online media: one quarter of the news stories are reproduced online in under 4 min. We show that this is accompanied by substantial copying, both at the extensive and at the intensive margins, which may constitute a severe threat to the commercial viability of the news media. Next, we estimate the returns to originality in online news production. Using article-level variations and media-level daily audience combined with article-level social media statistics, we find that original content producers tend to receive more viewers, thereby mitigating the newsgathering incentive problem raised by copying.

Key words: Internet, Information spreading, Copyright, Social media, Reputation.

JEL Codes: L11, L15, L82, L86

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 to which citizens have access. 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, 2020). 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]

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" (Athey *et al.*, 2018).

The editor in charge of this paper was Nicola Gennaioli.

This makes it extremely difficult for news content providers to distinguish, protect, and reap the benefits of the news stories they produce, especially as there is very little legal protection of intellectual property rights in news production.

In this article, we address the following question: given the limited intellectual property protection in news media, what is the extent of copying in online news production, and what are the incentives to produce original content? Understanding the different mechanisms at play has implications for the modern media industry and may help inform ongoing debates about the quality of 21st-century journalism. It has also clear relevance for the current concerns about ill-informed voters and the negative consequences for the functioning of electoral democracies.[2]

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

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 (or news events). We then study the timeline of each story. In particular, for each story, we first determine the media outlet that breaks the story, and then analyse the propagation of the story, second-by-second. Covering a news story does not necessarily imply providing original reporting on this story. We study how much each media outlet contributes to a story. More precisely, we develop a plagiarism detection algorithm to quantify the originality of each article compared to all the articles previously published within the event. The algorithm tracks small portions of text (verbatim) that are identical between documents. We distinguish between content copied from articles published by news agencies (to which media outlets subscribe) and content copied from competing media outlets.

Next, and most importantly, we attempt to better understand why, despite online copying, there are still incentives for original news production. In order to address this issue, 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 outlet fixed effects). This allows us to estimate the returns to originality in online news production.

2. On the way the market for news affects political outcomes, see among others George and Waldfogel (2006), Gentzkow (2006), Snyder and Stromberg (2010), Gavazza *et al.* (2019) Cagé (2020).

3. The reason for using French media in 2013 is mostly data driven. Content data for this research were constructed as part of the OTMedia research project, a unique data collection program conducted by the French National Audiovisual Institute and focusing on year 2013.

4. Other studies have taken a transmedia approach to investigate media consumption patterns. See in particular Prat (2018) who builds a media consumption matrix using survey data on the U.S. covering television, radio, printed media, websites, and social media. See also Kennedy and Prat (2019).

Our main findings are as follows. First, the speed and magnitude of online copying appear to be very large. We find very high reactivity of online media: on average, news is delivered to readers of different media outlets 169 min after first being published on the website of the news breaker, but in less than 4 min in 25% of cases. The reaction time is the shortest when the news breaker is a news agency, and the longest when it is a pure online media. We show that high reactivity comes with a high level of verbatim copying. We find that only 32.5% of online content is original, and that moving from the first ventile to the last ventile of the reactivity distribution nearly doubles the originality rate.

Copy can come either from articles previously published by the media itself (internal copying) or from articles published by competitors (external copying), including news agencies. Even considering only external copying and excluding copying from news agencies, we find that, on the extensive margin, 61.8% of the articles present at least some copy. In other words, most of the articles in events contain copy, and could possibly be a substitute for the original news producer. Conditional on copying, the average external copy rate (excluding copy from the news agencies) is equal to 25.7%. Moreover, this figure most probably underestimates the economic threat that copying may constitute, given that in practice media assumably reproduce the most relevant parts of the articles they copy, a dimension we cannot capture empirically here.

Given the magnitude of copying, we then explore why there are still incentives for original news production in the online world. From a theoretical perspective, the impact of copying on newsgathering incentives depends on a number of different parameters, including readers' mobility across media outlets, the quality of the copy with respect to the original, and consumers' valuation of originality. By using survey data on patterns of online readership, we first show that most consumers tend to consume news on multiple outlets online, thereby suggesting that switching behaviour can play an important role. We present a very stylized theoretical framework to understand the different forces at play. On the one hand, readers have a preference for a specific media to which they tend to be loyal (*e.g.* for ideological reasons). On the other hand, they also have a preference for original news production.[5] Depending on the relative strength of these individual-specific parameters, and depending on the extent to which the copying media offers a lower-quality coverage than the original news producer, they might decide to consume the online content from the media that has produced original information, either at the daily level or on a longer-term basis. The "quality of the copy" parameter appears to play a central role. With high mobility across media, high-copy quality can drastically reduce the incentives for original news production.

We attempt to estimate some of the model parameters in the following way. First, we present evidence showing that copy is of lower quality than the original. In particular, copy tends to be incomplete. On average, when an article is copied, "only" 9.4% of its content is reproduced; this means that nearly nine-tenths of the content of the original article is missing from the copy. As we will discuss, incomplete copying might itself be due (at least in part) to legal restrictions on how much one is allowed to copy-and-paste from other media outlets. Furthermore, besides incomplete copying, there might be other reasons why copying leads to lower-quality articles, *e.g.*, because the original is bundled with additional information that is absent from the copy.

Second, using article-level variations (with event, date, and media outlet fixed effects), we show that a 50-percentage-point increase in the originality rate of an article leads to a 40% increase in the number of times it is shared on Facebook, and to a 17% increase in the number of Tweets. We discuss a number of possible channels that may help rationalize these findings. Our preferred explanation is that consumers may favour originality. In particular, investigating

5. Consumers' preference for originality may be interpreted as a shortcut that captures the fact that the originality of content is correlated with its quality. Copy may indeed be a manifestation of lousy journalism.

whether the returns to originality vary depending on the characteristics of the media outlets, we show that originality has a stronger positive effect for the outlets that operate in a more competitive environment and are therefore more subject to switching. We also find that the returns to originality are lower for the media that are more copied by their competitors. These heterogeneous effects are consistent with the predictions of our simple theoretical framework.

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

Finally, we combine media-level daily audience data and article-level social media statistics to obtain an audience measure (number of views) at the article level. We first assume a simple linear relationship between the number of shares on social media and the number of article views. We then use a unique data set on the number of views and Facebook shares at the article level from leading daily newspaper *Le Monde* to characterize the joint distribution of the number of Facebook shares and the number of visitors. We use these different estimates to obtain a lower and an upper bound of the number of times each article is viewed. We show that a 50-percentage-point increase in the originality rate of an article leads to a 45% increase in its number of predicted readers. Lastly, depending on the specification we use, we find that the original content represents between 45.4% and 61.4% of online news consumption, *i.e.*, much more than its relative share in total online content (32.5%). This result holds regardless of the measure of copying we use.

In brief, one way to summarize our findings is the following. In case online audience was distributed randomly across the different websites and regardless of the originality of the articles, then the magnitude of copying would severely lessen the economic returns to original news production, which as a first approximation can be assumed to be proportional to audience and revenues.[6] Importantly, given that the section of an article that is copied is probably the most important one and that the most copied articles are arguably the most interesting ones, the average copy rate measure may underestimate the audience-stealing effect of copying. More efficiently than aggregators' snippets, articles containing some copy may be a substitute for original news content. Yet nearly two-third of the articles in news events contain at least some external copy from media other than the news agencies. The extent of the extensive margin of copying reflects the economic threat copying constitutes. However, due to the fact that readers are more likely to consume content on the website of the original producers, the latter capture a larger share of online audience and revenues than the share that original content represents out of overall content. In other words, the fact that readers partly favour the original producers helps to mitigate significantly the newsgathering incentive problem raised by copying.

Of course, our results do not imply that reputation effects and consumers' preference for originality alone can solve plagiarism issues. Greater intellectual property protection could also

6. Advertising pricing on the Internet is indeed based on audience, and advertising is the largest contributor to publishers' online revenues (Anderson, 2012).

play a role in raising the incentives for original news production, and we certainly do not mean to downplay the extent of this problem. Both in absolute terms and in relation to books, music, and movies, there is very little protection of intellectual property rights in news production. In 2013 France, the copyright law was governed by the 1992 "Code de la propriété intellectuelle" (French Intellectual Property Code). The French code—like the majority of the copyright laws around the world—does not protect facts but only the articles that are considered "original." However, what makes an article original is a complex issue (*e.g.* the amount of labour required is an open question). Similarly, while "substantial" copying is required in order to constitute a violation of the right of reproduction, there is no clear definition of what substantiality means in this context (Ginsburg, 2016).[7] In other words, opportunities for the infringement of intellectual protection abound in the context of the news media.[8] However, our results suggest that in order to effectively address the plagiarism issue, it is also important to study reputation effects and viewers' reaction to the newsgathering investment strategies of media outlets.

1.1. *Related literature*

Using micro data, Gentzkow (2007) estimates the relationship between the print and online newspapers in demand.[9] Our article is complementary to his. We investigate the production of original content and document the benefits of original information production. Franceschelli (2011) was the first to assess empirically the impact of the Internet on news coverage.[10] 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. Compared to this previous work, our contribution is threefold. First, we construct the set of news stories and study their timeline using the entire universe of French news media online, rather than two newspapers. To the extent of our knowledge, we are the first to study simultaneously the content produced by all the news media, whatever their offline format. Second, while Franceschelli (2011) relies restrictively on the mention of proper nouns to identify the news stories, we develop and run a state-of-the-art algorithm relying on word frequency without any restriction. Hence, our paper also contributes to the existing literature from a methodological point of view. In particular, we develop a new event detection algorithm that could be of future use to other researchers interested in text analysis and clustering. 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.*, 2018). Most of the literature on copyright online has centred on digitization and piracy within the music industry (Rob and Waldfogel, 2006; OberholzerłGee and Strumpf, 2007; Waldfogel, 2012, 2015).[11] With the exception of Chiou and Tucker (2017), there is little evidence on copying and intellectual property regarding online news media. Yet, the modern news media industry

7. See the Supplementary Appendix Section B for a discussion of copyright law in France and the United States.

8. In April 2019, an European Copyright Directive was adopted by the European Council. Article 15 of this Directive forces online platform and aggregators to pay press publishers to use their content. In September 2019, France was the first country to have transposed this directive into national law.

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

10. 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.

11. Recent work has also investigated the effect of digitization projects like Google Books (Reimers, 2019; Nagaraj, 2018).

shares a number of important characteristics with the cultural industry online; in particular, digital products just like news articles are non-rival, non-excludable, and can be copied at almost no cost (see *e.g.* Bae and Choi, 2006; Peitz and Waelbroeck, 2006a). We contribute to this literature by providing new empirical evidence on the extent of copying online and estimating the returns to originality. Our paper is a unique attempt to understand who is producing news, the character of what is produced and the propagation of information in the online world.[12]

The rest of the article is organized as follows. In Section 2 below, we describe the media universe and the content data we use in this paper, and briefly review the algorithms we develop to study the production and propagation of information online.[13] Section 3 provides new evidence on the speed of news dissemination and the importance of copying online, and discusses heterogeneity in the copying behaviour and media outlets' reputation. In Section 4, we discuss the mechanisms at play and the theoretical framework that we use to analyse the impact of originality and copying on consumer behaviour. In Section 5, we use article-level variations to investigate the relationship between originality and online audience and estimate the returns to originality in online news production. Section 6 performs a number of robustness checks and discusses the external validity of our main findings. Finally, Section 7 concludes.

2. DATA AND ALGORITHMS

2.1. *Media universe*

Our dataset covers 86 general information media outlets in France: 1 news agency; 59 newspapers (35 local daily, 7 national daily, 12 national weekly, 2 national monthly, and 3 free newspapers); 10 pure online media (*i.e.* online-only media outlets); 9 television channels; and 7 radio stations. The news agency is the Agence France Presse (AFP), the third largest news agency in the world. Moreover, our dataset also includes all the dispatches published in French by Reuters. For each of these media outlets, we gather all the content they published online in 2013.[14]

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

We choose a "transmedia" approach because, on the Internet, there is a tendency for different media to converge (see *e.g.* Peitz and Reisinger, 2016). On the web, media all offer texts, videos and photos. We include the AFP and Reuters even though they do not deliver news straight to

12. Sen and Yildirim (2015) investigate how popularity of online news stories affect editors' decisions.

13. See the working paper version of this research (Cagé *et al.*, 2017, pp. 6–17) and the Supplementary Appendix Section C for a more detailed description of the data sources and algorithms.

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

15. The only media outlets not included are some local daily newspapers that had no websites at the time, and some very small digital news media that could not be considered important information providers in 2013.

16. Moreover, as we will see in Section 6.3, the French media market is by and large very similar to other Western media markets, and it has remained steady since 2013.

individual consumers[17] because they are key providers 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.

Using their RSS feeds, we track every piece of content news media produced online in 2013. For the media outlets whose RSS feeds were not tracked by the INA, we complete the OTMedia data by scraping the Sitemaps of their website. We acquire all the AFP and Reuters dispatches directly from the agencies. Merging these datasets, we obtain the universe of all the articles published online by French news media in 2013. The articles we use in our database contain text and often photos, as well as videos. Our focus here is on text.[18]

Our dataset contains 2,552,442 documents for the year 2013; around 7,000 documents on average per day. 70.9% of the documents are from the websites of the print media; 4.5% from radio; 6.4% from television; 15.1% from the AFP and Reuters and the remaining documents from the pure online media. On average, these documents are 2,058 characters long.[19] Interestingly, while media outlets do not face the same space constraint online that they face offline (see *e.g.* Eisensee and Strömberg, 2007), the total amount of content produced on a daily basis is relatively stable through time. While space online is technically infinite, media outlets indeed still face an implicit space constraint which is the limited attention span of the readers.

2.2. *News events*

2.2.1. Event detection algorithm. Using the set of documents previously described, we perform an event detection algorithm to detect media events. This category of algorithm is often referred to as Topic Detection and Tracking (TDT) in the computer science community. These algorithms are based on natural language processing methods. An event is defined as a set of documents belonging to the same news story. Events are detected by our algorithm using the fact that the documents share sufficient semantic similarity. We only keep events with documents from at least two media outlets, and with more than 10 documents in our preferred specification.[20] We obtain a total number of 25,215 news events. Events can last more than one day; on average, they last 41 hours. The average number of documents per event is 34 and, on average, 15 media outlets refer to an event.[21] There are 182 events per day on average, with 69 new events beginning every day. These events are roughly equally distributed during the year. In this paper, given that our subject of interest is the propagation of news stories online and the importance of copying, we focus our main analysis on the 851,864 articles classified in our 25,215 events. Table 1 provides summary statistics on these articles.

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

18. We do not study the online production of videos and photos. Analysing the propagation of photos and videos online requires different technical tools and algorithms than those we develop here and will be the topic of future research.

19. See Supplementary Appendix Figures F.1, F.2a and F.3 and Table E.1.

20. 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 several robustness checks. In particular, we run it on a standard benchmark dataset (the Topic Detection and Tracking Pilot Study Corpus), and we compare our events to those obtained by the Europe Media Monitor (EMM) NewsExplorer. The EMM NewsExplorer provides on a daily basis the top 19 stories of the day. With our event detection algorithm, we match 92% of their stories in our sample. Full details about the algorithm are provided in Cagé *et al.* (2017, pp. 9–14) and in the Supplementary Appendix Section C.2.

21. Supplementary Appendix Table E.2.

TABLE 1
Summary statistics: articles (classified in events)

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

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

2.2.2. 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). 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.

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.[22] This does not mean that there is no article related to these topics, but that these topics are not associated with *events*.

2.3. *Timeline and plagiarism detection*

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

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

2.3.2. Plagiarism detection algorithm. The plagiarism detection algorithm efficiently tracks identical portions of text between documents. For each document, we determine the portions

22. See Supplementary Appendix Figure F.4.

TABLE 2
Summary statistics: media outlets

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

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

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. This allows us to determine for each document the number of times it is copied and the share of the document which is ultimately copied.

2.4. *Audience data*

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

2.4.1. 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[23]—we have information on the number of unique visitors, the number of visits, and the number of page views. 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.

2.4.2. Social media data. 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). We obtain information on this variable for all the documents in our sample, with the exception of the articles published by the AFP and Reuters that are not available online to the general audience. On average, articles are shared 64 times on Facebook; however, half of the articles are not shared (see Table 1 for summary statistics).

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

Facebook shares may 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. For each article, we have eight different measures of the number of times it its "shared" on Twitter: the number of direct tweets, retweets, likes, and replies, and then computing the statistics on the retweets and the replies, the indirect number of tweets, retweets, likes, and replies. Obviously, all these different measures are very strongly correlated.[24] For the sake of simplicity, in our preferred specification, the total number of times each article is shared on Twitter is defined as the sum of the values for these eight measures.

Lastly, note that there is a positive relationship between the number of shares on Facebook and the number of shares on Twitter thus defined. In Section 5, we use these social media statistics as a proxy for the number of views. To test the accuracy of this proxy, we rely on evidence from *Le Monde* newspaper and show that the relationship between the number of views and the number of social media shares is almost perfectly linear.[25]

3. THE SPEED AND MAGNITUDE OF ONLINE COPYING

In this section, we first study the speed of news dissemination online. That is, we investigate how quickly news is delivered to readers of different media outlets after being published first on the website of the news breaker. We then analyse the magnitude of online copying.

3.1. *The speed of news dissemination*

Studying the speed of news dissemination is of interest because the commercial value of a news item may depend on how long a news media retains exclusive use of it. We first study the time interval between the publication of the first document covering a story and the second one. We find that on average, it takes 169 min 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 min, of which less than 243 s in 25% of the cases and less than 6 s in 10% of the cases.

Table 3 reports the average reaction time depending on the offline format of the news breaker. If a news agency (the AFP or Reuters) is the first media outlet to publish some information, then the reaction time is shorter. When a news agency is the news breaker, we find that the second media outlet covers the story after 116 min on average, but after only 11 min in half of the cases and in 1 s or less in 5% of the cases. This rapidity comes from the fact that media outlets receive the news directly from the news agency; they do not 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 (*i.e.* AFP content of their choice is automatically integrated into their website). The average reaction time is longer when the news breaker is a media outlet other than the news agency.

This appears clearly in Figure 1, which represents the Kaplan–Meier survival functions depending on whether the news breaker is a news agency or another media outlet. We also find that the reaction time is the highest when the news breaker is a pure online media. 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.

24. Supplementary Appendix Table C.1.

25. Although this is an argument in favour of using such a proxy, it must be kept in mind that not all the online readers share articles on social media. In the Supplementary Appendix Table E.3, we show that while the readers who

TABLE 3
Reaction time

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

Notes: * $p<0.10$, ** $p<0.05$, *** $p<0.01$. The tables give summary statistics for the reaction time (in minutes).

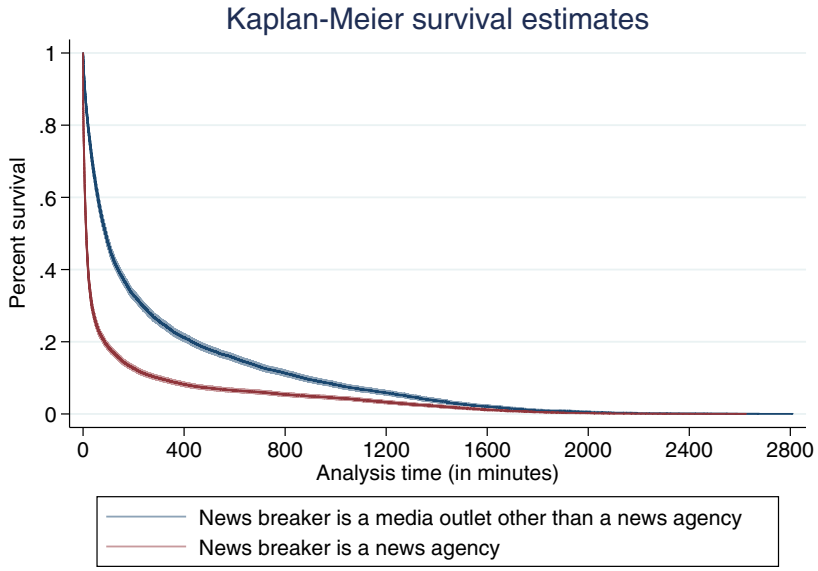


FIGURE 1
Average reaction time depending on whether the news breaker is a news agency: the Kaplan–Meier survival estimates

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

3.2. *The importance of copying online*

We now turn to an estimation of the originality of online articles. This is a key question because the high reactivity of the media may actually be the result of plagiarism, and the recourse to plagiarism may negatively affect newsgathering incentives.

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

share news articles on social media are selected, this selection seems to come mostly from age. See also Cagé *et al.* (2017, pp. 16–17).

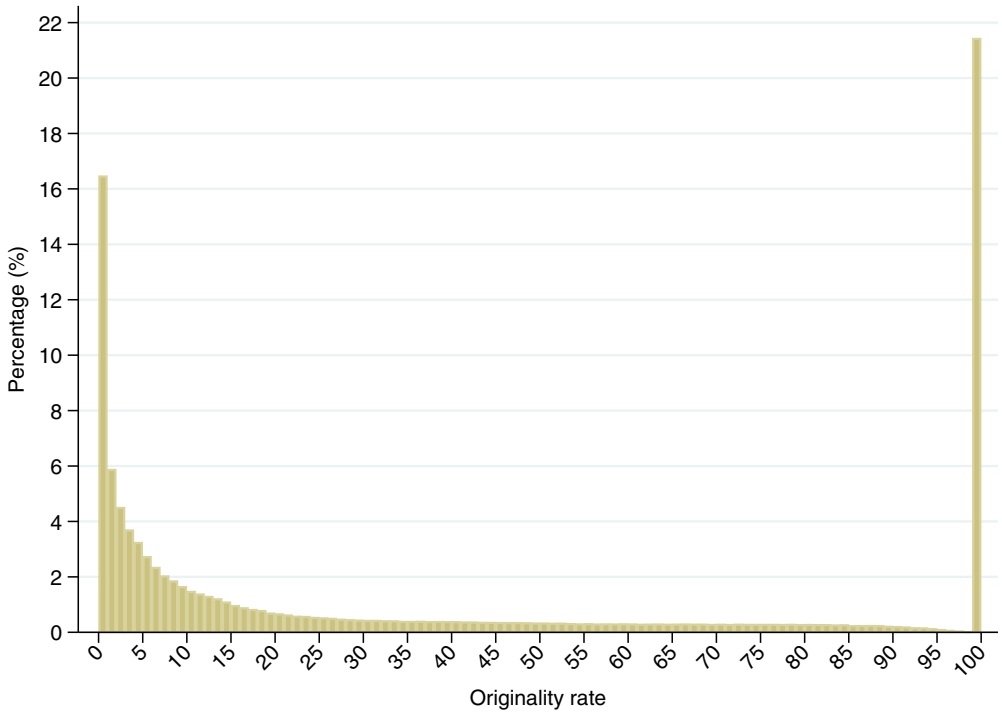


FIGURE 2
Originality rate

Notes: The figure plots the distribution of the originality rate (with bins equal to 1%).

On average, the originality rate of the documents classified in events is equal to 36.5%.[26] In Figure 2, we plot the distribution of the originality rate. The distribution is bimodal with one peak for the articles with less than 1% of original content (nearly 17% of the documents) and another peak for the 100% original articles (nearly 22% of the documents). The median is 14%. In other words, with the exception of the documents which are entirely original, the articles published within events consist mainly of verbatim copying: 54.6% of the articles classified in events have less than 20% originality.

Figure 3 shows the average originality rate of the articles for each ventile of the reactivity distribution. On average, the longer it takes for a media outlet to cover an event, the higher the originality rate of the article: moving from the first to the last ventile of the reactivity distribution increases the average originality rate from around 26% to around 40%.[27]

3.2.2. 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)

26. Given that documents are of different lengths, we also compute the ratio of original content in the dataset out of 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.

27. This finding is robust to dropping the articles published by the news agencies, and to computing the reactivity distribution at the media-outlet level (Supplementary Appendix Figures F.6 and F.7).

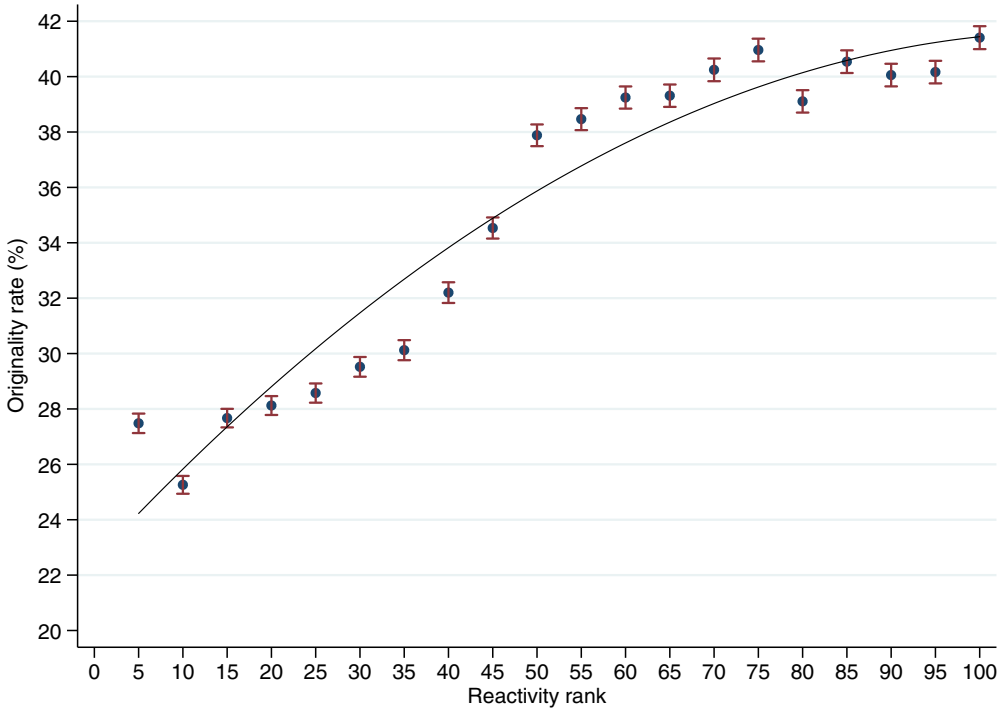


FIGURE 3

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

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

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. Table 4 presents the results for all the documents classified in events, with the exception of the press dispatches published by the AFP and Reuters.

Verbatim copying can be either "internal," when a media outlet copies and pastes content from documents it has itself previously published[28], or "external," when a media outlet reproduces content from articles published by its competitors. The mean originality rate appears to be equal to 35.1%, *i.e.*, 64.9% of the content is copied to other articles, including 3.9% from internal copying and 61.0% from external copying. External copy can come either from content published by the news agencies or from content published by media outlets other than the news agencies. Column (4) of Table 4 presents the results for all external copy, and Column (5) for external copy excluding copying from the news agencies.

Regarding the extensive margin of copying, we find that, when copying from the agencies is included, 77.5% of the articles in events contain at least some external copying. If we exclude content copied from the news agencies[29], we show that notwithstanding 61.8% of the articles present at least some copy. In other words, most articles in the news events contain some external copying and may potentially negatively affect the newsgathering incentives of the copied media.

28. For example, when it is updating previous versions of the same article.

29. All the AFP's clients are indeed allowed to reproduce the AFP content in its entirety, and the business model of the news agency is based on the reproduction of its content by other media outlets (similarly for Reuters).

On the intensive margin, we find that the average external copy rate is equal to 61.0% when copy from news agencies is included, and to 15.9% when we exclude content copied from the news agencies. Conditional on copying, the average external copy rate is equal to 78.7% when copy from news agencies is included, and to 25.7% when we exclude content copied from the news agencies. This is a very large number: this means that over a quarter of the content is directly taken from articles published in media that are not supposed to be copied (and that are not rewarded for it). Moreover, it is likely that the part that is copied is the most valuable one, although we cannot measure this properly.

On average, the share of a document that is copied is equal to 9.4%. But the copied part of an article is again very likely to be the most interesting one. Hence, the extensive margin of copying may be more informative than the average copy rate when it comes to estimating the audience-stealing effect of plagiarism. A copying article containing the most valuable part of the copied article may indeed be a substitute for the original article.

Finally, we find that each document is copied by 3.9 documents, 3.3 if we focus only on external copying. The articles copied most extensively are in all likelihood the most interesting ones. The main limitation of the empirical analysis performed here is that we can only measure the extent of online copying, not compute an exact estimate of its severity that would require the estimation of the relative importance of the content that is copied and of the content that is not. But from the figures presented in Table 4, online copying appears as a substantial threat to the economic viability of the news media, and may reduce media's newsgathering incentives. Furthermore, even though our estimation is imperfect, it is important to highlight that to the extent of our knowledge we are the very first to quantify in a systematic way the extent of online copying in news.

3.3. *Copying behaviour and reputation*

Overall, on average, media outlets tend to rely a great deal on plagiarism. But do all the news media display the same copying behaviour?

We compute the average copy rate separately for local newspapers, national newspapers, television stations, radio channels, and pure online media. It appears that local newspapers tend to produce less original content online than other types of media outlets.[30] A possible explanation is that, while local newspapers may compete on local news to attract readers, they may rely on copying when it comes to national news. Coherently with this assumption, we find that local newspapers tend to strongly rely on content produced by the news agencies.

We also find that pure online media tend to be on average more original than other media outlets. This may be due to the fact that these pure online media "seek to offer distinctive voices" (Nicholls *et al.*, 2016). Note however that pure online media only account for 3% of the documents in our dataset[31], and that we should not overstate these differences in editorial priorities. French pure online media indeed "closely approximate an online newspaper" (Nicholls *et al.*, 2016).

Besides, we have compared the media outlets depending on their "reputation" as defined by using their reliance on copy and the use of their content by others. Interestingly, the media outlets with the strongest reputation (*Le Monde*, *Les Echos*, France Television, *Ouest France*, etc.) are also those with the largest audience (average numbers of unique visitors per day) and the highest numbers of shares on social media. We find similar effects if we proxy the reputation of the media outlets by the number of "citations" they receive (*i.e.* the number of times their competitors refer

30. See Supplementary Appendix Figure F.8.
31. Moreover, only 25% of the pure online media documents are classified in events.

TABLE 4
Summary statistics: copy

	All copy			External copy	Excl. copy from agencies
	(1) Mean/SD	(2) Mean/SD	(3) Mean/SD	(4) Mean/SD	(5) Mean/SD
Originality rate	35.1				
	(39.0)				
Originality rate wghtd by		58.0			
nb of views (Facebook)		(38.2)			
Extensive margin					
Share of articles in events			80.0	77.5	61.8
with at least some copy					
Copying media					
Nb docs copied			4.1	3.9	2.3
			(5.0)	(4.7)	(3.3)
External copy rate				61.0	15.9
				(39.3)	(25.0)
External copy rate				78.7	25.7
conditional on copying				(24.5)	(27.6)
Copied media					
Nb copying docs			3.9	3.3	
			(9.1)	(8.2)	
% of the doc that is copied				3.9	
				(12.5)	
% of the doc that is copied				9.4	
conditional on being copied				(17.9)	
If copied media bk news				24.1	
% of the doc that is copied				(33.9)	
% of the doc that is copied				53.5	
conditional on being copied				(31.3)	

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

to them as the source of the information).[32] Obviously, these effects measured at the media level do not imply that reputation has a causal effect on audience, but they are nevertheless suggestive. In particular, they are consistent with the possibility of long-run reputation effects. We will later explore these issues by looking at the impact of originality on short-run audience at the article level.

4. COPYING AND NEWSGATHERING INCENTIVES: EXPLORING THE MECHANISMS

In the previous section, we have quantified the speed of news propagation online and the magnitude of copying. These results lead us to the following paradox: given all this copying, why is there any original news production at all? From a theoretical perspective, the impact of copying on newsgathering incentives depends on a number of different parameters, including readers' mobility across media outlets, the quality of the copy with respect to the original, and consumers' valuation of originality.

In this section, we start by using survey data in order to document the patterns of online readership. We show that most consumers tend to consume news on multiple outlets online,

32. See Supplementary Appendix Figures F.9–F.11, and Cagé *et al.* (2017, pp. 22–26) for a detailed discussion.

thereby suggesting that switching behaviour can be substantial. We then present a very stylized theoretical framework to understand the different forces at play in a setting where mobility across media plays an important role. It becomes clear that a key parameter is the extent to which copied articles are of lower quality than the original (which will depend on copyright law and other factors). With high mobility across media outlets, high copy quality can drastically reduce the incentives for original news production.

In the next section, we will then attempt to estimate some of the model parameters. Using article-level variations and media-level daily audience combined with article-level social media statistics, we find that readers are more likely to consume news on the website of the original producers. As we will see, this can be rationalized by the fact that the quality of the copy appears to be relatively low. This also reflects consumers' strong taste for originality, particularly for the media outlets operating in highly competitive environments. In turn, this consumption behaviour helps both to mitigate the newsgathering incentive problem raised by copying and to solve our paradox.

4.1. *Readers' mobility across the media*

We first document the extent of consumer switching across media outlets. In the event that readers were sparsely mobile across outlets, being original or being copied should have little impact on a media outlet's audience. Another possibility is that readers are mobile and shop for the best news across media outlets. This potentially raises the incentives for original news production, but also makes copying more problematic.

Recent studies of audience news consumption behaviour have indicated that news users increasingly rely on multiple news media (see *e.g.* Pew Research Center, 2016; Reuters Institute, 2017). Given that "people have more power to navigate the news content they want to use, when, where and how" (Swart *et al.*, 2017), they seem to shop for the best news across outlets online. As a consequence, they follow the news on multiple media platforms (Picone *et al.*, 2015; Yuan, 2011).

In order to document the patterns of online readership and the extent of readers' mobility in France, we use survey data from the *2013 Digital News Report* (Reuters Institute, 2013).[33] The sample includes 1,016 individuals for France for the year 2013. Among the survey questions, respondents are asked whether they followed different media outlets online.[34] Out of the nine television channels included in our sample, five are covered by this question regarding online news consumption; thirteen national newspapers (out of twenty-four); and five pure online media (out of ten). Furthermore, radio stations in our sample are grouped into two categories: private radio and public radio. Finally, from the "other" category, we compute a measure of the online consumption of local newspapers.

We see that nearly two-thirds of the surveyed individuals consume at least one media outlet online. Among those who consume at least one news media, the average number of outlets consumed is equal to 2.35; in other words, users spread their news consumption over multiple platforms online. We also use this survey data to build a matrix of proximity across media outlets. That is, we compute the probability that a respondent accessing one website also accesses the other. Three main media ensembles appear: one including center-left and left media outlets (*Le Monde*, *Libération*, Mediapart, etc.); one including center-right and right media outlets (*Le Figaro*, *Le Point*, *L'Express*, etc.); and one including free newspapers (*20 Minutes*, *Direct Matin*), TF1, the

33. Similar data have been used by Kennedy and Prat (2019).

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

private radios, etc. Respondents reading *Le Monde* also tend to consume news from *Libération*, and the same applies to the other two groups.[35]

In other words, we find that the patterns of online readership reflect both mobility and loyalty. That is, most readers consume multiple media outlets (and are therefore likely to switch to outlets with higher quality content), but at the same time they have specific ideological or preference-based loyalty for particular groups of media.

4.2. *Quality of copying, valuation of originality: a simple model*

Taking as given these patterns of online readership, it is useful to outline a simple theoretical framework in order to clarify the main mechanisms through which copying may negatively affect newsgathering incentives and assist in interpreting the empirical results. We summarize below some of the main forces and parameters at play, and refer the reader to the Supplementary Appendix for a formal description of the model.[36]

The three key parameters of this simple framework are the consumers' loyalty to a particular media, the consumers' taste for originality, and the quality of the copy with respect to the original. Consumers are heterogeneous with respect to their taste for originality, and face a trade-off between their loyalty to their preferred media and their taste for originality, depending on the quality of the copy. As long as the ratio of the average taste for originality over loyalty is high enough compared to the relative quality of the copy, at least some readers will switch across the media. Furthermore, we show that when media outlets are more "isolated" (in the sense that they are in competition with fewer outlets), there are lower returns to originality, a prediction that we will test in the next section, where we use media-level daily audience and article-level social media statistics to quantify the returns to originality.

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

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

Second, online, articles tend to be published along with photographs, videos or other kinds of illustrations, *e.g.*, data visualizations, visual stories, and graphics. In this article, we only

35. See Supplementary Appendix Figures F.12 and F.13, and the discussion in Cagé *et al.* (2017, pp. 27–29).
36. Supplementary Section A.

consider text. However, having "manually" analysed the websites of a number of media outlets and discussed the question with a number of publishers, our educated guess is that while plagiarism is a common practice regarding text, it is very uncommon to reproduce the illustrating images alongside, in particular when it comes to visualizations. While text plagiarism falls in a grey area in terms of copyright enforcement (due to the right-to-quote exemption, the issue of the substantiality of copying and of the originality of the copied work, etc.), photos and data visualizations are more clearly copyrighted, and it is much easier to identify the infringement. In other words, the original may be of higher quality than the copy because the original is bundled with photographs, visual stories, etc. that are absent from the copy.[37]

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

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

The arguments detailed above help to rationalize the positive relationship we describe below between originality and news consumption as proxied by the number of shares on social media. Consumers may favour original content over copy because the original content is of higher quality. Moreover, original news producers may also benefit from an increase in their audience through a sampling effect. In the context of the music industry, this effect corresponds to the fact that "downloaders use the downloaded files for sampling in order to make more informed purchasing decisions" (Peitz and Waelbroeck, 2006b). In the case of the news media industry, this could take the form of readers discovering a new media outlet by reading its original content reproduced on the website of its competitors. Finally, note that despite consumers' valuation of originality, some readers may consume the copy as long as the ratio of the average taste for originality over the consumers' loyalty to media brands is high enough (in our very simple theoretical framework, $\frac{\bar{v}}{\bar{u}} > \frac{1}{2(1-\lambda)}$). Ultimately, quantifying the returns to originality is an empirical issue.

5. ONLINE AUDIENCE AND THE RETURNS TO ORIGINALITY

In this section, we attempt to address the following key question: given the magnitude of online copying, what are the incentives to produce original content? Using article-level variations and media-level daily audience combined with article-level social media statistics, we show that an increase in originality leads to an increase in audience, thereby mitigating the newsgathering incentive problem raised by copying.

37. Although copyright enforcement appears to be relatively low regarding text plagiarism, incomplete copying might also be due (at least in part) to legal restrictions on how much one is allowed to copy-and-paste from other media. That is, in most legal systems only short quotations are usually allowed. In the event that some media outlets tried to systematically copy-and-paste 100% of the original text content produced by other outlets (rather than merely 10% or 20%), it is likely that there would be greater pressures to carry out effective enforcement.

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). We attempt to overcome this limitation by using alternative article-level information that we collect from Facebook and Twitter. Furthermore, we use an additional dataset from *Le Monde* to relate article-level Facebook shares and article-level numbers of views.

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

5.1. *Social media statistics and number of views*

5.1.1. Evidence from *Le Monde* data. What is the relationship between the number of times an article is viewed and the number of times it is shared on Facebook? Answering this question is of particular importance for us given that our approach uses this relationship to compute statistics on the number of views per article.[38] To understand the mapping between article views and number of Facebook shares, we obtain 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 the number of shares on Facebook. On average, during this time period, each *Le Monde* article is viewed by 19,656 unique visitors and shared 1,015 times on Facebook.[39]

Figure 4 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 4a). 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 4b shows the results. We approximate the relationship between the rank in the number of visitors distribution (P_{Vadn}) and the average number of visitors (as a multiple of the mean number of daily visitors) by a polynomial of degree six (so as to obtain the best possible

38. A number of articles in the literature simply assume that exposure is proportional to Facebook shares (see *e.g.* Allcott and Gentzkow, 2017). However, this assumption is questionable and we therefore made the choice here to document this relationship empirically.

39. In the Supplementary Appendix, we provide detailed summary statistics for these two variables (Supplementary Appendix Table E.4) and plot their spread and skewness functions (Supplementary Appendix Figure F.14).

fit). We also use alternative non-linear specifications and show that this has a limited impact on our main results (see below).

5.1.2. Article-level estimation of the audience. In what follows, we use the relationship uncovered thanks to the *Le Monde* data as our preferred specification to estimate article-level audience. It will also be useful to compare our findings with simple "naive" and "linear" estimates of article-level 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).

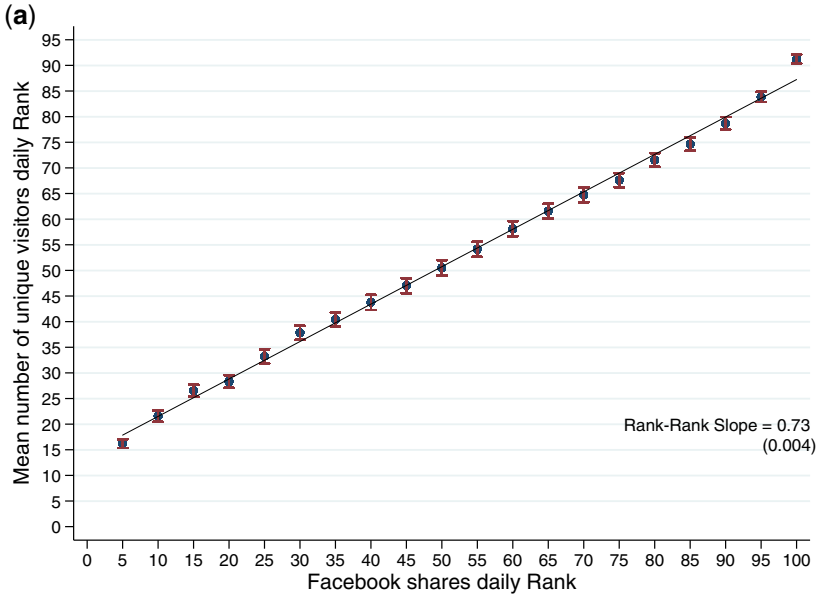
Social media approach, assuming linear relationship. In the linear social media approach, we use the information on the number of Facebook shares (respectively on the number of Tweets) to obtain a less naive measure of the audience of each article. More precisely, we compute for each media/day the total number of shares and then attribute a number of views to each article as a function of its relative number of shares.

Social media approach, using estimates from Le Monde (rank-rank approach). In our favoured approach, we use the estimated parameters from *Le Monde*'s article-level data to approximate the number of views of each article.[40] For the sake of robustness, we use two different methodologies: a rank–rank approach and a blinder approach simply regressing the share of the total number of daily views represented by each article on its share of the total number of Facebook shares.

The rank–rank approach relies on the findings described above (Section 5.1.1). First, for each article, we compute its rank in the Facebook shares distribution (P_{FBadn}) and then use the estimated coefficients from *Le Monde* (slope equal to 0.73 and intercept equal to 14.20) to impute its rank in the number of visitors distribution ($\widehat{P_{Vadn}}$). Then, from the total number of views received by the media outlet n on date d, we estimate the number of views of each article by using the parameters obtained when estimating the following relationship using *Le Monde* 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* data, we perform the following estimation:

40. This approach relies on two assumptions. First, we assume that the relationship between the number of Facebook shares and the number of views we uncover for 2017 also holds in 2013. We think this assumption is relevant given the stability of the media market we discuss in Section 6.3. Second, we assume that this relationship is constant across all the media outlets in our sample. While we cannot test this assumption, in the Supplementary Appendix Section 1.4 we provide additional evidence of the shape of the relationship using a dataset we obtain from Parsely. This dataset covers the year 2017 and contains information on the number of clicks and on the number of shares on social media originating from the U.S. for 1,363,308 articles published in English. These data rely on articles published by a large number of different media outlets and, in line with the evidence we obtain using *Le Monde* data, we find that the relationship between the number of clicks and the number of shares is almost perfectly linear.



Association between number of Unique visitors' and Facebook shares' Percentile Ranks

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

FIGURE 4

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 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 4a, we plot the relationship between the articles' Facebook shares' percentile rank and the average value of the visitors percentile rank (error bars represent the 95% confidence interval). The slope of this relationship is equal to 0.73. In the bottom Figure 4b, 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 represent the 95% confidence interval). The line is the predicted value of the average number of visitors when this relationship is approximated by a polynomial of degree six.

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

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

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

To estimate the returns to originality using article-level estimations, we consider three dependent variables: the number of times an article is shared on Facebook, the number of times it is shared on Twitter, and its predicted number of readers, and estimate how they vary with its originality and reactivity. We then investigate whether the returns to originality vary depending on the characteristics of the media outlets, which allows us to rationalize the positive relationship between originality and news consumption we obtain.

5.2.1. Number of Facebook shares. We use article-level data to investigate how the number of times an article is shared on Facebook varies with its originality and reactivity. Given that the distribution of the number of Facebook shares is right-skewed, we perform a log-linear estimation. Equation (5.1) describes our preferred identification equation (the observations are at the article level):

$$\text{Facebook shares}_{aedn} = \alpha + \mathbf{Z}'_{\mathbf{aedn}}\beta + \lambda_e + \gamma_n + \delta_d + \epsilon_{aedn}, \tag{5.1}$$

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.[41]

$\mathbf{Z}'_{\mathbf{aedn}}$ 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, then equal to 2 for the article published next in the event, then to 3,...); (ii) the reaction time (which is equal to 0 for the breaking news article and is then a measure of the time interval between the publication time of the considered article and that of the breaking news article); (iii) the originality rate of the article (in percentage: the variable varies from 0% to 100%); (iv) the length of the article (total number of characters in thousands); (vi) the original content (also by number of thousand characters); and (vii) the non-original content. Alternatively, we use an indicator variable equal to one for the breaking news article, and to zero otherwise, and then only control for the length of the article. Regarding the rank and reactivity measures, we anticipate 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 anticipate a positive sign for our measures of originality (the originality rate and the original content), as well as for the breaking news indicator variable.

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

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

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

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

Columns (1)–(3) of Table 5 present the results. Regarding originality, we find that an increase of 1,000 in the number of original characters leads to a 22% increase in the number of Facebook shares. If we consider the originality rate instead, we show that a 50-percentage-point increase in the originality rate of an article (*e.g.* moving from an article with no original content to an article with 50% originality) leads to a 40.5% increase in the number of Facebook shares. If we now turn to reactivity, we find that both the publication rank and the reaction time matter. The effect is economically small, however: taking 41 h (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%. Yet, we observe high returns from being the news breaker: according to our estimates, being the breaking news article more than doubles the number of Facebook shares received by an article. Given that our specification uses within media outlet-event-date variation, we interpret these estimates as causal.

Robustness. In order to take into account non-linear 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 (5.1) using as independent variables these categorical variables rather than the continuous originality rate measure. Figure 5 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%.

Equation (5.1) 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). We find that the effect is statistically significant at the 1% level, and the coefficients on the different measures of originality are unchanged.[42]

Finally, as an alternative strategy to deal with the skewness of the Facebook shares variable distribution, we use a winsorized version of the variable at the 99th percentile. We then perform a linear estimation. The results are consistent with the ones we obtain when performing the log-linear estimation.[43]

5.2.2. Number of Twitter shares. As a measure of the returns of original news production, the number of times an article is shared on Facebook suffers from a number of caveats, in particular the fact that this number is partially filtered through the Facebook News Feed algorithm. While we cannot directly correct for this filtering, we show that our findings are robust to the use of the number of shares on Twitter. Columns (4)–(6) of Table 5 present the results of the estimation of equation (5.1) where the number of shares on Twitter is the dependent variable.

The results we obtain are consistent with the findings using the number of Facebook shares. On the one hand, social media audience increases with the number of original characters: an increase of 1,000 in the number of original characters leads to a 11.4% increase in the number of Tweets. If we instead consider the originality rate, a 50-percentage-point increase in the originality rate of an article leads to a 17.3% increase in the number of Tweets. Moreover, as before, both the

42. Supplementary Appendix Table E.5.
43. Supplementary Appendix Table E.6.

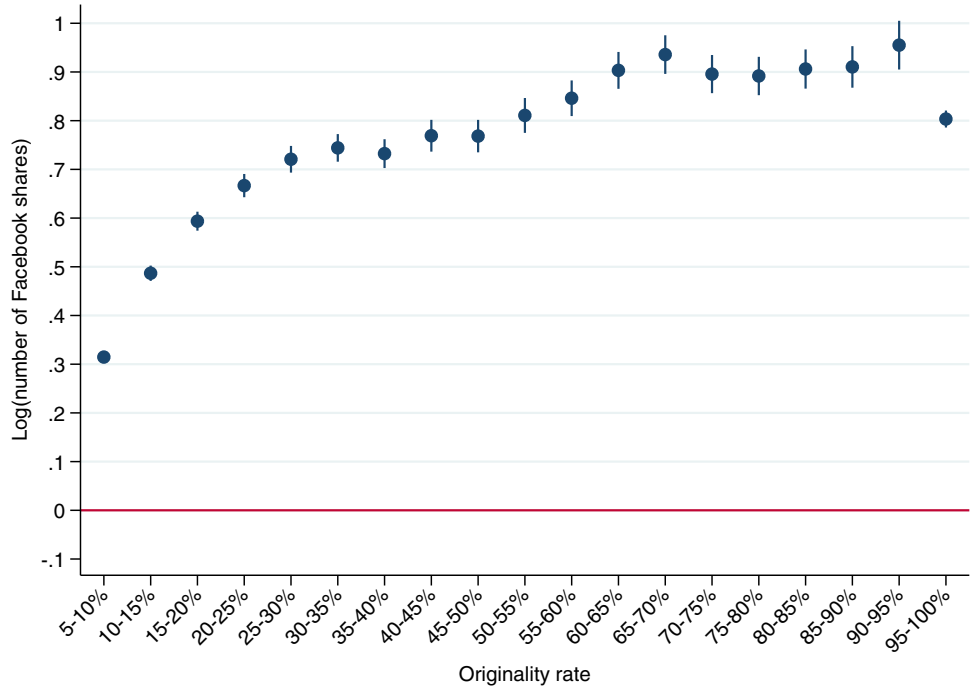


FIGURE 5

Facebook shares and originality 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.

publication time and the publication rank matter regarding reactivity. Taking 41 h to cover an event rather than writing about it from the beginning decreases the number of Tweets by 8.2%.

We have constructed the number of times an article is shared on Twitter variable as the sum of different measures (number of direct tweets, reweets, likes, etc.) By aggregating these correlated measures, we may overemphasize the extent to which a story is likable on Twitter. In the Supplementary Appendix, we show that the magnitude and statistical significance of the coefficients is unchanged if we instead consider the number of (direct) tweets independently as a dependent variable. For example, we find that a 50-percentage-point increase in the originality rate of an article increases the number of times it is tweeted by 13.3%.[44]

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

5.2.3. Predicted number of readers. Finally, in Columns (7)–(9) of Table 5, we present the results of the estimations when we use the number of times an article is viewed (using

44. Supplementary Appendix Table E.7.

the Facebook approach detailed above) rather than the number of shares on social media as a dependent variable. Although this measure is imperfect—it is a predicted measure of the number of readers based on the estimates we obtain from *Le Monde* data rather than the actual number of readers—it may be considered the more telling variable to estimate the returns to originality in terms of audience.

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

To summarize using various measures, we find strong evidence that readers are more likely to consume more original articles, which serves as a positive incentive for the media to produce original content. In light of our theoretical model, this finding is consistent with the poor quality of copying that we observe in the data, and with a relatively strong taste for originality (as compared to consumer loyalty).

5.2.4. Heterogeneity of the effects. We now investigate whether the returns to originality vary depending on the characteristics of the media outlets and of the events they cover. We consider different dimensions of heterogeneity: first, the competitiveness of the media environment; second, the extent to which the media outlets are copied by other outlets; and finally, the topic of the events (*e.g.* sport or economy) and their "general interest." Doing so allows us to improve our understanding of the mechanisms at play behind the positive returns to original news production and to better test some of the predictions of our theoretical framework.

Competitiveness of the media environment. We estimate equation (5.1) with an interacted "high competition" indicator variable equal to one for the media outlets that are in a "more competitive" media environment and to zero for those that are in a "less competitive" media environment. The competitiveness of the environment is measured with respect to the average number of other media outlets consumed by the readers who access a given media.[45] In the spirit of our simple theoretical framework, a highly competitive environment is an environment in which $\bar{u}$—the utility users derive from consuming a specific media—is low, while a less competitive environment is an environment where consumers' loyalty to certain media brands is high. Obviously, as we have highlighted, none of the media outlets is "in isolation" online, but it is nonetheless of interest to exploit the heterogeneity of their competitive environment.

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

45. See the discussion in Section 4.2 above and the Supplementary Appendix Figure F.12. The "low-competition" media outlets are TF1, BFM TV, France Television, *20 minutes*, Mediapart, *Le Monde*, Europe1, RMC, RTL, LCI, *Le Figaro*, France24, France Culture, France Info, France Inter, Metro, and Rue89. The "high-competition" media outlets are *Les Echos*, *Direct Matin*, *Courrier International*, I-TELE, *Liberation*, Slate, *La Croix*, *Marianne*, *L'Express*, *Le Point*, *Le Nouvel Obs*, and Atlantico.

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

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

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is the log of the number of times an article is shared on Facebook in Columns (1) and (2), the log of the number of times an article is shared on Twitter in Columns (3) and (4), and the log of the number of views per article in Columns (5) and (6). The number of views per article is computed by combining media-level information on the daily number of page views with article-level information on the number of times an article is shared on Facebook (as detailed in Section 5.3). Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. "High competition" is an indicator variable equal to one for the media outlets that are in a "more competitive" media environment and to zero for those that are in a "less competitive" media environment. The competitiveness of the environment is measured with respect to the average number of other media outlets consumed by the readers who access a given media (see the text for more details). All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters."

of an article leads to a 52.2% increase in the number of Facebook shares—than for the outlets that are in a less competitive environment (29.7% increase).

However, it should be noted that while we believe these results are of interest and serve to highlight the mechanisms at play, they should be interpreted with caution given the limits of the survey data we use to distinguish between "high-competition" and "low-competition" outlets.

Extent to which the media are copied. The second dimension of heterogeneity we consider is the extent to which the media outlets are copied by their competitors. To do so, we rely on the results of Section 3.3 where we have computed, for each of the media outlets in our sample, the average share of their content that was copied in 2013. Using the median, we split our sample into two groups and estimate equation (5.1) with an interaction term between the different explanatory

TABLE 7

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

	Facebook shares		Tweets		Number of views	
	(1)	(2)	(3)	(4)	(5)	(6)
Publication rank	−0.0005***	−0.0005***	−0.0002***	−0.0002***	−0.0005***	−0.0005***
	(0.0001)	(0.0001)	(0.0000)	(0.0000)	(0.0001)	(0.0001)
Publication rank *		−0.0000		0.0000		−0.0001*
Highly copied		(0.0000)		(0.0000)		(0.0001)
Reaction time	−0.0023***	−0.0019***	−0.0021***	−0.0018***	−0.0045***	−0.0044***
	(0.0006)	(0.0006)	(0.0003)	(0.0003)	(0.0007)	(0.0007)
Reaction time *		−0.0003*		−0.0003***		0.0001
Highly copied		(0.0002)		(0.0001)		(0.0002)
Originality rate	0.0068***	0.0083***	0.0032***	0.0052***	0.0074***	0.0087***
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0002)
Originality rate *		−0.0022***		−0.0030***		−0.0019***
Highly copied		(0.0002)		(0.0001)		(0.0002)
Length	0.0855***	0.0644***	0.0593***	0.0505***	0.0796***	0.0680***
	(0.0025)	(0.0030)	(0.0014)	(0.0017)	(0.0030)	(0.0041)
Length * Highly copied		0.0404***		0.0174***		0.0212***
		(0.0038)		(0.0022)		(0.0049)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes	Yes	Yes
Event FE	Yes	Yes	Yes	Yes	Yes	Yes
R^2	0.40	0.41	0.54	0.55	0.42	0.42
Observations	664,650	663,825	656,129	655,304	509,378	508,665
Clusters (event)	25,200	25,200	25,200	25,200	25,109	25,109

textitNotes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is the log of the number of times an article is shared on Facebook in Columns (1) and (2), the log of the number of times an article is shared on Twitter in Columns (3) and (4), and the log of the number of views per article in Columns (5) and (6). The number of views per article is computed by combining media-level information on the daily number of page views with article-level information on the number of times an article is shared on Facebook (as detailed in Section 5.3). Standard errors in parentheses are clustered by event. Models are estimated using OLS estimations. The unit of observation is an article. "Highly copied" is an indicator variable equal to 1 for the media outlets that have been highly copied in 2013, *i.e.*, whose share of the content that has been copied is higher than the median, and to 0 otherwise (see the text for more details). All the estimations include media outlets, date, and event fixed effects. Variables are described in more details in the text. "(thsd ch)" stands for "thousand characters."

variables and a "highly copied" indicator variable equal to one for the media outlets whose share of the content that has been copied is above the median (25.5%), and to zero otherwise.[46]

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

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

of views) than for the media outlets that suffer less from copying (49% increase in the number of views). This finding is also consistent with our simple theoretical framework.[47]

Furthermore, in the Supplementary Appendix, we estimate equation (5.1) adding to the vector $\mathbf{Z}'_{\mathbf{aedn}}$ an additional characteristic of the article a published by media n on date d and included in the event e, namely the share of its content that has been copied by other media outlets.[48] Evidently, this characteristic is hard to interpret given that not only the articles published first in an event—and the most original ones—tend to be the most copied, but also because the most copied articles may be the ones that are of higher "quality." The only proxy we have here for the "quality" of an article is its originality (originality rate or original content), and we probably miss out on other important dimensions. However, it is interesting to note that the share of the article that is copied is negatively correlated with our different measures of the article's audience once we control for the other characteristics of the article (although the effect is not statistically significant for the number of predicted readers).

Topic of the event. Finally, do the characteristics of the events, and in particular their topic, affect the originality premium? In Table 8, we estimate equation (5.1) separately for the different events depending on their topic (politics, economy, sport, etc.). We find that the returns to originality—as measured by the effect of an increase in the originality rate on the number of times an article is shared on Facebook—are higher for "Crime, law and justice" events (a 50-percentage-point increase in the originality rate of an article leads to a 49.2% increase in the number of Facebook shares) as well as for "Politics" events (45.5% increase), and lower for events about "Economy, business and finance" (31.6%) and for "Sport" events (27.1%). Moreover, the difference in the magnitude of the effects is statistically significant. In other words, topics that generate less attention on social media, such as sport and economy[49], also seem to have lower returns to originality. In light of our very simple theoretical framework, this heterogeneity in the returns to originality can be interpreted in terms of the "easiness to find scoops" for a given investment in newsgathering. One can indeed assume that finding a "Sport event" scoop (*e.g.* reporting the results of a soccer game) may be "easier" than finding a "Politics event" scoop (*e.g.* reporting a political scandal). Assuming that readers are more willing to switch to the website of the original news producer when they acknowledge the "rareness" of the scoop found, we show that the returns to originality in terms of audience are lower for relatively more easy-to-find events.

We obtain similar results if we investigate heterogeneity in the returns to originality depending on the "general interest" of the events, as proxied by the total number of shares received by all the articles in an event. We generate a "High general interest" indicator variable equal to 1 for the events whose total number of shares received is higher than the median (201), and to 0 otherwise. We find that while originality always matters, the returns to originality are higher for the events with greater general interest, and that the difference is statistically significant. Furthermore, this effect holds even within topics, *i.e.*, if we perform the estimation separately for the different events depending on their topic.[50]

47. In the Supplementary Appendix Table E.8, we perform the same analysis separately for low-competition and high-competition media outlets (as defined above). We obtain in both cases lower returns to originality for highly copied media outlets. Given that we do not observe the low- vs. high-competition status for all the media outlets in our sample, these results should be interpreted with care, however.

48. Supplementary Appendix Table E.9.

49. "Sport" and "Economy, business and finance" events indeed tend to generate fewer shares on Facebook than "Crime, law and justice" events (Supplementary Appendix Figure F.5).

50. See Supplementary Appendix Tables E.10 and E.11.

TABLE 8
Article-level analysis: number of Facebook shares (log-linear estimation), heterogeneity of the effects depending on the topic of the events

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

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

5.2.5. Short-run audience effects vs long-run reputation effects. Ultimately, how can one rationalize the positive relationship between originality and news consumption as proxied by the number of shares on social media? Our preferred explanation is that consumers favour originality, and that the quality of the copy is lower than that of the original. The evidence we present in this section is consistent with the predictions of our simple theoretical framework on copying and returns to originality. It is also consistent with the fact that, as highlighted by Boczkowski and Mitchelstein (2013), consumption choices are "often made at the story level" (p. 9). Hence, consumers willing to learn about a news event may decide to read the most original piece because they value its originality, or simply because this is the first article published within an event and so the first they have a chance to see.

Note moreover that, while until now we have only considered consumers' switching behaviour at the short-run level (using article-level estimations and media-outlet-event date variations), it is also possible that longer-run reputation effects allow original producers to recoup an even larger share of the audience. In the Supplementary Appendix, we estimate the correlation between the average daily number of unique visitors (we compute this average over the year 2013) and the average content produced. We find that audience is positively correlated (with a statistically significant relationship) with the quantity of content classified in events, with the originality of the content produced, and with the number of breaking news stories.[51] There is no statistically significant correlation between the quantity of content not classified in events and the number of unique visitors, however. We also perform a similar estimation but using the daily-level variations in audience and controlling for media and date fixed effects (the unit of observation is a media outlet-date and standard errors are clustered at the media outlet level). We find that the only characteristic of the content produced on a daily basis by a media outlet that has a statistically significant impact on the daily variations in its audience is the originality rate. The magnitude of the effect is small; however, a 50-percentage-point increase in the originality rate of the content published by a media outlet on a given date is associated with a 2.5% increase in its number of daily visitors.[52] These results should be interpreted carefully though, given that these daily-level variations in the production of information and in the audience share of each media outlet allow us to estimate only correlations, not to identify causal effects.

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

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

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

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

Figure 6 presents the results. First, for the sake of comparison, we compute the share of original content in the dataset. This share is equal to 32.5%.[53] 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.

51. Supplementary Appendix Table E.12.
52. Supplementary Appendix Table E.13.
53. We only consider here the articles for which we have audience data, and in particular we drop the AFP and Reuters. If we were to consider all the articles, then the share of original content in the dataset is equal to 32.6%. The difference with the average originality rate of the articles comes from the fact that articles are of different length.

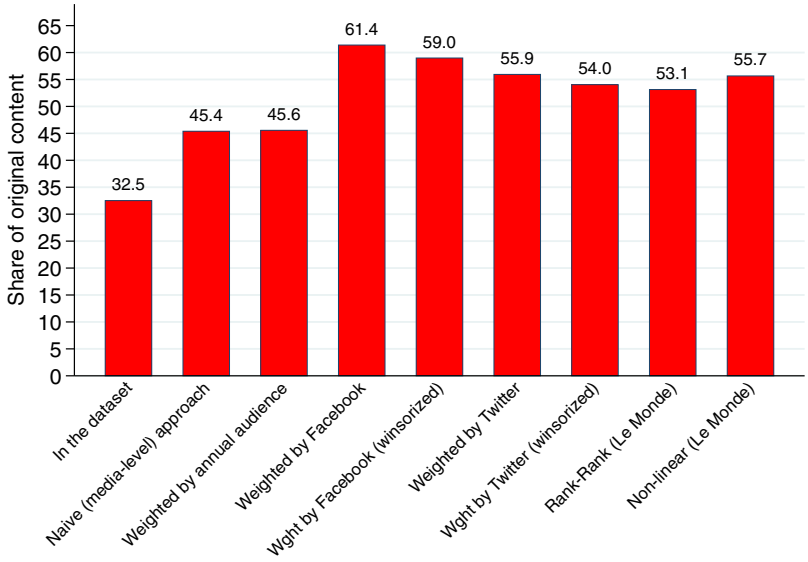


FIGURE 6

The audience-weighted share of original content

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

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

5.4. *The returns to originality*

The key question this article attempts to address is the following: given the limited legal protection of intellectual property rights in news production, what is the extent of copying online and what are the incentives to produce original content? We document the severity of copying, both on the extensive and on the intensive margins, and show that online readers are more likely to read articles from the website of the media outlets with more original content, thereby rewarding originality and mitigating the newsgathering incentive problem raised by copying (at least in part). Ideally,

we would like to estimate what fraction of the returns to original news content production is appropriated by the original news producers thanks to consumer behaviour. Although our data sources do not allow us to fully address this question, our results can be used to provide some orders of magnitude.

Our basic result is that only 32.5% of the online content is original. Every time an original piece of content is published on the Internet, it is actually published three times: once by the original producer, and twice by media outlets who simply copy-and-paste this original content. In the event that Internet audience was distributed randomly on the different websites and on the original and copied version of the articles, this result would imply that the original producer captures only one-third 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), and that the copiers capture up to two-thirds of the returns.[54]

However, as we have just shown, audience is not randomly distributed on the Internet. First, if we weight content by media-level daily audience shares (using the naive approach), we find that original content represents 45.4% of online news consumption. This may reflect the fact that media outlets with a larger fraction of original content tend to attract a higher audience, possibly because they have a stronger reputation and/or because on days when more original content is published there is also a higher audience. This may also be partly due to the way news aggregators work. For example, 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.[55] To further investigate this issue, we weight the content of each article by the average annual audience of the media outlet in which it was published (assuming all the articles published on the website of an outlet in 2013 received the same number of views). When we do so, we find that the original content represents 45.6% of the online news consumption, *i.e.* almost the same share as when we use the naive approach (weighting the content by media-level daily audience shares). This shows that the daily-level audience hardly varies on average with daily-level average originality. This result—which is also consistent with the very small magnitude of the impact of original copy on audience we find when we only consider daily-level variations in audience and control for media and date fixed effects—suggests that media-level reputation effects play an important role.

Most importantly, if we weight content by media-level audience shares and article-level Facebook shares or number of Tweets, we find that the original content represents between 53.1% and 61.4% of online news consumption, depending on the approach chosen. That is, within a given media outlet, the articles that get more views (as approximated by the number of shares on social media) are those with more original content. In effect, thanks to the combination of media-level reputation effects and consumers' preference for originality at the article level, the audience share of original content jumps from 32.5% to between 53.1% and 61.4%.[56]

54. These figures rely on all the copied content in our dataset (see below for similar estimations without the content reproduced from the news agencies). The objective of this section is to document the relative importance of original news production and of original news consumption. As detailed in Section 3.2, copied content can come from a number of different sources and our estimation of the "magnitude" of copying is imperfect given that we cannot measure the relative editorial importance of the parts that are copied and of the parts that are not. But, with these limitations in mind, the computation of the audience-weighted share of original content improves our understanding of the incentives to produce original content.

55. 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*, 21 February 2013.

56. 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 (Supplementary Appendix Figure F.15).

As a robustness check on this estimation of the returns to originality, we perform the same analysis but after having dropped all the content copied from the news agencies. More precisely, we define the *total content* of an article as its content minus the content reproduced from the news agencies, and the *original content* of an article as its content minus the content reproduced from the news agencies and the content reproduced from other media outlets (excluding itself). Doing so, we find that on average documents are 1,311 characters long, and that 69.3% of the online content is original (higher originality is not surprising given that media outlets mainly rely on content copied from the news agencies). The audience-weighted share of original content is equal to 79.7% when we use the naive approach, and to between 81.9% and 84.4% when we allocate the number of views as a function of the number of shares on social media.[57] Hence, despite a lower reliance on copying, media-level reputation and consumers' preference for originality still lead to a consumption of original content that is higher than its relative production.

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

6. ROBUSTNESS CHECKS AND DISCUSSION

In this section, we perform a number of robustness checks and discuss the external validity of our main findings.

6.1. *Relaxing the "10 documents condition"*

Not surprisingly, the total number of media events identified by the algorithm strongly increases when we relax the ten documents condition. We obtain a total number of 113,959 news events. The events are on average much shorter and comprise a lower number of documents. Regarding the documents classified in the events, their originality rate is equal on average to 42.6%. The distribution of the originality rate is bimodal, with one peak for the articles with less than 1% original content and another peak for the 100% original articles. 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.

If we turn to the ratio of original content over the total content, it is equal to nearly 39%. Following the same empirical strategy as before, we show that the audience-weighted share of original content (which varies between 57.1% and 66% depending on the specification) is much higher than the production of original content. Finally, we re-estimate equation (5.1) and show that the order of magnitude of the estimated coefficients is unchanged. For example, we find that an increase of 1,000 in the number of original characters leads to a 21.2% increase in the number of Facebook shares, a 10.6% increase in the number of Tweets, and a 22.4% increase

57. Supplementary Appendix Figure F.16.

in the predicted number of views. Hence, the main findings of this article do not depend on the threshold we impose regarding the number of articles to define an event.[58]

6.2. *Alternative event detection algorithms*

The event detection algorithm—which we have developed to identify the media events—is a key element of this article. The algorithm is composed of two main parts: the clustering algorithm and the semantic features for the text representation. In the past few years, the Natural Language Processing (NLP) research field has made great progress in several tasks by using new text representation schemes that better model the language and thus the semantic.[59] These new text representations have been used to replace the standard TF-IDF scheme (which we use in the event detection algorithm described in Section 2.2) in several NLP tasks and have brought significant improvements in terms of the performances of the algorithm.

Mazoyer *et al.* (2019) explore the potential benefits of these new word embedding models for the Topic Detection and Tracking task. In particular, they test the accuracy of the approach we use in this article against Word2Vec and Doc2Vec-based methods, using the dataset of media events we have created manually from our 2013 French corpus. First, they find that Doc2vec has much lower performances than the TF-IDF scheme; we thus decide to abandon this approach. Second, they show that the best Word2Vec document representation is obtained by using the TF-IDF weighting of word vectors instead of a simple mean. Third, they show that even the TF-IDF-weighted Word2Vec method does not perform better than the simple TF-IDF representation. Consequently, in our core specification, we use the TF-IDF scheme. As an additional robustness check, we have checked that our main findings are robust to using the TF-IDF-weighted Word2Vec approach.[60]

6.3. *External validity*

The results presented in this article are based on French data for the year 2013. Hence, one final question is whether we should expect the patterns we have uncovered in the case of 2013 France to be repeated in other contexts. First, should these patterns hold in other countries? And second, should they still hold today? There are good reasons to think this could be the case. First, while the French media market certainly presents specific features, it is by and large very similar to other Western media markets, whether we consider Internet penetration (87%, like Italy and Spain and only slightly below Belgium—88%—and Germany—90%), the use of social media for news (36%, compared to 31% for Germany and 39% for the U.K.), or the proportion of the population who paid for online news (11%, like in Spain, slightly above Germany or Canada—8%—but below Italy—12%) (Reuters Institute, 2018). In France, like in other Western media markets, many publishers offer online news for free and largely rely on advertising. Moreover France, like the U.S., has an international news agency, the AFP, which is the third leading agency in the world after Reuters and Associated Press. From this point of view, the French market is more similar to the U.S. market than the Spanish, Italian, or German markets. Therefore, overall, we believe the patterns we uncover regarding the propagation of online information, the importance of copying and the valuation of originality using French data would hold in other contexts.

58. See the Supplementary Appendix Section G for the associated tables and figures and Cagé *et al.* (2017, pp. 48–50) for a more detailed discussion.

59. For example, Word2Vec (Mikolov *et al.*, 2013), Doc2Vec (Le and Mikolov, 2014), and Glove (Pennington *et al.*, 2014).

60. See the Supplementary Appendix Section H for the associates tables and figures, and Cagé *et al.* (2017, pp. 51–52) for a detailed discussion.

Obviously, we are also well aware of the fact that the digital news market has evolved since 2013. In particular, pay models are becoming an important part of the business of digital news, while they were just in their infancy in 2013. In most markets, however, there are still many publishers who offer online news for free. As a consequence, digital advertising revenues are still the main source of revenue for the media online. In 2018, only 16% of consumers paid for online news content in the U.S.[61], 12% in Italy, 11% in France, and 7% in the U.K. (Reuters Institute, 2018). Hence, even if new paywall systems have developed since 2013 and may further develop in the future, it is important to highlight that digital advertising remains a critical source of revenue. Furthermore, the growing importance of the paywall systems, while modifying media outlets' sources of revenues, should not significantly modify the impact of copying on newsgathering incentives.[62]

Note also that, while in this article we estimate the economic returns to originality in terms of audience, we know that the objective function of the media certainly includes other dimensions than audience size alone. On the one hand, public-service broadcasters have public service obligations. On the other hand, a number of media outlets may have political motives entering their production function. Media owners may derive utility from influencing the political tastes of their readers and find the verbatim copying of their content useful for spreading their editorial line across different outlets. However, even public-service broadcasters in France depend on advertising funding, and audiences also enter the objective function of politically driven media owners. Hence, we think that our approach, despite its focus on the monetary profits of the different media outlets, is relevant to investigate the returns to originality.[63]

Lastly, it should be noted that even though the news market has changed online in recent years, the overall structure of the French media market has not changed radically since 2013. First, according to Reuters, there was no change between 2013 and 2018 in the use of the Internet as a source of news (68%) (Reuters Institute, 2013, 2018). Second, the French media landscape has remained fairly stable, in particular if we consider the main media outlets in terms of audience.

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

7. CONCLUSION

This article documents the extent of copying online and estimates the returns to originality in online news production. It builds a unique dataset combining the online production of information of the French news media during the year 2013 with micro audience data, and develops a number of algorithms which could be of future use to other researchers studying media content.

First, we investigate the speed of news dissemination and distinguish between original information production and copy-and-paste. We find that only 32.5% of online news content is original. Even if we focus only on external copying and exclude content copied from the news agencies, we show that, on the extensive margin, nearly two-thirds of the articles in events contain at least some copying. Furthermore, the copied parts are arguably the most valuable parts of the

61. This relatively high share in the U.S. in 2018 comes from the so-called "Trump Bump."

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

63. See Cagé *et al.* (2017, pp. 31–35) for a detailed discussion.

copied articles, and may be a substitute for the original content. This scale of copying online might be related to the observed drop in media companies' employment of journalists in recent years, raising growing concerns about the industry's ability to produce high-quality information (see *e.g.* Angelucci and Cagé, 2019)[64]. In the event that online audience was distributed randomly and revenues were proportional to audience, our results would imply that the original news producers only capture a small share of the economic returns to the original news content they provide.

Next, this article seeks to better understand why, in spite of massive online copying, there is still original news production in online media. Using article-level variations and media-level daily audience combined with article-level social media statistics, we find that readers are more likely to consume news on the website of the original producers, thereby mitigating the newsgathering incentive problem raised by copying. We show that long-term reputation mechanisms and the short-run behaviour of Internet viewers—in particular their preference for original content at the article level—make it possible to mitigate a significant part of the plagiarism problem. We indeed find that original content represents up to 61.4% of online news consumption, *i.e.*, much more than its share of online news production.

Of course, greater intellectual property protection could also play a role in reducing copyright violation and raising the incentives for original news production, and we certainly do not mean to downplay the extent of this problem. In 2010, the Federal Trade Commission (FTC) in the U.S. issued a discussion paper outlining the enactment of "Federal Hot News Legislation" as a proposal aimed at reinventing journalism and addressing newspapers' revenue problems. Whether or not a stricter enforcement of copyright laws regarding online news media is feasible or desirable is very much an open issue at this stage. It is possible that new policy tools need to be developed, including a more favourable legal and fiscal status for investigative journalism (Cagé, 2016). In any case, our results suggest that in order to effectively address these issues, it is important to study how viewers react to the newsgathering investment strategies of media outlets and how much they care about originality, quality, and reputation.

Finally, we think that our results—as well as the algorithms we developed for this study—may help to improve our future understanding of "where people get their news," combining consumption and production data. Prat (2018) and Kennedy and Prat (2019) have documented news consumption across platforms; a complementary strategy to estimate media power would be to weight the influence of media companies by their supply of original news and how much other companies rely on that news. It would also be of interest to investigate how news production has evolved over time, and in particular to study the heterogeneous impact of the Internet on media investment in in-depth reporting. More research is still needed, but we hope this article will inform the debate on concentration in media power.

Acknowledgments. We are grateful to the Editor (Nicola Gennaioli) and to five anonymous referees for insightful comments that substantially improved the article. We gratefully acknowledge the many helpful comments and suggestions from Charles Angelucci, Yasmine Bekkouche, Filipe Campante, Lucien Castex, Etienne Fize, Matthew Gentzkow, Sergei Guriev, Emeric Henry, Ali Hortacsu, Elise Huillery, Laurent Joyeux, Benjamin Marx, Petra Moser, Elisa Mougin, Aurélie Ouss, Arnaud Philippe, Thomas Piketty, Andrea Prat, Valeria Rueda, Agnès Saulnier, and Katia Zhuravskaya. We are grateful to participants at the Barcelona GSE Summer Forum, the Big Data for Media Analysis Conference, the CEPR Public Economics Annual Symposium, the Economics of Media and Communication Conference, the IEA World Congress, the NBER Political Economy Meeting, the NET Institute Conference on Network Economics, and SIOE 2017, and to seminar participants at Banque de France, the Paris School of Economics, Sciences Po Paris, the Toulouse School of Economics, and the University Carlos III of Madrid. We thank Jérôme Fenoglio and Pierre Buffet for sharing *Le Monde*'s data on the number of views per article, Andrew Montalenti for giving us access to Parsely's data, and Richard Fletcher for providing us the survey data from the 2013 and 2018 Reuters' Digital News Reports. Edgard Dewitte, Anais

64. Note that the causality could go both ways: a smaller number of journalists can in turn contribute to more copying-and-pasting.

Galdin, Béatrice Mazoyer, Lucile Rogissart and Jeanne Sorin provided outstanding research assistance. This research was generously supported by the NET Institute, the Paris School of Economics, the Banque de France, Sciences Po's Scientific advisory board (SAB), and the French National Research Agency (ANR-17-CE26-0004). Since November 2015, Julia Cagé has been a Board member of the Agence France Presse; this paper does not reflect the views of the AFP and responsibility for the results presented lies entirely with the authors.

Supplementary Data

Supplementary data are available at *Review of Economic Studies* online.

REFERENCES

ALLCOTT, H. and GENTZKOW, M. (2017), "Social Media and Fake News in the 2016 Election", *Journal of Economic Perspectives* **31**, 211–236.
ANDERSON, S. P. (2012), "Advertising on the Internet", in Peitz, M. and Waldfogel, J. (eds.) *The Oxford Handbook of the Digital Economy* (Oxford, UK: Oxford University Press).
ANGELUCCI, C. and CAGÉ, J. (2019), "Newspapers in Times of Low Advertising Revenues", *American Economic Journal: Microeconomics* **11**, 319–364.
ATHEY, S., CALVANO, E. and GANS, J. S. (2018), "The Impact of Consumer Multi-homing on Advertising Markets and Media Competition", *Management Science* **64**, 1574–1590.
BAE, S. H. and CHOI, J. P. (2006), "A Model of Piracy," *Information Economics and Policy* **18**, 303–320.
BIASI, B. and MOSER, P. "Effects of Copyrights on Science: Evidence from the WWII Book Replication Program" (Working Paper 2015).
BOCZKOWSKI, P. J. (2010), *News at Work: Imitation in an Age of Information Abundance* (Chicago, IL: University of Chicago Press).
BOCZKOWSKI, P. J. and MITCHELSTEIN, E. (2013), *The News Gap: When the Information Preferences of the Media and the Public Diverge* The News Gap (Cambridge, MA: MIT Press).
CAGÉ, J. (2016), *Saving the Media* (Cambridge, MA: The Belknap Press of Harvard University Press).
CAGÉ, J. (2020), "Media Competition, Information Provision and Political Participation: Evidence from French Local Newspapers and Elections, 1944-2014," *Journal of Public Economics*, forthcoming.
CAGÉ, J., N. HERVÉ, and VIAUD, M.-L. (2017), "The Production of Information in an Online World: Is Copy Right?" (Discussion Papers 12066, CEPR).
CHIOU, L. and TUCKER, C. (2017), "Content Aggregation by Platforms: The Case of the News Media," *Journal of Economics & Management Strategy* **26**, 782–805.
EISENSEE, T. and STRÖMBERG, D. (2007), "News Droughts, News Floods, and U. S. Disaster Relief," *The Quarterly Journal of Economics* **122**, 693–728.
FRANCESCHELLI, I. (2011) "When the Ink is Gone: The Transition from Print to Online Editions" (Technical Report, Northwestern University).
GAVAZZA, A., NARDOTTO, M. and VALLETTI, T. (2019), "Internet and Politics: Evidence from U.K. Local Elections and Local Government Policies", *The Review of Economic Studies* **86**, 2092–2135.
GENTZKOW, M. (2006), "Television and Voter Turnout," *Quarterly Journal of Economics* **121**, 931–972.
GENTZKOW, M. (2007), "Valuing New Goods in a Model with Complementarity: Online Newspapers", *American Economic Review* 97, 713–744.
GENTZKOW, M. and SHAPIRO, J. M. (2008) "Competition and Truth in the Market for News", *Journal of Economic Perspectives* **22**, 133–154.
GEORGE, L. M. (2008) "The Internet and the Market for Daily Newspapers", *The B.E. Journal of Economic Analysis & Policy* **8**, 1–33.
GEORGE, L. M. and WALDFOGEL, J. (2006), "The New York Times and the Market for Local Newspapers", *American Economic Review* **96**, 435–447.
GINSBURG, J. C. (2016), "Overview of Copyright Law", in Dreyfuss, R. and Pila, J. (eds) *Oxford Handbook of Intellectual Property* (Oxford: Oxford University Press).
GIORCELLI, M. and MOSER, P. (2015), "Copyright and Creativity: Evidence from Italian Operas" (Working Paper).
KENNEDY, P. J. and PRAT, A. (2019), "Where Do People Get Their News?", *Economic Policy* **34**, 5–47.
LE, Q. and MIKOLOV, T. (2014), "Distributed Representations of Sentences and Documents", in *Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML'14*, Vol. 32 (JMLR.org) 1188–1196.
LI, X., MACGARVIE, M. and MOSER, P. (2018), "Dead poets' property - How Does Copyright Influence Price?", *The RAND Journal of Economics* **49**, 181–205.
MACGARVIE, M. and MOSER, P. (2014), "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.) 357–379.
MAZOYER, B., HERVÉ, N., EVRARD, M. *et al.* (2019), "Word Embeddings for Topic Detection and Tracking" (Technical Report, INA Working Paper).

MIKOLOV, T., I. SUTSKEVER, CHEN, K. *et al.* (2013), "Distributed Representations of Words and Phrases and their Compositionality", in *Advances in Neural Information Processing Systems* 3111–3119.
NAGARAJ, A. (2018), "Does Copyright Affect Reuse? Evidence from Google Books and Wikipedia," *Management Science* **64**, 3091–3107.
NICHOLLS, T., SHABBIR, N. and NIELSEN, R. K. "Digital-Born News Media in Europe" (Techreport, Reuters Institute for the Study of Journalism 2016).
OBERHOLZER‐GEE, F. and STRUMPF, K. (2007), "The Effect of File Sharing on Record Sales: An Empirical Analysis," *Journal of Political Economy* **115**, 1–42.
PEITZ, M. and REISINGER, M. (2016), "Chapter 10 - The Economics of Internet Media," in Simon, J.W., Anderson, P. and Strömberg, D. (eds) *Handbook of Media Economics*, Vol. 1 of *Handbook of Media Economics* (North-Holland) 445–530.
PEITZ, M. and WAELBROECK, P. (2006), "Piracy of Digital Products: A Critical Review of the Theoretical Literature," *Information Economics and Policy* **18**, 449–476.
PEITZ, M. and WAELBROECK, P. (2006), "Why the Music Industry May Gain from Free Downloading The Role of Sampling," *International Journal of Industrial Organization* **24**, 907–913.
PENNINGTON, J., SOCHER, R. and MANNING, C. D. (2014), "Glove: Global Vectors for Word Representation", in *EMNLP*, Vol. 14, 1532–1543.
PEW RESEARCH CENTER (2016), "State of the News Media Report 2016" (Report 2016).
PICONE, I., COURTOIS, C. and PAULUSSEN, S. (2015) "When News is Everywhere," *Journalism Practice* **9**, 35–49.
PRAT, A. (2018), "Media Power," *Journal of Political Economy* **126**, 1747–1783.
REIMERS, I. (2019), "Copyright and Generic Entry in Book Publishing," *American Economic Journal: Microeconomics* **11**, 257–284.
REUTERS INSTITUTE (2013), "Digital News Report 2013" (Annual Report 2013).
REUTERS INSTITUTE (2017), "Digital News Report 2017" (Annual Report 2017).
REUTERS INSTITUTE (2018), "Digital News Report 2018" (Annual Report 2018).
ROB, R. and WALDFOGEL, J. (2006), "Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students," *Journal of Law and Economics* **49**, 29–62.
SALAMI, A. and SEAMANS, R. (2014), "The Effect of the Internet on Newspaper Readability" (Working Papers 14-13, NET Institute).
SCHUDSON, M. (1981), *Discovering the News: A Social History of American Newspapers* (New York: Basic Books).
SEN, A. and YILDIRIM, P. (2015), "Clicks and Editorial Decisions: How Does Popularity Shape Online News Coverage?" (Working Paper).
SNYDER, J. M. and STROMBERG, D. (2010), "Press Coverage and Political Accountability" *Journal of Political Economy* **118**, 355–408.
SWART, J., PETERS, C. and BROERSMA, M. (2017) "Navigating Cross-media News Use," *Journalism Studies* **18**, 1343–1362.
WALDFOGEL, J. (2012), "Copyright Protection, Technological Change, and the Quality of New Products: Evidence from Recorded Music since Napster," *The Journal of Law & Economics* **55**, 715–740.
WALDFOGEL, J. (2015), "Digitization and the Quality of New Media Products: The Case of Music", in *Economic Analysis of the Digital Economy* (University of Chicago Press) 407–442.
YUAN, E. (2011), "News Consumption Across Multiple Media Platforms," *Information, Communication & Society* **14**, 998–1016.