The Production of Information in an Online World*†

Julia Cagé[‡1], Nicolas Hervé[§2], and Marie-Luce Viaud[¶2]

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

September 30, 2015

Abstract

Information is costly to produce but cheap to reproduce. Who are the main providers of original news in the online world, and are they rewarded for this? What are the benefits of breaking out a story, and how does information propagate? This paper addresses these issues by exploiting a unique dataset including all online content produced by general information media outlets in France during year 2013. Tracking every piece of content produced by these outlets, we develop a topic detection algorithm to construct the set of news stories. We study the timeline of each story and distinguish between original reporting and copy-and-paste. We then merge this content data with data on investment in news gathering and daily audience to investigate the costs and benefits of information production. This paper offers a typology of online media outlets and associated business models. We first highlight the specific role played by news agencies. AFP has the largest news desk and is the main provider of original information, reflecting the use of an adequate copyright system. We then find a quasi-linear relationship between the number of journalists, the quantity of original news production, and online audience. This positive correlation hold for all the media outlets independently of their offline support; hence the relevance of a transmedia approach. However online audience does not translate into significant revenues. This illustrates the need to develop new paywall or copyright models.

Keywords: Internet, information production, paywall, copyright, online audience
JEL No: L11, L15, L82, L86.

*We are particularly indebted to Lucien Castex, Laurent Joyeux and Agnès Saulnier for numerous comments and discussions. We also gratefully acknowledge the many helpful comments and suggestions from Elise Huillery and Thomas Piketty. We are grateful to participants at the "Big Data for Media Analysis" conference. This research was generously supported by the NET Institute (www.NETinst.org) and the Paris School of Economics. All errors remain our own.

†An online Appendix with additional empirical material is available here. A Technical Annex providing additional details on the tracking of information across transmedia sources as part of the OTMedia project is available here. This Annex will be updated on a regular basis.

‡Corresponding author. julia [dot] cage [at] sciencespo [dot] fr.
§nherve [at] ina [dot] fr.
¶mlviaud [at] ina [dot] fr.

1 Introduction

Information is costly to produce but cheap to reproduce. Who are the main providers of original news in an online world? What are the benefits of breaking out a story? And how does information change as it propagates? This paper aims at tackling these questions.

The production of information has always being characterized by large fixed costs (in particular the size of the newsroom) and increasing returns to scale (see e.g. Cagé, 2014). Newspapers have been willing to bear such a fixed cost in order to reap a profit from the original news content they provide. As highlighted by Gentzkow and Shapiro (2008), the incentive to beat competitors to a story has driven investments in news gathering since newspapers' earliest days: on a day where a newspaper had a big story and its competitors did not, demand was higher for the breaking news media (see also Schudson, 1981). In today's online world, utilizing other people's work has became easy and instantaneous, however.[1] This makes it extremely difficult for news content providers to protect and distinguish their content, and reap a profit from it where such profit is due.

In this paper, we examine a large set of French general information media outlets (including newspapers, television channels, radio stations and a news agency) and track every piece of content these outlets produced online in 2013. Using these documents, 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.[2] We then study the timeline of each story. In particular, for each story, we determine first the media that breaks out the story, and then study the propagation of this story, second-by-second. Beyond determining the first mover, we are indeed interested in how much different media outlets contribute to information production with respect to the story. We develop a plagiarism detection algorithm in order to quantify the copy rate between an article and all the articles previously published within the event. Our outcomes of interest are thus: (i) the probability of breaking out a news story; (ii) the reactivity of the outlet; and (iii) for a given publication time, the addition of original reporting.

We establish a number of descriptive facts on the propagation of information online. First, the dissemination of news is very fast. On average, it takes two hours for an information published by a media oulet to be published on the website of another oulet; but less than 45 minutes in half of the cases, of which less than 5 minutes in 25% of the cases. This very high reactivity comes with the use of copy-and-paste, however. According to our lower bound estimation, half of online information production is copy-and-paste. Most outlets simply echoe others work without adding original reporting. Moreover, a number of them do not obey the formal procedures for citing and crediting.

[1] While print editions have simultaneous daily updates, online editions can indeed be updated anytime.

[2] The system itself creates the clusters. We describe it in more details below.

We then merge this content data with data on the characteristics of the different media outlets. We collect information on the investments made by the outlets, in particular the size of the newsroom and the total wage bill. This allows us to investigate the quantity effects of investments in news gathering on the production of information. We offer a typology of online media outlets and associated business models.

We first highlight the specific role played by news agencies and syndicated news production. The French news agency, Agence France Presse (AFP), has the largest news desk and is the main provider of original information. It initiates one third of the news stories in 2013. This may reflect the use of an adequate copyright system: AFP is indeed the only actor which gets paid for the use of its content.

We then find a quasi-linear relationship between the number of journalists and the quantity of original news production (the probability of breaking a news and the provision of original information). A one percent increase in the size of the newsroom increases the probability of breaking out a news story by one percent. This positive correlation hold for all the media outlets independently of their offline support (newspaper, radio, or television). Radio and television stations – some of them publicly funded – have indeed invested massively in online news production. Their newsroom is of similar size than the one of a newspaper, and they behave on the Internet the same way than traditional print media outlets. Moreover, they compete online with the websites of newspapers. Future debates on media financing must therefore take a transmedia approach.

What are the benefits of breaking out a story? We investigate to which extent getting scoops benefits media in terms of audience on the web, circulation of their print edition for newspapers, and advertising revenues. We collect audience data at the daily level. Using a media reference detection algorithm, we also study to which extent providing original content can be used as a brand building strategy (which may be the case if a media is referred to as the source of the information). Reputation can indeed provide one way to understand why firms invest in information production (Gentzkow and Shapiro, 2008). We show that the relationship between the number of journalists and the online audience is also quasi-linear. However online audience does not translate into significant revenues.

Our findings have important implications for the financing of the media and their business models. As we noted above, information is costly to produce but cheap to reproduce. There is a "free rider" issue: the rapid spillover of information may lead to a situation where no firms would invest in gathering information. Hence how to encourage media outlets to produce original news? We finally explore the relevance of introducing copyright laws for news property and the need to develop new paywall models.[3] In the absence of such incentives, we may assist

[3]A paywall refers to any type of digital mechanism that separates free content from paid content on a website (Chiou and Tucker, 2013).

to an overall decrease in the production of information.

Literature review This paper contributes to three strands of literature in economics: the impact of the Internet on news coverage, intellectual property and copyright online, and the production and consumption of information in today's changing media landscape. It also contributes to the computer science literature on topic detection and on the structure of Internet diffusions.

Using micro data for newspapers in Washington, DC, Gentzkow (2007) studies competition between print and online newspapers: he estimates the relationship between the print and online papers in demand.[4] Franceschelli (2011) has been the first to assess empirically the impact of the Internet on news coverage. Using a dataset that includes every article published by the two main Argentinean newspapers (*Clarin* and *La Nacion*) in their online and print editions, he reconstructs the typical timeline of a news story in a print and in an online world.[5] Compared to this previous work, our contribution is threefold. First, we construct the timeline of the entire set of news stories using a sample including nearly all the French general information media outlets, rather than two newspapers. Second, while in order to identify the different news stories, Franceschelli (2011) relies restrictively on the mention of proper nouns, the algorithm we develop and run relies on word frequency without any restriction. Third and most importantly, we provide the first analysis of both the costs and benefits for news providers to provide original news content. Our paper is a unique attempt at trying to understand who is producing news, the character of what is produced and the benefits of news production.

We find that the rapid spillover of information may lead to a situation where media outlets do not have economic incentives to produce original reporting. This raises the question of the relevance of introducing copyright laws for news property, and relates our work to the research on intellectual property and copyright online. Giorcelli and Moser (2015) exploit historical variation in the adoption of copyright laws within the 19th century Italy to examine the effects of copyrights on creativity. They find that the adoption of copyrights led to a significant increase in the number of both new operas premiered and high-quality operas. While most of the literature has centered on digitization and piracy within the music industry (Rob and Waldfogel, 2006; OberholzerGee and Strumpf, 2007) or on the use of trademarks, Chiou and Tucker (2011) focus on the reproduction of content for information.[6] They exploit

[4] On the effect of the Internet on the demand for traditional media, see also George (2008). Salami and Seamans (2014) study the effect of the Internet on newspaper content, and in particular newspaper readability.

[5] Franceschelli (2011) investigates in particular how fast would news be delivered to readers in a print editions only world relatively to an online editions only world. He finds that, due to reduced reporter staff, online editions are slower at genuinely discovering news stories, but that nonetheless, thanks to continuous updates and rewriting, in a world populated only with online editions, the last reader to learn about a story would do it faster than the first reader to do it in a world populated only with print editions.

[6] For an assessment of the impact of copyright laws on the magazine industry in America during the 18th

a contract dispute that led a major aggregator to remove content from a content provider to quantify the impact on content aggregation on users' search for information.[7] They argue that producers of primary content may actually benefit from relaxing their restrictions on copyright and by allowing others to disseminate their content. Their empirical analysis relies on the very specific case of an aggregator, however. Aggregators only display small extracts of information, so that aggregator users visit content websites after visiting an aggregator. On the contrary, we show that half of online production is copy-and-paste. One media outlet can be seen as a perfect substitute for another outlet. Our paper thus investigates to which extent providing original news content leads to benefits for media outlets. According to our findings, the rapid spillover of information leads to a "free rider" issue. Hence the need to encourage media outlets to produce original news.[8]

Our work is more broadly related to the literature on citizens information in the new media landscape (see e.g. Popkin, 2007; Prior, 2005, 2007). It also relates to the literature on the production of information (see e.g. Cagé, 2014, on media competition and the provision of information). The way in which the Internet may affect political information patterns is an important question, with consequences in terms of political participation and the accountability of governments. By studying the production of information in an online world, we hope to inform this debate.

Finally, our work relates to the computer science literature on topic detection. The goal of topic detection is to organize a constantly arriving stream of news articles by the events they discuss (see e.g. Allan et al., 2005). In this paper, we develop a topic detection algorithm with state of the art performances to determine the set of news stories. By constructing the timeline of these stories, we also contribute to a recent strand of the literature that investigate the structure of Internet diffusions (Golub et al., 2010). But while the focus of this literature is on the large-scale diffusions of information like chain letters (Liben-Nowell and Kleinberg, 2008), we focus on the propagation of news stories. This has political economy implications absent the existing literature on the structure of Internet diffusions.

---

and 19th centuries, see Haveman and Kluttz (2014).

[7] More specifically, they exploit a contract dispute between Google News and the Associated Press as a discontinuous shift in the provision of copyrighted content by an aggregator, using Yahoo! News as a control group. Online aggregators assert that their practice is protected by copyright law because they only display small extracts of information and often this information is factual. Producers of content challenge this assertion because they fear that consumers may use these extracts of content as a substitute for accessing and reading the full content. Are consumers using aggregators to reduce search costs and terminate their search for content that they would already seek, or are consumers using aggregators to seek new content that they would not otherwise obtain? Chiou and Tucker (2011) show that users do not view an aggregator as a perfect substitute for copyrighted content: when users encounter content summarized by an aggregator, they are more likely to be provoked to seek additional sources and read further rather than merely being satisfied with a summary.

[8] Our paper also relates to the literature on the economics of innovation. The main premise of this literature is that, when knowledge is cheap to imitate, innovative rewards are vulnerable to expost expropriation by imitation (Henry and Ponce, 2011). Hence the need for patents. The same reasoning could be applied in the case of original information production and the need for copyright as legal means of protecting media outlets' profit.

The rest of the paper is organized as follows. Section 2 below describes the media universe and the content data we use in this paper. It reviews the algorithms we develop to construct the set of news stories and presents descriptive evidence on the production of online information. Section 3 provides an illustration of the propagation of information. In Section 4, we study empirically the costs and benefits of original information production. Finally, Section 5 discusses the need to develop new paywall and/or copyright models and concludes.

2 Data and descriptive statistics

2.1 Media universe

Our dataset covers 84 general information media outlets in France. The outlets included in the dataset can be classified into five categories: 1 news agency (AFP); 52 newspapers (local and national daily, national weekly, and free newspapers); 10 pure Internet players (online-only media outlets); 13 television channels; and 8 radio stations.

Using their RSS feeds, we track every piece of content these outlets produced online in 2013.[9,10] This content data is from the OTMedia research projet (2010-2013).[11] INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French radio and television audiovisual archives) was the project and technical leader.

2.2 Content data

Documents Our dataset contains 1,872,248 documents for the year 2013; 5,130 documents on average per day. Figure 1 plots this number on a daily basis. 47% of the documents are from the websites of the print media; 18% from radio; 13% from television; 17% from the news agency (AFP) and the remaining documents from the pure Internet players. On average, these documents are 2,010 characters long, but with a very large variance. Table D.1 in the online Appendix provides summary statistics for the entire sample, as well as by media type (print media, television, radio, pure Internet player and news agency).

Topic detection algorithm Using this set of documents, we perform a "media event detection" or "topic detection" algorithm to detect events. We first need to define our concept

[9] More details on the RSS feeds we capture are provided in the Technical Annex to this paper.

[10] Obviously, by only using the information published online by the different outlets, we are not capturing the whole production of information of these outlets – to the exception of the pure players. Some newspapers may for example choose to publish a number of their articles only in their printed version (offline). We discuss the extent to which this selection issue may bias our results in Section 4.3 below. Future work will consist in considering both the online and offline production of general information media outlets.

[11] This projet was subsidized by *Agence Nationale de la Recherche* (ANR – National Agency for Research), a French institution tasked with funding scientific research.

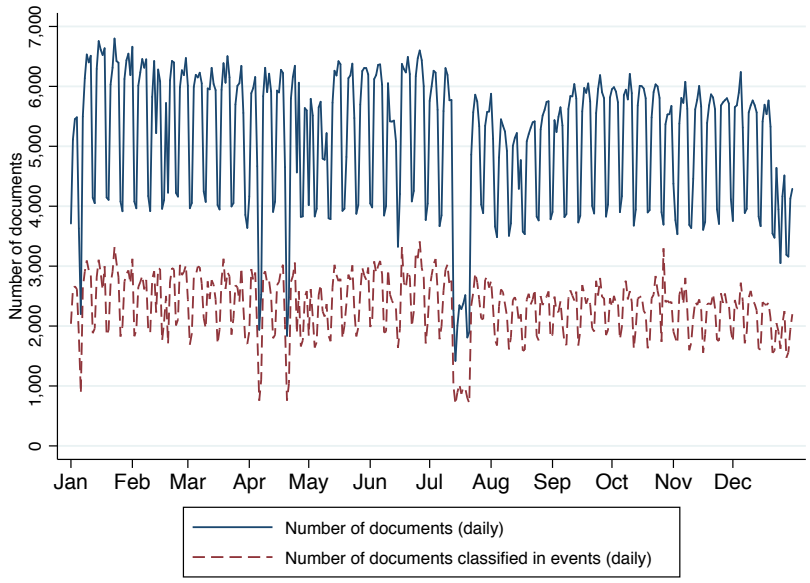


Figure 1: Daily repartition of the number of documents and of the number of documents classified in events in the dataset

of a "media event". Extracting manually the events on a given day leads to standard ambiguities: granularity, boundaries, event size limit. The computer science literature provides some approaches. But they are applied on very specific corpora, not publicly available. In this research project, we choose to implement a new approach that could be improved in the future and used by other researchers on different datasets. It could be compared to other detection systems by its ability to put all stories in a single topic together.[12]

As any natural language processing algorithm, the media event detection algorithm we develop here is based on text analysis methods. Its goal is to place all the documents into appropriate bins (clusters), such that each bin includes documents that talk about the same topic (event) and only that topic. It consists in the following steps:

- Each document is described by a semantic vector which takes into accounts both the title and the text (through standard TF/IDF approach).
- The documents are clustered in a bottom-up fashion to form the events based on their semantic similarity (we use the vector angle distance) on a daily basis.[13]
- This iterative agglomerative clustering algorithm is stopped when the distance between documents reaches a given threshold.[14]

[12] Evaluation is in terms of errors (misses and false alarms) and the tradeoff between them.

[13] We are working on detecting events that last more than one day. Results based on this new algorithm will be presented in an updated verison of the paper.

[14] We have determined empirically this threshold based on manully created media events.

- We determine the top most informative keywords for these events. We use these keywords to name the clusters.

To ensure consistency, we keep only the events with documents from at least two different media outlets; and the events whose number of documents is higher than 10.

We make here the simplifying assumption that a document discusses exactly one topic. As highlighted by Allan et al. (2005), this assumption is not true and may give rise to problems. For example, if a document discusses two topics, the simplifying assumption requires that it will be put in a single cluster (topic) and that if it were put in the other, it would be an error. Only a small percentage of documents cover multiple stories, however. Hence we think this assumption is a reasonable one.

Media events We obtain a total number of 25,877 events. On average, there are 71 events per day, roughly equally distributed during the year. Figure 2 plots the number of events per day. Out of the 1,872,248 documents in the dataset, 851,411 (45%) are classified in an event (for a daily plot of this ratio, see Figure E.1 in the online Appendix; Table D.2 provides the share of documents classified by media category).

The average number of documents per event is 16; but this number can vary strongly (Figure E.2 in the online Appendix plots the distribution of this number for the events which count less than 200 documents – 99% of the sample). The number of documents associated with an event can be seen as a proxy for the "importance" of the event. Another possible proxy for this "importance" is the number of media outlets talking about the event. On average, 10 media outlets refer to an event, but, as for the number of documents, this number varies depending on the event (online Appendix Figure E.3 plots the distribution of this number).

The remaining documents are not classified in events but are nevertheless of interest for us. Part of the job of some journalists is indeed to write feature articles about not hot news topics. Such articles would not be classified in events.

Finally, we classify the events according to their topic. In order to do so, we use information from AFP. AFP indeed includes a lot of metadata we capture with its dispatches, and in particular the subject of the dispatch. AFP uses the 17 IPTC classes to classify the dispatches.[15] These topics are: (i) Arts, culture and entertainment; (ii) Crime, law and justice; (iii) Disaster and accident; (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.

[15] More precisely, to define the subject, AFP uses URI, available as QCodes, designing IPTC media topics (the IPTC is the International Press Telecommunications Council). These topics are defined precisely in the online Appendix (Section A.2).

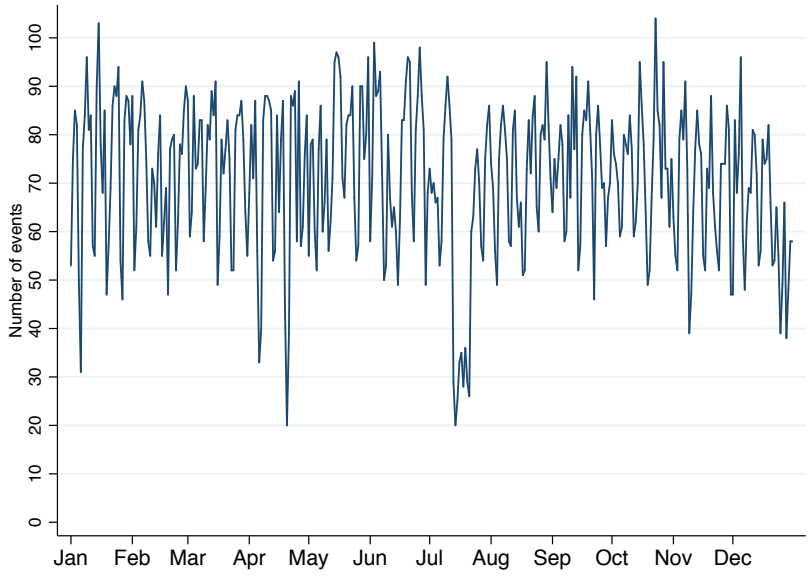


Figure 2: Daily repartition of the number of events

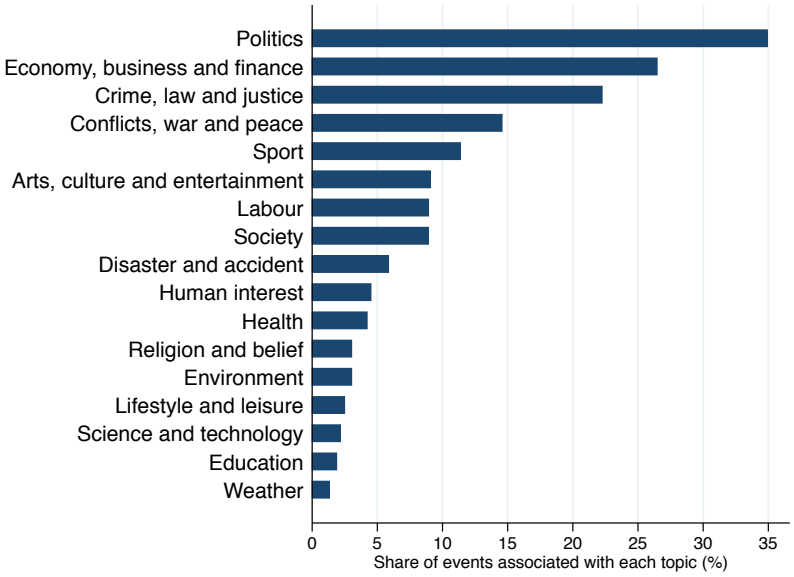
To define the topic associated with an event, we rely on the metadata associated with AFP dispatches included in the event.[16] Figure 3 plots the share of events associated with each media topic.[17] It appears clearly that the vast majority of events are about politics (35%), "economy, business and finance" (26%) and "crime, law and justice" (22% of the events), as well as as "conflicts, war and peace" (14%), and sport (11%). The other topics like weather, health or "lifestyle and leisure" have much less importance.[18] This does not mean that there is no article related to these topics, but that these topics are not associated with any *events*.

Timeline and Plagiarism detection There is a set of events $e \in [1, E]$. Each event e is characterized by a set of documents $n \in [1, N_e]$. Let T_e be the length of event e. For each document n we know the media $m(n)$ which published the document, as well as the exact time $t(n)$ at which the document has been published. For each event, we can thus order the documents depending on the timing of their publication and rank them. We obtain an ordered set of documents $1, 2, 3, ..., n', ..., n, ..., N_e$. By construction, $t(1) < t(2) < t(3) < ... < t(n') < ... < t(n) < t(N_e)$. Hence, for each news story, we determine the media outlet who breaks out the story, and then rank the other outlets. Moreover, not only we know whether a media is for example the second or the third to cover the story, but also how long it took to the media to talk about the story (the time interval between the publication of the

[16] There is at least one AFP dispatch in 86% of our events.

[17] Given that some events are associated with more than one IPTC topic, the sum of the shares is higher than 100%.

[18] Note that, to the exception of sport, the relative importance of each topic in the events corresponds to the relative importance of these topics in the set of AFP dispatches (see online Appendix Figure E.4).



Notes: The Figure shows the share of events associated with each media topic. The topics correspond to the IPTC media topics described in the text and defined in the online Appendix.

Figure 3: Share of events associated with each media topic

first document and the publication of the article by the media).

We investigate the speed of news dissemination. On average, it takes two hours for an information published by a media oulet to be published on the website of another oulet; but less than 45 minutes in half of the cases, of which less than 5 minutes in 25% of the cases. It is not because a media outlet is talking about a story that the media outlet is providing original reporting on this 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 copy rate between two articles.

Consider a document n. To compute the copy rate of this document we proceed as follows. First, $\forall n' \in [1, n[$, we compute the IP rate (Identical Portion) defined as:

$$IP\left(n, n'\right) = \frac{\mid n \cap n' \mid}{\mid n \mid} \tag{1}$$

For every couple of documents (n, n'), $IP\left(n, n'\right)$ gives us the share of document n that is in document n'. We develop a plagiarism detection algorithm to compute this IP rate. The algorithm – which has state of the art performances – tracks efficiently small portions of text that are identical between documents.[19] We then aggregate all the identical portions of text between the two documents. Currently, we focus on exact copies only.[20]

[19]Note that $IP\left(n, n'\right) \neq IP\left(n', n\right)$.

[20]Technically, the algorithm is based on hashing techniques of n-grams for speedup (the n-grams consist in sets of n consecutive words, we use 5-grams) and a threshold on the minimal length of a shared text portion to consider there is a copy (we use 100 characters).

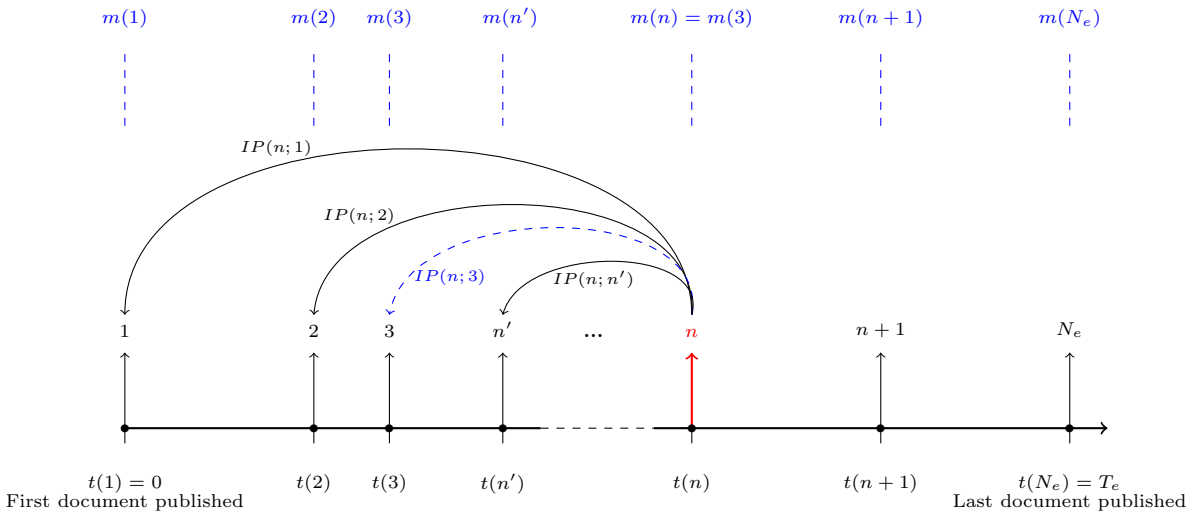


Figure 4: Timeline: illustration

We then compute the copy rate as the maximum of all the IP rates[21], i.e.:

$$copy\,(n) = \max_{1 \leq n' < n} \left(IP\left(n, n'\right) \right) \tag{2}$$

Figure 4 illustrates how the timeline of each news story is build and the way we compute the IP rates.

A media can copy documents that it has itself previously published (in particular when it is updating previous versions of the same article, for example adding new elements). Conditional on publishing at least one document related to the event, a media publishes on average 2.28 documents per event (but in 59% of the cases, media outlets only publish one document per event). Hence we also compute an *external* copy rate, excluding the documents published by the media itself:

$$copy^{external}\,(n) = \max_{1 \leq n' < n, m(n) \neq m(n')} \left(IP\left(n, n'\right) \right) \tag{3}$$

as well as an *internal* copy rate, considering only the documents published by the media itself:

$$copy^{internal}\,(n) = \max_{1 \leq n' < n, m(n) = m(n')} \left(IP\left(n, n'\right) \right) \tag{4}$$

Table 1 provides descriptive statistics on copy rates. Out of the 851,411 documents classified in events in our sample, only 41% do not have any portion of text identical to what has been previously published by another media outlet (Table 1a). On average, the copy rate is equal to 45%; it is equal to 77% if we only consider documents that present at least some

[21] We are now working on defining the copy rate as $\frac{|n \cap (n')_{n' < n}|}{|n|}$, i.e. at taking into account Identical Portions from all previously published documents, rather than just considering the maximum of all the IP rates.

Table 1: Summary statistics: Copy

(a)

Share of documents without copy (%)	41
Share of documents without external copy (%)	47
Share of documents without internal copy (%)	74
Observations	851,411

(b)

	Mean	sd	Median	Min	Max
All					
Copy rate (%)	45	43	43	0.00	100
Copy rate conditional on copy (%)	77	26	89	0.34	100
External					
External copy rate (%)	39	41	19	0.00	100
External copy rate conditional on copy (%)	73	27	85	0.34	100
Internal					
Internal copy rate (%)	17	34	0	0.00	100
Internal copy rate conditional on copy (%)	68	32	78	1.43	100
Observations	851,411				

Notes: The Table gives summary statistics for copy rates for the 851,411 documents in our sample that are classified in events.

copy (Table 1b). In other words, half of online information production is copy-and-paste. Do media outlets obey the formal procedures for citing and crediting? To answer this question, we study media references.

Media reference detection We apply an algorithm to detect media references in an article refering to a media as the source of the information. This text analysis algorithm is based on dictionnaries and grammars. The grammars gathers representation of linguistic phenomena and are based on recursive transition networks, a formalism closely related to finite state automata. Dictionnaries contain words (media names and synonyms, declarative verbs, journalist activities, ...) and their metadata (nature and description), and grammatical rules are expressed by linguistic graphs. The algorithm allows the detection of specific linguistic forms determined by these graphs.

In order to create the dictionnaries and grammar rules, we first detect media names in each document and iteratively proceed to a manual analysis of the media name used in order to extract the linguistic rules. We proceed until all media references are covered by a linguistic

Table 2: Summary statistics: Total number of references in 2013

	Mean	sd	Median	Min	Max
Total number of references	9,077	37,175	1,516	7	326,475
Print media	5,239	11,624	1,256	25	73,383
Television	6,413	5,558	5,224	112	15,277
Radio	6,796	5,420	5,448	926	14,693
Pure internet player	1,380	2,962	396	7	9,689
News agency	326,475	.	326,475	326,475	326,475
Observations	80				

Notes: The Table gives summary statistics for the total number of references to each media outlet in the dataset in 2013.

rule. These precise grammar rules allow us to distinguish when a media is referenced as a source of information or when the information is about the media itself (appointment, take over, ...). This collection of linguistic forms is updated as needed by processing new input sources.

Using this algorithm, for each event e and media outlet n, we determine how many times the other media outlets refer to outlet n. On average, for each event, 1.22 media outlets are mentioned by other media outlets. The main outlet to be refered to is AFP: more than half of all media references refer to AFP (see online Appendix Table D.3). Table 2 provides summary statistics on the total number of references to the media outlets in our dataset, aggregated over the year 2013. If we consider the total number of references over the year, for each media outlet, we find on average 9,077 references to the media in 2013. This number is much higher if we only consider AFP, with a total of 326,475 references in 2013.

2.3 Explanatory variables

Different media outlets produce different quantity of original information online. Who are the main providers of original news? To answer this question, we combine the content data described above with data on investments in news gathering at the outlet level. We focus in particular on the size of the newsroom.

Size of the newsroom We collect information on the number of journalists working for each media outlet, as well as on the total number of employees. (For a sub-sample of the media outlets, we also have information on the number of journalists specifically working for the website of the outlet, but there is only a few outlets that consider their digital newsroom separately). We complement this data on the number of journalists and employees with data on the total payroll of the outlets, as well as data on the share of the payroll devoted to

Table 3: Summary statistics: Size of the newsroom

	Mean	sd	Median	Min	Max
Number of journalists	122	131	88	0	780
Print media	123	107	88	1	562
Television	161	125	122	5	394
Radio	116	100	105	12	335
Pure internet player	8	9	4	0	28
News agency	780	.	780	780	780
Observations	84				

Notes: The Table gives summary statistics for the size of the newsroom.

the journalists. For each media outlet, we indeed have information on the total number of journalists and the wage of each journalist, along with information on the specific occupation of the journalist (e.g. editor, international correspondent,...), its age and experience.[22] Table 3 provides summary statistics for the size of the newsroom. AFP has the largest news desk with 780 journalists. On average, there are 123 journalists working in the newsroom of a print media, 161 for a television channel and 116 for a radio station. Pure players have much smaller newsrooms, with some of them having no professional journalists at all.

What is driving these investments in news gathering ? We first consider audience.

Audience For newspapers, we collect data on circulation, readership (annual data) as well as on online audience. We measure online audience using data from the OJD: for each website, 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.[23] We also have this information for pure players, as well as for the websites of television channels and radio stations.

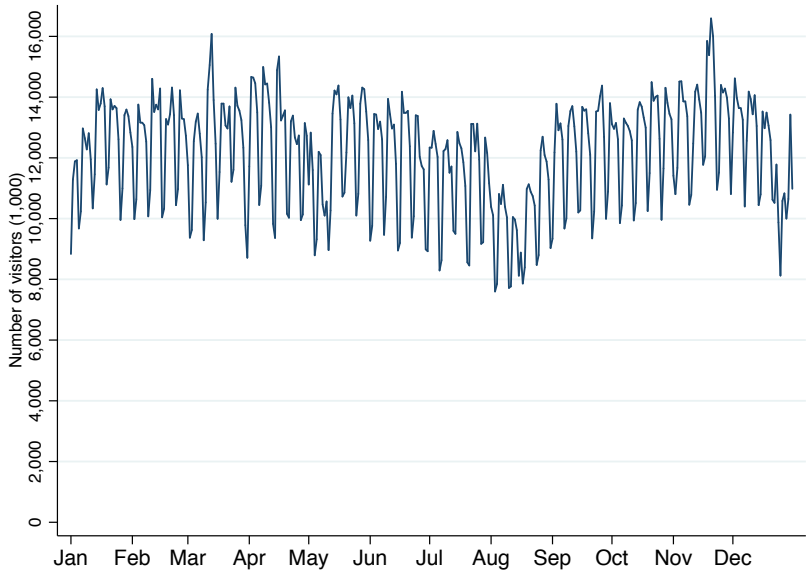
Figure 5 shows the number of unique visitors on a daily basis, aggregated over all the websites of the media outlets included in our sample. This number varies strongly from one day to the other, with a number of peak audiences. We use this daily audience data to investigate to which extent demand is higher for the breaking news media or for the media outlets that provide original reporting.

Providing original content can also be used as a brand building strategy, which may be the case if a media is referred to as the source of the information. To see whether it is the case, we finally use as an explanatory variable the number of times the media is refered to.

Before turning to the empirical analysis, we provide an illustration of the propagation of

[22]This very detailed data on the characteristics of the newsroom of every French general information media outlets is from Cagé (2015a).

[23]These three measures are very strongly correlated: the coefficient of correlation is higher than .9 and significant at the 1% level.



Notes: The Figure shows the number of unique visitors on a daily basis, aggregated over all the websites of the media outlets included in our sample.

Figure 5: Number of unique visitors

information.

3 The propagation of information: illustration

On Monday October 21, 2013, the national daily newspaper *Le Monde* reported that the National Security Agency (NSA) accessed more than 70 million phone records of French citizens in a single month, from December 10, 2012 to January 8, 2013. This big story, a worlwide exclusive entitled *"Comment la NSA espionnne la France"* ("France in the NSA's crosshair") was published on the newspaper's website at 06:01:13 am.[24] *Le Monde* published at the same time a second article on the topic, entitled *"L'ampleur de l'espionnage mondial par la NSA"* ("Inside the NSA's web of surveillance").

30 seconds later, at 06:01:43 am, the French news agency AFP published a dispatch on the same topic (*"La NSA a recolté des millions de données en France"*). How to explain such a high reactivity? The dispatch was very short (494 characters) and 40% of its content was copy-and-paste of *Le Monde*'s original article, as it appears in Figure 6a which illustrates our plagiarism detection algorithm. At 06:01:48 am (35 seconds after the publication of the first article by *Le Monde*), AFP published a second dispatch, much longer than the first one (3,177 characters). 75% of the content of this dispatch was copy-and-paste of the content of *Le Monde*'s article (Figure 6b). In both cases, AFP refers to *Le Monde* a number of times as

[24] Online Appendix Figure F.1 provides a screenshot of the first paragraphs of the article.

the source of the information.

Half an hour later, the first non-news agency media outlet to report online on this extensive electronic eaversdropping with France is a radio station, RTL (at 06:29:00 am). 81% of the 2,976 character-long article published by RTL is simple copy of the longest AFP dispatch. When 12 minutes later (at 06:40:58am), *Le Nouvel Observateur* (a national weekly newspaper) reports the story, 89% of its 3,526 character-long article is copy-and-paste from AFP (Figure 6c). Both media outlets refer to *Le Monde*.

Overall, on October 21, 2013, the story broken out by *Le Monde*, classified as "Politics" using the IPTC topics, gave rise to 213 articles by 52 media outlets. Just within three hours after the publication of the first article by *Le Monde*, 55 articles related to the event had already been published.

Le Monde has been refered to 304 times by other media outlets, but AFP has also received some credit for the story with 46 references. At 8am, in a rapid reaction, Interior Minister Manuel Valls spoke out againt US spy practives on the radio station Europe 1, explicitely refering to *Le Monde* as the source of the information.[25] Thanks to Manuel Valls' reaction, the radio station was refered to 42 times on October 21. Figure 7 illustrates our reference detection algorithm.

This exemple is meant to illustrate the propagation of information online and the algorithms we develop to study it. It reflects both how fast information is spreading, and the importance of copy-and-paste. The next section studies empirically the determinants of information production and the benefits of breaking out a story.

4 Empirical analysis

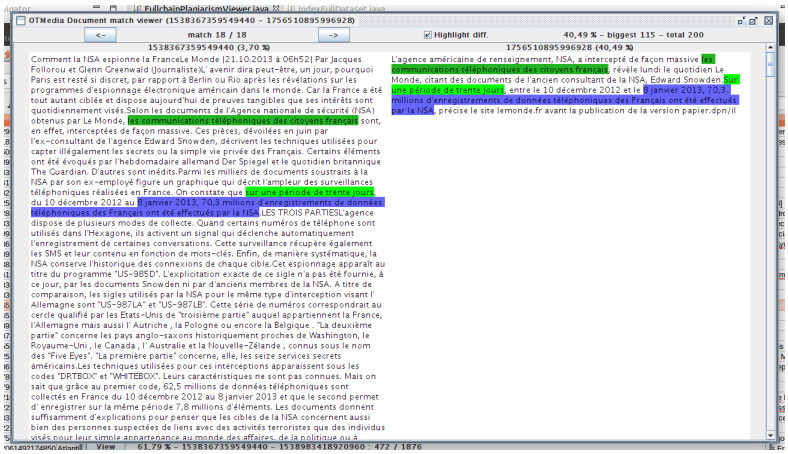
Who are the main providers of original news in an online world? Before turning to the empirical estimations, we present some cross-sectional graphical evidence on the relationship between the size of the newsroom and information production by media outlets.

4.1 Mapping the production function of information

We use seven different measures of information production: (i) the total number of documents produced by the media outlet in 2013; (ii) the total number of documents classified in events; (iii) the total content (total number of characters); (iv) the total content classified in events; (v) the total original content[26]; and (vi) the share of breaking news.

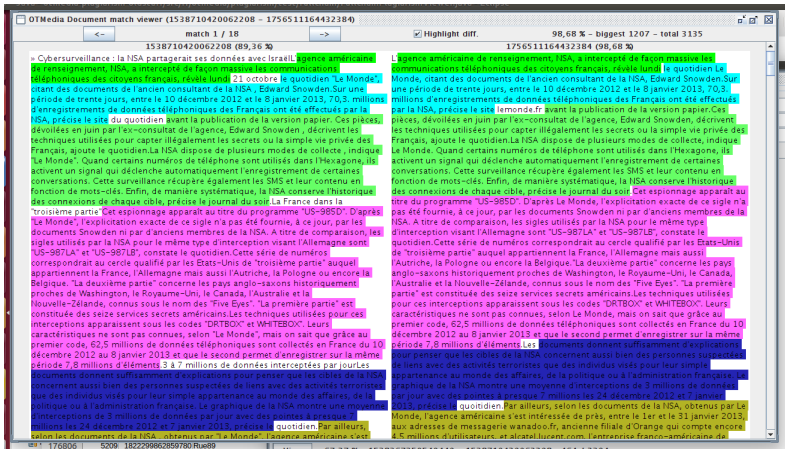
[25] *"The revelations on Le Monde are shocking and demand adequate explanations from the American authorities in the coming hours".*

[26] The total original content of an article is the total number of characters of this article minus the characters that are copy-and-paste from the previously published article with the highest copy rate. Given that an article can reproduce portions of more than one previously published article, this measure overestimates the



(a) Copy rate between 1st AFP dispatch and *Le Monde*'s exclusive

(b) Copy rate between 2nd AFP dispatch and *Le Monde*'s exclusive



(c) Copy rate between 2nd AFP dispatch and *Le Nouvel Observateur*

Figure 6: Illustration of the plagiarism detection algorithm: the NSA spying scandal on October 21, 2013

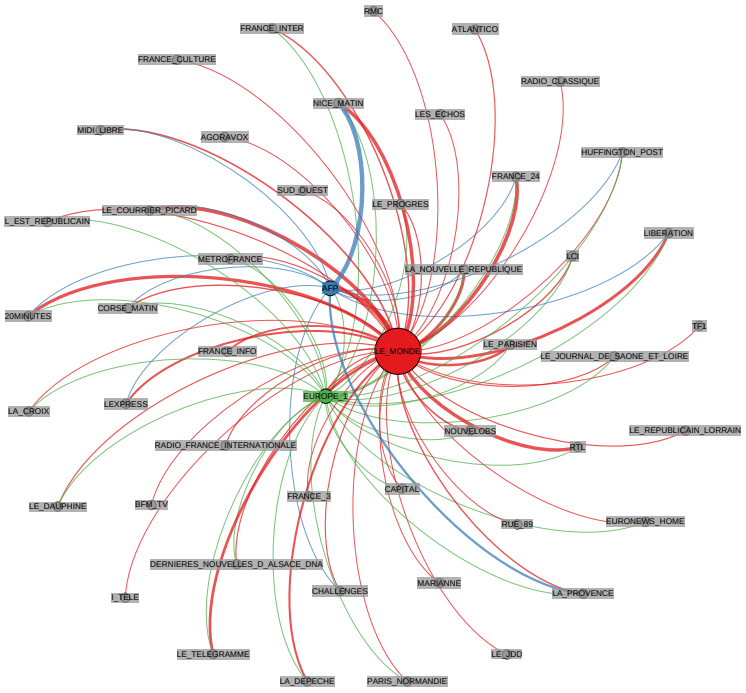


Figure 7: Illustration of the reference detection algorithm: the NSA spying scandal on October 21, 2013

The role of press agencies Figure 8 plots the relationship between these different measures of information production and the size of the newsroom. For all the different measures, there is a clear outlier: AFP, which has the largest news desk and is the main provider of original information, whether we consider original reporting or breaking news. The AFP initiates one third of the news. Moreover, there is at least one AFP dispatch in 86% of the events.

How to explain this predominant role played by AFP? It may reflect the use of an adequate copyright system. AFP is indeed the only actor which gets paid for the use of its content. But this actor is also characterized by the specificity of its business model: the goal of AFP is not to maximize the audience of its website in order to obtain advertising revenues. AFP does not rely on advertising revenues, and while it has a website, it does aim at maximizing audience. The revenues of AFP are from the subscriptions of other media outlets.[27] Hence what AFP aims at maximizing is the number of subscriptions, which depends on the quality – and in particular the novelty – of the information provided in AFP dispatches.[28]

"originality" of the article. To fix this overestimation bias, we are now working on defining the originality rate with respect to all previously published articles.

[27] As well as from the French government which is the primary client of AFP. Government subscriptions to AFP (which amounts to €120 million in 2013, i.e. nearly one fourth of the total government budget for the press) are a way to subsidize news production by the agency (AFP statutes prohibits direct government subsidies).

[28] Given the price of the subscription – which is sometimes prohibitive – some newspapers have chosen to terminate it. In 2010 for example, the free newspaper *20 Minutes* terminated its subscription to AFP because

In the rest of the empirical analysis, we drop AFP from our sample. We come back to it in Section 4.3, when discussing the relevance of introducing copyright laws.

A transmedia approach Beyond AFP, there is a diversity of online actors and business models. Figure 9 shows the correlation between the size of the newsroom and the production of information when AFP is excluded from our sample of media outlets. First, we see that there is positive correlation between the number of journalists and information production. We estimate the economic and statistical significance of this relationship below (Section 4.2). Second, we see that, at least graphically, there is not a lot of differences between a newspaper, a radio station and a television channel online. Newspapers appear with blue dots, radio station with green squares and television station with red triangles in the Figure. The size of the newsroom is not a function of the offline support, nor is the quantity of information produced offline.[29] Only pure Internet players (with yellow diamonds) are odd. To the exception of one of them, Mediapart, they have a very small newsroom (the median number of journalists is 4 for pure players). This may change in the future, but at least up to the evidence we have for 2013, they do not play an important role in the online production of information in France, nor in the online audience, as we will see below.

4.2 Econometric estimates

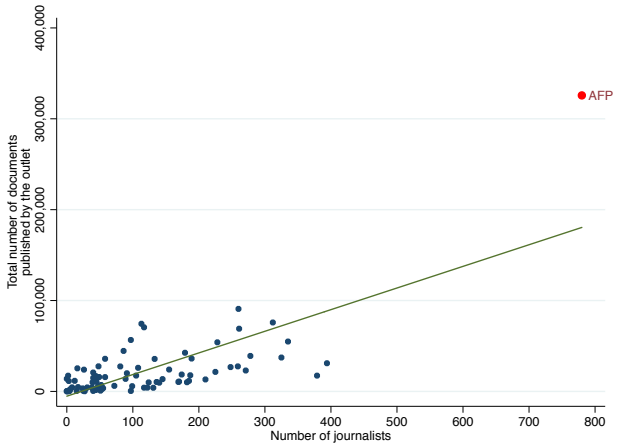
Cross-sectional analysis To estimate the determinants of information production, we first perform a cross-sectional estimation by aggregating the information produced by the media outlets in 2013. Equation 5 describes our identification equation:

$$\ln(\text{information production})_n = \alpha + \beta_1 \ln(\text{Number of journalists})_n + \gamma_{\text{media}} + \epsilon_n \qquad (5)$$

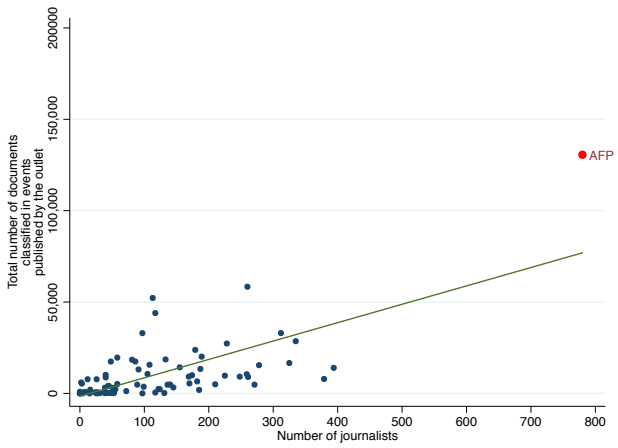
where n index the media outlets. The outcome of interest, $\ln(\text{information production})_n$, is alternatively the logarithm of (i) the number of events covered by the media outlet in 2013; (ii) the total content (total number of characters); (iii) the total content classified in events; (iv) the total original content; and (v) the share of breaking news. (In the online Appendix Table C.1, we present additional results on the total number of documents produced by the media outlet and the total number of documents classified in events.) γ_{media} are media category (newspapers, television, radio, pure Internet players) fixed effects. Standard errors are robust.

of its price.

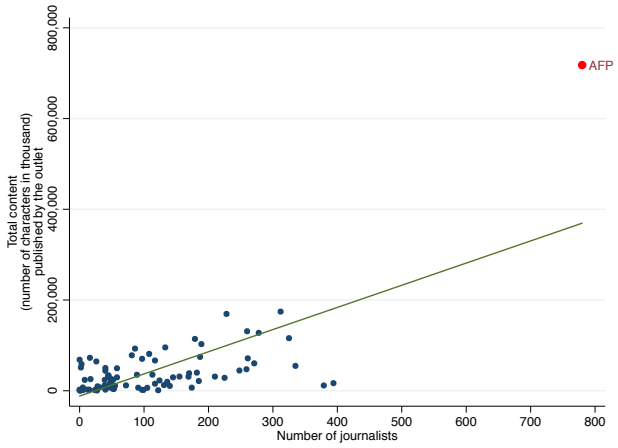
[29] The two television channels (red triangles) with the highest number of journalists seem to produce much less content than their number of documents. This comes from a technical issue we faced when capturing the RSS feeds of these stations. This issue will be fixed in an updated version of the paper.



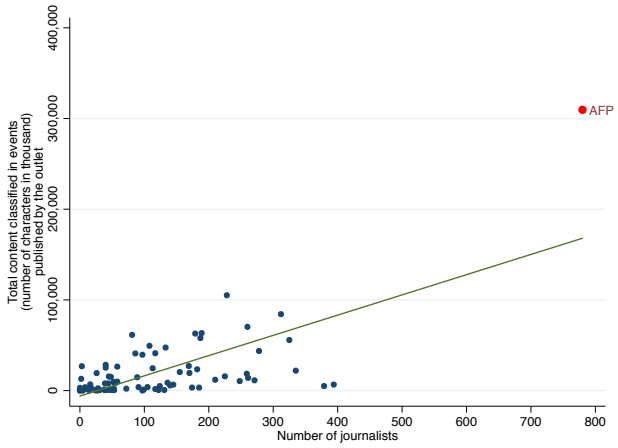
(a) Number of documents



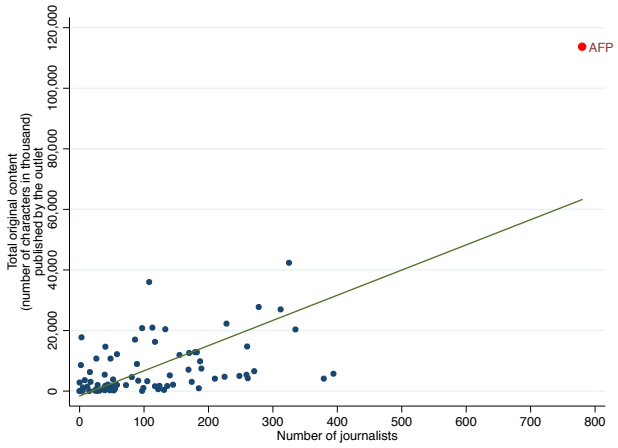
(b) Number of documents classified in events



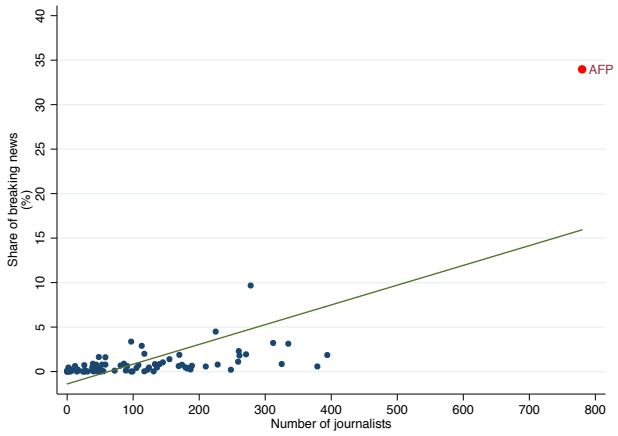
(c) Total content (number of characters)



(d) Total content (number of characters) classified in events



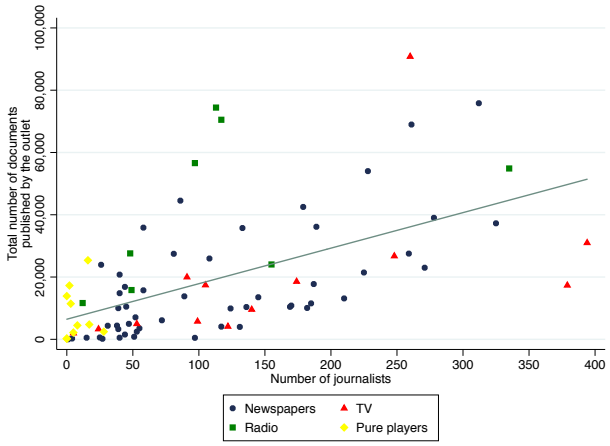
(e) Total original content (number of characters)



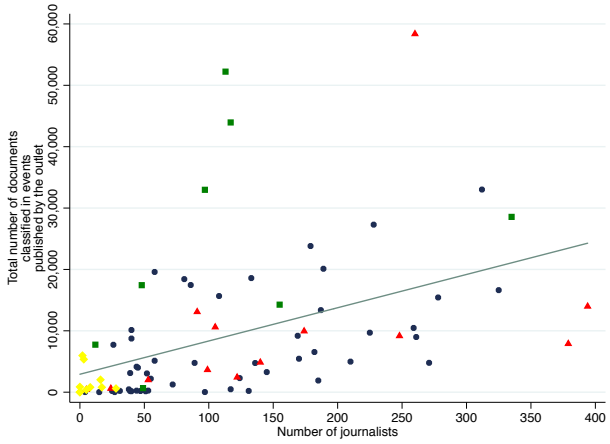
(f) Share of breaking news (%)

Notes: The Figures show the correlation between the size of the newsroom and different measures of information production for all the media outlets included in our sample and described in more details in the online Appendix.

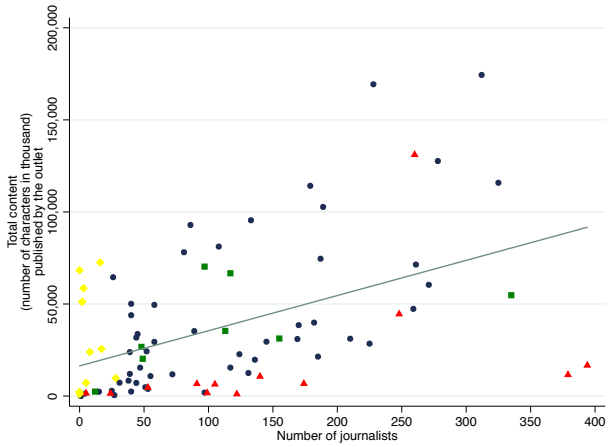
Figure 8: Information production and journalists: The role of AFP



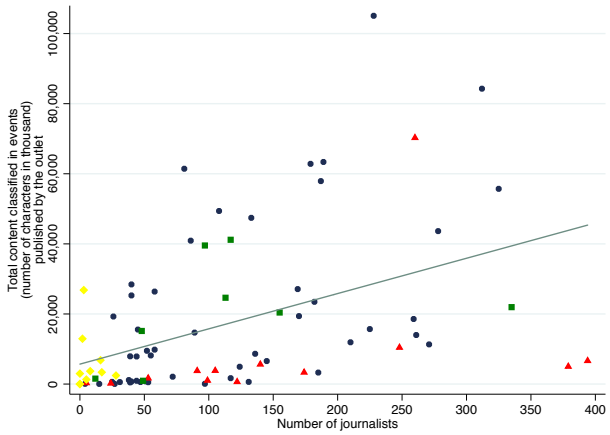
(a) Number of documents



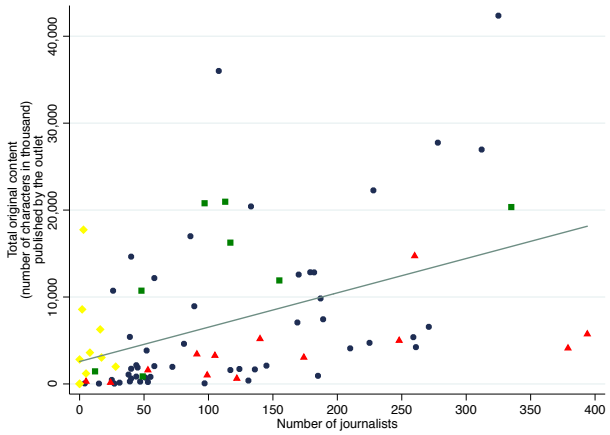
(b) Number of documents classified in events



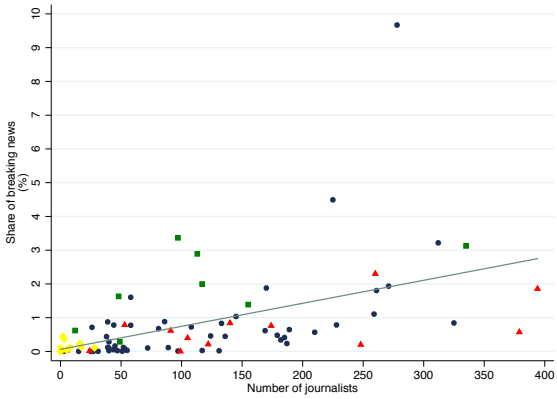
(c) Total content (number of characters)



(d) Total content (number of characters) classified in events



(e) Total original content (number of characters)



(f) Share of breaking news (%)

Notes: The Figures show the correlation between the size of the newsroom and different measures of information production for all the media outlets included in our sample excluding AFP.

Figure 9: Information production and journalists: A transmedia approach

Table 5 presents the results. Columns 1 and 2 provide the results of the estimations when all the media outlets are included (column 2 includes category fixed effects). First, we find that the higher the number of journalists working for a media outlet, the more events it covers (sub-Table 4a): a 1 percent increase in the number of journalists increases the number of events covered by .9 to 1.2 percent (columns 1 and 2). This effect is stronger for newspapers: a 1 percent increase in the number of journalist increases the number of events covered by 1.5 percent (column 3).

Second, we find that the total content produced by media outlets (using the total number of characters), and in particular the total original content, also increases with the size of the newsroom. When controlling for media category fixed effects, a 1 percent increase in the number of journalits increases the total original content produced by 1.1 percent (column 2, sub-Table 4d). In other words, there is a quasi-linear relationship between the number of journalists and the quantity of information produced.

Interestingly, when we turn to the share of breaking news (sub-Table 4e), we find that there are increasing return to scales, in particular for newspapers: a 1 percent increase in the number of journalists increases the probability that a newspapers breaks out a news by 5.2 percents (column 2).

Event-level analysis Performing the empirical estimation at the event level allows us to control for a number of important factors, in particular the topic of the event (different media outlets may devot more efforts on certain topics than on others). We can also introduce a number of controls at the event- and at the day-level. Equation 6 describes our prefered identification equation:

$$\ln(\text{information production})_{end} = \alpha + \delta_1 \ln(\text{Number of journalists})_n + \mathbf{X}'_{\mathbf{ed}}\delta_{\mathbf{2}} + \mathbf{Y}'_{\mathbf{d}}\delta_{\mathbf{3}} + \gamma_{\text{media}} + \lambda_{\text{topic}} + \epsilon_{end} \quad (6)$$

where n index the media outlets as before, e the event and d the date of the event. The vector of event-level controls $\mathbf{X}'_{\mathbf{ed}}$ include the number of documents in the event and the number of media outlets talking about the event. The vector of daily-level controls $\mathbf{Y}'_{\mathbf{d}}$ includes the number of events during the day. Standard errors are clustered at the media outlet level.

Table 5 presents the results. We find as before a positive relationship between the size of the newsroom and the production of information. A 1 percent increase in the size of the newsroom increases the total original content produced by a newspaper in an event by 0.3 to 0.4 percent (sub-Table 5b columns 1 to 4). Note that the order of magnitude is the same when we only focus on newspapers (colums 5 and 6). The results are robust to controlling or

Table 4: The effect of the number of journalists on the production of information: Cross-sectional estimation

(a) **Number of events covered (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	1.0***	1.2***	1.5***	0.8***	0.5	0.9
	(0.2)	(0.2)	(0.1)	(0.1)	(0.3)	(0.6)
Category FE	No	Yes	No	No	No	No
R-sq	0.41	0.51	0.55	0.77	0.19	0.20
Observations	82	82	51	13	8	10

(b) **Total content (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.6***	1.0***	1.3***	0.8***	0.9**	0.3
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.4)
Category FE	No	Yes	No	No	No	No
R-sq	0.25	0.53	0.63	0.43	0.73	0.09
Observations	82	82	51	13	8	10

(c) **Total content classified in events (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.9***	1.3***	1.5***	1.0***	1.0**	1.1
	(0.2)	(0.2)	(0.2)	(0.2)	(0.3)	(0.7)
Category FE	No	Yes	No	No	No	No
R-sq	0.32	0.43	0.42	0.57	0.50	0.24
Observations	81	81	50	13	8	10

(d) **Total original content (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.7***	1.1***	1.3***	0.9***	1.0***	1.1
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)	(0.7)
Category FE	No	Yes	No	No	No	No
R-sq	0.27	0.43	0.42	0.65	0.55	0.25
Observations	81	81	50	13	8	10

(e) **Share of breaking news (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	3.6***	4.5***	5.2***	0.9***	0.6***	7.7*
	(1.1)	(1.3)	(1.6)	(0.2)	(0.2)	(4.1)
Category FE	No	Yes	No	No	No	No
R-sq	0.29	0.33	0.34	0.34	0.43	0.35
Observations	82	82	51	13	8	10

Notes: * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors in parentheses. Models are estimated using OLS. Variables are described in more details in the text.

not for topic fixed effects. The fact that the magnitude of the effect is smaller than when we perform the cross-sectional analysis comes from the propensity of media outlets with bigger newsroom to cover more events, as we saw in sub-Table 4a. Given that a media outlet with more journalists cover more events, the total original content produced by the outlet over the year increases more strongly with the number of journalists than the total original content produced *by event*.

The benefits of original information production Hence, information is costly to produce for media outlets. Is there any benefits for original news producers? To answer this question, we investigate the relationship between the number of journalists and the audience of the website. Table 6 presents the results for the three measures of audience we have: the number of unique visitors, the total number of visits and the total number of page views. We find that the audience of the websites increases linearly with the number of journalists: a 1 percent increase in the number of journalists working for a newspaper increases the number of unique visitors on the website of this newspaper by 1 percent.

However, online audience does not translate into significant revenues. Online advertising revenues accounted on average for less than 5% of total revenues in 2013. Hence there is a "free rider" issue: information is costly to produce but cheap to reproduce. We discuss in Section 5 the policy implications of this finding.

4.3 Discussion

This paper is a first attempt at trying to understand who is producing news, the character of what is produced and the benefits of news production. It is worth underlying some caveats of the empirical analysis. In particular, concerning the cross-sectional analysis, we are well aware that we are only able to identify correlations, not causality. With only one year of data, there is no exogenous shock that could be used for identification. But we think that our results can nevertheless improve our understanding of the costs and benefits of information production.

The second limit of this analysis is the focus on online information. For now, we do not capture the offline production of information of media outlets. Obviously, offline and online content can sometimes coincide, but there is also a number of occurences where a media chooses to publish an information offline but not online (especially mediat outlets that do not have a paywall online). We are now working on introducing the offline content in the dataset.

Moreover, we only take into account here the "traditional" general information media outlets, not the expanding universe of new media, including blogs, Twitter or Facebook. However it has been shown (see e.g. Pew Research Center, 2010) that these new social media play only a limited role in the original production of information. They are mainly an alert

Table 5: The effect of the number of journalists on the production of information: Event-level estimation

(a) **Number of documents (log)**

	All media outlets				Newspapers		Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.0***	0.0***	0.0***	0.0***	0.1**	0.1**	0.2*	0.2**	-0.0
	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.0)
Nb docs in event (log)		0.1***	0.2***	0.2***	0.6***	0.6***	0.7***	0.9***	0.3***
		(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
Nb media in event (log)		-0.0	-0.1***	-0.1***	-0.6***	-0.5***	-0.5***	-0.5***	-0.3***
		(0.0)	(0.0)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.0)
Nb events in day (log)		0.0	0.0*	0.0	0.0**	0.0**	0.0	-0.1	0.1***
		(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)	(0.0)
Category FE	No	No	Yes	Yes	No	No	No	No	No
Topic FE	No	No	No	Yes	No	Yes	No	No	No
R-sq	0.02	0.12	0.14	0.14	0.25	0.23	0.27	0.42	0.16
Observations	2,090,426	2,090,426	1,805,640	1,805,640	215,620	204,117	48,455	56,648	11,868
Clusters	82	82	82	82	51	51	13	8	10

(b) **Total original content (log)**

	All media outlets				Newspapers		Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.3***	0.3***	0.4***	0.4***	0.4*	0.4*	0.2	0.5***	0.4***
	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.1)	(0.1)
Nb docs in event (log)		0.5***	0.7***	0.7***	1.1***	1.0***	1.1***	1.0***	0.7***
		(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)	(0.1)	(0.1)
Nb media in event (log)		0.8***	0.6***	0.6***	-1.1***	-0.9***	-0.6***	-0.4**	-0.8***
		(0.1)	(0.1)	(0.1)	(0.2)	(0.2)	(0.1)	(0.2)	(0.1)
Nb events in day (log)		0.1	0.1*	0.1*	0.5***	0.4***	-0.0	0.2*	0.3*
		(0.1)	(0.1)	(0.0)	(0.1)	(0.1)	(0.1)	(0.1)	(0.2)
Category FE	No	No	Yes	Yes	No	No	No	No	No
Topic FE	No	No	No	Yes	No	Yes	No	No	No
R-sq	0.04	0.15	0.16	0.16	0.10	0.11	0.24	0.26	0.19
Observations	2,090,426	2,090,426	1,805,640	1,805,640	215,620	204,117	48,455	56,648	11,868
Clusters	82	82	82	82	51	51	13	8	10

Notes: * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors in parentheses. Models are estimated using OLS. Variables are described in more details in the text.

Table 6: The effect of the number of journalists on audience:
Cross-sectional estimation

(a) **Number of unique visitors (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.4**	0.7***	1.0***	0.2**	0.8***	-0.1
	(0.2)	(0.1)	(0.2)	(0.0)	(0.1)	(.)
Category FE	No	Yes	No	No	No	No
R-sq	0.19	0.36	0.44	0.09	0.82	1.00
Observations	50	50	35	7	6	2

(b) **Number of visits (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.5**	0.7***	1.0***	0.2***	0.8***	-0.1
	(0.2)	(0.1)	(0.2)	(0.0)	(0.1)	(.)
Category FE	No	Yes	No	No	No	No
R-sq	0.21	0.38	0.46	0.11	0.85	1.00
Observations	50	50	35	7	6	2

(c) **Number of page views (log)**

	All media outlets		Newspapers	Television	Radio	Pure player
	(1)	(2)	(3)	(4)	(5)	(6)
	b/se	b/se	b/se	b/se	b/se	b/se
Number of journalists (log)	0.5**	0.7***	1.1***	0.2**	0.7***	-0.1
	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)	(.)
Category FE	No	Yes	No	No	No	No
R-sq	0.21	0.37	0.48	0.13	0.72	1.00
Observations	50	50	35	7	6	2

Notes: * p<0.10, ** p<0.05, *** p<0.01. Robust standard errors in parentheses. Models are estimated using OLS. Variables are described in more details in the text.

system and a way to disseminate stories from other places.

Finally, we are still working on improving the algorithms we develop to study the propagation of online information. In particular, to construct the set of news stories, we clustered the documents in a bottom-up fashion to form the events based on their semantic similarity on a daily basis. This may bias our results given that some events last more than one day. The direction of the bias is not clear. An updated version of the paper will soon be available where the documents will be clustered with no time constraint (so that we will capture all the events lasting more than one day). We are also working on improving the plagiarism detection algorithm in two directions: (i) computing copy with respect to all previously published documents rather than simply with the document with the highest copy rate; (ii) capturing reformulations on top of exact copies. Note that we are also improving the capture of a number of RSS feeds. These further steps should not change the main results of the paper, but they will allow us to investigate more precisely the production of online information.

5 Conclusion and policy implications

Where does the news come from in today's changing media? With the decrease in the size of the newsrooms in France and the United States (see e.g. Cagé, 2015b), there is a feeling that general information media outlets are scaling back on original reporting while increasingly reproducing other people's work. This paper provides evidence of the existence of such a "free rider" issue. Producing information is costly for media outlets. They benefit from it through audience, but are not able to monetize this audience online. In particular, online audience does not translate into significant revenues. Hence the need to develop new paywall and/or copyright models.

Traditionally, copyright violations occurred when someone manually recopied, then reprinted, large portions of someone else's story. While in the past the so-called "fair use" doctrine allowed a newspaper to comment on its competitors' day before storyline, making some selective quotes, copyrights are violated with the click of a mouse nowadays. Currently, the copyright law is governed in France by the *"Code de la propriété intellectuelle"* (French Intellectual Property Code) of 1992. In the United States, it is governed by the Copyright Act of 1976 in the United States (see e.g. Fox, 2009). To receive protection, a work must be original, fixed, and an expression; in particular, the copyright law does not protect facts. In other words, a news article, as expressed by the author's sentences and structure, is copyrighted, but the facts underlying the story are not – with the notable exception of the misappropriation or "hot news" doctrine. The "hot news" doctrine was announced in a Supreme Court decision in 1918; the decision of the Court rested on the idea that – though the Associated Press had no copyright in the facts underlying their stories – *"unfair competition in business"* could

lead the AP to lose incentives to publish news reports in the first place. Hence the Supreme Court granted a limited right to the AP enforceable against the International News Service (another newswire) allowing it to prevent the INS from publishing stories based on the news AP had discovered for a limited time. This right existed for the time necessary to allow AP to make a sufficient profit.

But the scope of misappropriation narrowed after the passage of the 1976 Copyright Act. Moreover, if hot news protection does exist, its scope has to be considered in light of recent technological changes. How could one grant a media rights against various digital "competitors" – other newspapers, aggregators, bloggers, and the public? The digital revolution means that consumers can now quickly and easily access content that is aggregated from many online sources. The results of this paper provide additional argument in favor of introducing copyright laws. Indeed "unfair" competition does not stem only from aggregators, but also from competing media outlets.

Instead of strengthening copyright laws to give the news industry more protection, some are proposing that the government asks content producers to give up copyright rights in exchange for a direct financial subsidy.[30] The case of AFP is of interest from this point of view. As we highlighted above, not only AFP is the only actor which gets paid for the use of its content, but it also receives indirect subsidies through government subscriptions.

Governments also subsidize original news production through the radio stations and television channels that are publicly funded. In France, we show that publicly funding media outlets have invested massively in online news production. In the United Kingdom, the public financing of audiovisual media exceeds 80 euros per capita, primarily from the license fee. According to the BBC Annual Report 2014-2015, 5% of this licence fee is spent on BBC Online and BBC iPlayer. Moreover, while television and radio remain the most popular source of news, more and more people are using the BBC's online service to access it.[31]

Finally, in this online world, another interesting issue is the one of the paywalls. Paywall can be seen as the most efficient way for media outlets to monetize their content online. Recently, several media outlets, including the *New York Times* in the US and *Le Monde* in France, have moved from providing online content free of charge to implementing paywalls where readers are charged a fee for accessing content online. Chiou and Tucker (2013), investigating the effect of a paywall on newspaper readership, find that the introduction of a paywall leads to a 51% decrease in the overall visits. However, this lost of audience

[30] Then the question is who should pay for the subsidy? It could be the government but it could also be firms that extract profits through using content produced by others. Hence, in the recent past, a number of disputes took places between Google News and news organizations in different countries. In France, in 2013, Google set up a 60 million euros fund to finance digital publishing innovation. In 2015, Google announced a new digital partnership with eight European publishers (the "Digital News Initiative"). In this new partnership, Google is to establish a working group to focus on product development, and to providing a €150 million innovation fund over three years, alongside additional training and research.

[31] Around 50% of the UK adult population access BBC Online each week.

does not imply that the media outlets that introduce a paywall face an overall decrease in their incentives to produce original information. It won't be the case in particular if the revenues from online subscriptions where high enough to compensate the decrease in online advertising revenues. Yet if competitors copy-and-paste the content that is behind a paywall, then paywalls may not be sufficient to provide media outlets with incentives to produce original news. In particular, media outlets may suffer from a decrease in their online subscriptions. We hope this paper will inform the debate on the optimal business models and/or policy interventions.

References

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

Cagé, Julia, "Media Competition, Information Provision and Political Participation," Working Paper 2014.

_ , "Do Journalists Drive Media Bias? Payroll and Inequality within the Newsroom," Working Paper 2015.

_ , *Sauver les médias: Capitalisme, financement participatif et démocratie* La République des idées, Seuil (English version forthcoming: *Saving the Media. Capitalism, Crowdfunding and Democracy*, Harvard University Press), 2015.

Chiou, Lesley and Catherine Tucker, "How Does Content Aggregation Affect Users' Search for Information?," Working Papers 11-18, NET Institute October 2011.

_ **and** _ , "Paywalls and the demand for news," *Information Economics and Policy*, 2013, *25* (2), 61–69.

Fox, Ariel, "Copyright, Competition and Publishers' Pursuit of Online Compensation," Technical Report 2009.

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

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

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

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

Giorcelli, Michela and Petra Moser, "Copyright and Creativity: Evidence from Italian Operas," Working Paper 2015.

Golub, Benjamin, Matthew O Jackson, and Ronald L Graham, "Using selection bias to explain the observed structure of Internet diffusions," *Proceedings of the National Academy of Sciences of the United States of America*, 2010, *107* (24), pp. 10833–10836.

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

Henry, Emeric and Carlos J Ponce, "Waiting to Imitate: On the Dynamic Pricing of Knowledge," *Journal of Political Economy*, 2011, *119* (5), 959–981.

Liben-Nowell, David and Jon Kleinberg, "Tracing Information Flow on a Global Scale Using Internet Chain-Letter Data," *Proceedings of the National Academy of Sciences of the United States of America*, 2008, *105* (12), pp. 4633–4638.

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

Popkin, Samuel L., "Changing Media and Changing Political Organization: Delegation, Representation and News," *Japanese Journal of Political Science*, 2007, *8* (01), 71–93.

Prior, Markus, "News vs. Entertainment: How Increasing Media Choice Widens Gaps in Political Knowledge and Turnout," *American Journal of Political Science*, 2005, *49* (3), pp. 577–592.

_ , *Post-Broadcast Democracy: How Media Choice Increases Inequality in Political Involvement and Polarizes Elections* Cambridge Studies in Public Opinion and Political Psychology, Cambridge University Press, 2007.

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

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

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