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

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

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

November 26, 2016

Abstract

Could greater intellectual property protection raise the incentives for original information production? In this paper, we build a unique dataset including all online content produced by the universe of news media (newspaper, television, radio, pure online media, and a news agency) in France during year 2013. We develop a topic detection algorithm that identifies each news story, trace the timeline of each story and study news propagation. We show that in 25% of the cases, once a news is broken by a media outlet, it is covered online by another outlet in less than 4 minutes. High reactivity comes with verbatim copying. Only 36% of the online content is original, and copying hardly comes with acknowledgment. Finally, using daily-level variations, we show that breaking news only slightly increases the number of online viewers. This evidence shades new light on the current debate on property in news.

Keywords: Internet, Information spreading, Copyright, Investigative journalism
JEL No: L11, L15, L82, L86

[*] We gratefully acknowledge the many helpful comments and suggestions from Filipe Campante, Lucien Castex, Matthew Gentzkow, Sergei Guriev, Emeric Henry, Elise Huillery, Laurent Joyeux, Petra Moser, Aurélie Ouss, Arnaud Philippe, Thomas Piketty, Valeria Rueda, Agnès Saulnier and Katia Zhuravskaya. We are grateful to participants at the "Big Data for Media Analysis" conference, the CEPR Public Economics Annual Symposium, the NBER Political Economy Meeting, and the NET Institute Conference on Network Economics, and to seminar participants at Banque de France, the Paris School of Economics and Sciences Po Paris. Edgard Dewitte, Anais Galdin, Lucile Rogissart and Jeanne Sorin provided outsanding research assistance. This research was generously supported by the NET Institute, the Paris School of Economics, and the Banque de France. Since November 2015, Julia Cagé has been a Board member of the Agence France Presse; this paper does not reflect the views of the AFP and responsability for the results presented lies entirely with the authors.

[†] An online Appendix with additional empirical material is available here.

[‡] Corresponding author. `julia [dot] cage [at] sciencespo [dot] fr`.

1 Introduction

While online media have dramatically increased access to information, the impact of the Internet on news coverage has spurred concerns regarding the quality of news citizens have access to. The switch to digital media has indeed affected the news production technology. The production of information is characterized by large fixed costs and increasing returns to scale (Cagé, 2014). 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 became instantaneous.[1] This makes it extremely difficult for news content providers to distinguish, protect and reap a profit from the news stories they produce.[2] One central issue is whether greater intellectual property protection could address some of these difficulties and raise the incentives for original information production.

What is the extent of online copyright infringement? Despite the intrinsic policy significance of the news industry and the growing importance of online news consumption, there is very little empirical evidence, in particular at the micro level, on the production of online information. In this paper, we attempt to open up this black box by using new micro data and relying on a machine-learning approach. We examine the universe of French news media – including newspapers, television channels, radio stations, pure online media and the French news agency Agence France Presse (AFP) – and track every piece of content these outlets produced online in 2013. Our dataset contains more than 2.5 million documents. To the extent of our knowledge, it is the very first time one adopts such a transmedia approach and covers the integrality of the content produced by media online, whatever their offline format.[3] The general structure of our dataset is illustrated by Figure 1, where we plot the total original content and number of journalists for each media outlet. One can see the general positive relationship between these two variables, and the special role played by the AFP both in terms of number of journalists and original output. In our companion paper (Cagé et al., 2016), we further investigate the structure of the production and demand functions for online news. In the present paper, we leave aside the input side (number of journalists, etc.), and focus upon the pattern of copying between media outlets.

[FIGURE 1 HERE]

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

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

[3] As we will see, on the Internet, there is a tendency of different media to converge, and it becomes increasingly difficult to classify a media outlet as belonging purely to one of the traditional media formats.

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, comprised of 850,000 documents.[4] 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 analyze the propagation of the story, second-by-second. We investigate the speed of news dissemination and the length of the stories, depending on the topic and other story characteristics, in particular its importance as proxied by the number of outlets covering it.

Because a media covers a news story, it does not necessarily imply that it is 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. To which extent does such verbatim copying fall outside the bounds of copyright law? National copyright laws act within the framework imposed by international agreements (Ginsburg, 2016). They rely on two main principles. First, copyright excludes ideas; it protects only the form of expression in which the ideas are communicated. The form of expression is what we capture in this paper by identifying and quantifying verbatim copying. Second, to violate the exclusive right of reproduction, copying should be "substantial". In this article, we quantify substantiality quantitatively, both from the point of view of the copying and of the copied media outlet.

Some stories are not the product of original reporting; e.g. in the case of a government press release giving rise to a number of articles, the first media covering the story cannot be considered as a news breaker providing exclusive news. Moreover in this case, we may overestimate verbatim copying by attributing the release to the first media and counting as copy the reproduction of the release by other outlets. To deal with this issue, we code manually all the news stories in our sample and isolate the stories that are the result of a piece of original reporting. We call these stories exclusive news events. The remaining stories are either non-exclusive news events or short news items with multiple witnesses.

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

We show that on average, news are delivered to readers of different media outlets 171 minutes after having been published first on the website of the news breaker, but in less than 230 seconds in 25% of the cases. The reaction time is the shortest when the news breaker is

[4] All the documents are not classified in events. Unclassified documents include "contextual reporting" (Schudson, 2015), one-off reports, editorial and opinion pieces, as well as local news stories that are only covered by one outlet and thus not classified in events.

the news agency, and the longest when it is a pure online media, most likely because of the need for verification. The reaction time is also shorter on average for short news items with multiple witnesses.

High reactivity comes with verbatim copying. We find that only 36% of the online content is original, and that 56% of the articles have less than 20% originality. Originality is the lowest for non-exclusive news events (due to the reproduction of releases). Overall, more than 70% of the documents classified in events present at least some external copy[5] and conditional on copying, the average external copy rate is 84.1%. Moreover, verbatim copying is substantial not only from the point of view of the copying but also of the copied media: on average, conditional on being copied, 21% of the content of a document is copied. Finally, despite the substantiality of copying, media outlets hardly name the outlets they copy.

Do breaking news outlets nonetheless benefit from their investment in newsgathering? To answer this question, we collect audience data at the daily level that we merge with the content data. Using daily-level variations, we show that the number of unique visitors of a media outlet increases with its production of original news content and whether it breaks out a news, but that the effect is economically small. This evidence shades new light on the current debate on property in news. Should the current copyright law be enforced more strictly, or should it be drastically reformed? It is well beyond the scope of our paper to provide a full answer to such a complex question. More modestly, we are the first to quantify the extent of violation and conclude that the problem is so massive that it needs to be addressed one way or another. Maybe a much extended version of our plagiarithm detection algorithm could be used to apply more strictly a clearer law. Or maybe it is simply too difficult to apply copyright laws to online news production, and one should find other ways to compensate the producers of original content, especially as copyright is a second-best solution to intellectual property provision. In our future research, we hope to build upon this work in order to provide better answers to these questions.

Related literature Using micro data, Gentzkow (2007) estimates the relationship between the print and online newspapers in demand.[6] The focus of our paper is on the production rather than on the consumption of news. Franceschelli (2011) has been the first to assess empirically the impact of the Internet on news coverage.[7] Using a dataset that includes every article published by the two main Argentinean newspapers, he reconstructs the typical

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

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

[7] 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.

timeline of a news story in the online world.[8] 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.[9] Moreover, we distinguish between "investigative stories" and "non-exclusive news events". Second, while Franceschelli (2011) relies restrictively on the mention of proper nouns to identify the news stories, we develop and run a state-of-the-art algorithm relying on word frequency without any restriction. Third and most importantly, we quantify the importance of plagiarism online, identify references to media outlets as the source of the story, and use this new evidence to discuss the need for greater intellectual property protection for news online.

Our results complement a growing literature on copyright (Boldrin and Levine, 2002, 2008; Henry and Ponce, 2011). Empirically, there is evidence that providing basic level of copyright protection may encourage artistic creativity (MacGarvie and Moser, 2014; Giorcelli and Moser, 2015; Li et al., 2015). Copyrights, however, increase the cost of accessing creative works. Biasi and Moser (2015) argue that policies that reduce the costs of accessing copyrighted content can encourage diffusion and the production of new knowledge. There is thus a tradeoff between the provision of sufficient incentives to invent and the negative effect these incentives may have on the diffusion of existing and production of new knowledge.

Most of the literature on copyright online has centered on digitization and piracy within the music industry (Rob and Waldfogel, 2006; OberholzerGee and Strumpf, 2007; Waldfogel, 2012, 2015), and there is little evidence on copying and intellectual property regarding online news media.[10] An exception is Chiou and Tucker (2015) who focus on the reproduction of content for information. They exploit a contract dispute between an aggregator and a content provider[11] They argue that producers of primary content may actually benefit from relaxing their restrictions on copyright and allowing others to disseminate their content. Their empirical analysis relies on the specific case of an aggregator, however. Aggregators only display small extracts of information, and aggregator users visit content websites after visiting an aggregator.[12] Moreover, aggregators do not produce much original content, but rather

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

[9] This media universe includes newspapers, radio stations, television channels, pure online media, as well as a news agency. This is of particular importance given that on the Internet, there is a tendency of different media to converge, independently of their offline format. To the extent of our knowledge, we are the first to study simultaneously the content produced by all the news media, independently of their offline format.

[10] Recent work has also investigated the effect of digitization projects like Google Books (Reimers, 2015; Nagaraj, 2016). For an assessment of the impact of copyright laws on the magazine industry in America during the 18th and 19th centuries, see Haveman and Kluttz (2014) and Haveman (2015).

[11] See also Athey and Mobius (2012) who analyze the impact of news aggregators on the quantity and composition of internet news consumption using a case analysis of the inclusion of local content by Google News.

[12] An aggregator may have two effects on the quality choices of competing newspapers on the Internet: a "business-stealing" effect and a "readership-expansion" effect (Jeon and Esfahani, 2012). On the theory front, see also Dellarocas et al. (2013) who examine the role of news aggregators in the competition between media outlets.

curate content created by others. On the contrary, our focus in this paper is on the universe of news media which, on the one hand, supposedly create original news content and, on the other hand, may become perfect substitute by using plagiarism.

While there is no variation in copyright we can use to identify the causal impact of copyright on the production of original news content, this article makes empirical progress on the question of copyright by providing new empirical evidence on the extent of copying online. It is a unique attempt at understanding who is producing news, the character of what is produced and the propagation of information in the online world.[13]

The rest of the paper is organized as follows. Section 2 below provides a quick overview of copyright laws in the context of the production of information by news media. In Section 3 we describe the media universe and the content data we use in this paper, and review the algorithms we develop to study the production and propagation of information online. These algorithms are illustrated in Section 4 with the example of a news event. Section 5 provides new evidence on the speed of news dissemination and the importance of copying online, and quantitifies verbatim copying without acknowledgement. In Section 6, we use daily-level variations to investigate the audience share captured by breaking news media outlets. Finally, Section 7 concludes.

2 Copying and news property

Traditionally, copyright violations occurred when someone manually recopied, then reprinted, large portions of someone else's story.[14] 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. A journalist is protected as the author of her articles the same way a writer is protected for the content of the books she writes.

In the United States, copyright law is governed by the Copyright Act of 1976. To receive protection, a work must be original, fixed, and an expression; in particular, the copyright law does not protect facts.[15] 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

[13] Sen and Yildirim (2015) use the case of an Indian English daily newspaper in 2012 to investigate how popularity of online news stories affect editors' decisions. Athey et al. (2013) provides a model of advertising markets for news media.

[14] Our focus here is on the American andh the French copyright laws. Both countries have joined the Berne Convention for the Protection of Literary and Artistic Works, the international agreement governing copyright. The main difference between the two systems comes from the fact that while the system of exceptions is "closed" in France, the U.S. system is "open-ended"(**?**).

[15] Copyright exludes data, as well as ideas and processes (Ginsburg, 2016).

exception of the misappropriation or “hot news” doctrine. The “hot news” doctrine refers to a cause of action for the misappropriation of time-sensitive factual information that state laws afford purveyors of news against free riding by a direct competitor (Balganesh, 2011).[16] The scope of hot news protection has to be considered in light of recent technological changes, however (Fox, 2009). In a world where consumers can quickly and easily access content that is aggregated from many online sources, how long could be the exclusive use time granted to a media outlet?

Moreover, if the reproduction right is not limited to verbatim copying (it also protects against paraphrasing), in all events, the second author must have copied *protected material* (Ginsburg, 2016). This raises the question of the originality of copied articles. What makes a work original is a complicated issue. While it is clear that creativity is not required to make a work original, an open question is the amount of labour required.[17] In particular, in the case of news media, the identity of the information issuer is key. If the issuer is a third party (e.g. authorized journalists simply attend a press conference), then not only multiple media outlets will all copy from the same source, but one can question whether journalists summarizing the press conference are actually using their skill and judgment or simply performing a purely mechanical exercise. We will come back to this point in Section 3.5.

Finally, to violate the exclusive right of reproduction, the second author's copying must be “substantial.” Substantiality of copying depends very much on context; even a small, but qualitatively important, extract from a larger work may be found to infringe, depending on the nature of the copyrighted work and of the portion copied (Ginsburg, 2016). Substantiality can indeed be defined with respect to quality and/or to quantity. Decisions on substantiality must inevitably be ad hoc, as must be decisions on originality.

An illustration: Moneyweb vs. Fin24 A copyright infringement case illustrating the different aspects we just reviewed has taken place in South Africa from 2013 to 2016. The case opposed the financial media websites Moneyweb.co.za (Moneyweb) and Fin24.com (Fin24). Moneyweb brought the application against Fin24 for the breach of copyright after Fin24 copied extensively a number of Moneyweb's articles and published reworked version on its website. Moneyweb contends that Media24 infringed its copyright under the Copyright Act

[16] The “hot news” doctrine was announced in a Supreme Court decision in 1918 (International News Service v. Associated Press); the decision of the Court rested on the idea that – though the Associated Press had no copyright in the facts underlying its 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 (INS, another newswire) allowing it to prevent the INS from publishing stories based on the news the AP had discovered for a limited time. This right existed for the time necessary to allow the AP to make a sufficient profit.

[17] According to the Supreme Court of Canada, *“an original work must be the product of an author's exercise of skill and judgment. The exercise of skill and judgment required to produce the work must not be so trivial that it could be characterized as purely mechanical exercise.”* (CCH Canadian Ltd v Law Society of Upper Canada).

98 of 1978. Three issues were at the centre of the case: (i) the originality of the articles; (ii) the substantiality of the copying; (iii) and section 12(8)(a) of the Act which provides that *"no copyright shall subsist (...) in news of the day that are mere items of press information"*[18]. In the view of the judge (reported in the judgment of the High Court of May 5th, 2016), items of press information include *"all information communicated to the media in material form or subsequently reduced to material form. This would include, but not be limited to, press statements and press interviews concerning "news of the day" which journalists, and anyone else, would be free to use (...) without restriction and without authorisation being required from anyone."*

The judge found that Fin24 infringed Moneyweb's copyright, but only regarding one article of seven that were the subject of the dispute. Regarding the 1st article considered, written by a Moneyweb's journalist after she had participated in a press conference in Parliament, the judge decided that more evidence was required to establish that it is an original work, the article been based on the information that was made available at the press conference.[19] The judge ruled similarly for the second article under consideration, written after a journalist had participated in a conference call with a CEO together with other journalists, and for the third, one written after a Moneyweb's journalist and two other journalists from other media groups had attended a media visit at a restaurant. He claimed that he was *"not able to discern the nature and extent of her contribution"*. Moreover, he notes that the information disclosed at the press conference, given in the conference call, or gathered during the media visit at the restaurant was *"given to the media with full knowledge that [it] would be put into the public domain."*

For the fourth article, the judge also ruled against Moneyweb but based on substantiality. In this case, the source of the article was a press release issued by a company, but the journalist also interviewed company's employees and sourced additional material. However, the judge considered that if the article *"contains more infromation than the press release, the difference is insubstantial. Indeed, it is quite trivial."* Furthemore, in this instance alike for the previously described articles, he considered that the information – in this case the press release – was given to the media knowing that it would be put into the public domain.

To wrap-up, even if decisions on whether an article is protected by copyright law tend to be ad-hoc, four criteria should be considered. First, substantiality: we will study it in Section 5.2 when quantifying the importance of copying online both from the point of view of the copying and of the copied media. Second, originality: to be protected, an article should be

[18] News of the day here mean current news; they are not limited to a 24-hour news cycle.

[19] The judge pointed out that he *"do[es] not know how much of the article is [the journalist]'s own work or simply a repetition of what was said in her presence or contained in a written press release. (...) The fact that she attended the press conference, took notes and asked questions does not mean that the article is not a mechanical repetition of existing material."*

original, which depends on the identity of the information issuer. We determine and discuss this identity in Section 3.5. Third, the time interval between the publication of the original article and the publication of the copying article. According to the "hot news" doctrine, a news media indeed retains exclusive use of its product so long as it has a commercial value. We study the speed of news dissemination in Section 5.1. Finally, the "fair use" in the United States or the "right to quote" in France is subject to the mention of the copied media, including the name of the author. Section 5.3 investigates whether verbatim copying comes with acknoledgment. Overall, we aim at quantifying the importance of verbatim copying falling outside the bounds of copyright law in the online world.

3 Data and algorithms

3.1 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 (or pure internet players, i.e. online-only media outlets); 9 television channels; and 7 radio stations. The news agency is the Agence France Presse (AFP), the third largest news agency in the world (after the Associated Press and Reuters).

We choose this "transmedia" approach because on the Internet, there is a tendency of different media to converge (see e.g. Peitz and Reisinger, 2016). Users interested in news balance and compare multiple sources, independently of the offline format of the media. One cannot infer the offline format of a media by visiting a website, as illustred in Figure 2. On the web, media all offer text, video and photo. We include the AFP despite it does not deliver news straight to individual consumers[20] because it is a key provider of original information in the online world. We think it is essential to consider news agencies when investigating newsgathering and copyright infringement. To the extent of our knowledge, we are the very first to perform such an inclusive empirical analysis of original news production.[21]

[FIGURE 2 HERE]

Using their RSS feeds, we track every piece of content news media produced online in 2013. This content data is from the OTMedia research projet conducted by the INA (*Institut National de l'Audiovisuel* – National Audiovisual Institute, a repository of all French radio

[20] News agencies are based on a Business-to-Business model (they sell news to other media outlets), not on a Business-to-Consumer model. We provide more details on the specifics of the AFP in our companion paper, Cagé et al. (2016).

[21] We do not consider news aggregators and curators, however, nor do we investigate information dissemination on social media. Doing so is well beyond the scope of this paper whose focus is on original news *producers*. On the effect of aggregators, see Athey and Mobius (2012); George and Hogendorn (2012, 2013); Chiou and Tucker (2015); Calzada and Gil (2016).

and television audiovisual archives). For the media outlets whose RSS feeds were not tracked by INA, we complete the OTMedia data by scrapping the Sitemaps of their website. Finally, we get all the AFP dispatches directly from the agency. Merging these datasets, we obtain the universe of all the articles published online by French news media in 2013. These articles contain text and often photos, as well as videos. Our focus here is on text.[22]

Our dataset contains $2,548,634$ documents for the year 2013; around $7,000$ documents on average per day. Figure 3 plots this number on a daily basis. On average, more documents are published during the week, and we observe a drop in this number during the week-end.[23] 72.8% of the documents are from the websites of the print media; 4.7% from radio; 6.6% from television; 12.6% from the AFP and the remaining documents from the pure online media. On average, these documents are 1,865 characters long. Table C.1 in the online Appendix provides summary statistics for the entire sample, as well as by media type (print media, television, radio, pure online media and news agency).

[FIGURE 3 HERE]

In the rest of this section, we review the algorithms we develop to study the production and propagation of information online. In Section 4, we illustrate these different algorithms by taking the example of a specific news event.

3.2 Event detection

Event detection algorithm Using the set of documents previously described, we perform an event detection algorithm to detect media events. This category of algorithm is often refered to as Topic Detection and Tracking (TDT) in the computer science community. These algorithms are based on natural language processing methods (NLP). The goal of online topic detection is to organize a constantly arriving stream of news articles by the events they discuss. The algorithms place all the documents into appropriate and coherent clusters. Consistency is ensured both at the temporal and the semantic levels. As a result, each cluster provided by the algorithm covers the same topic (event) and only that topic. Several approaches have been proposed to solve this problem. Following the conclusions of Allan et al. (2005) who have experienced their TDT system in a real world situation, we adopted a simple but robust approach. Our implementation consists of the following steps :

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

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

1. Each document is described by a semantic vector which takes into account both the headline and the text.[24] A semantic vector represents the relative importance of each word of the document compared to the full dataset. A standard scheme is TF-IDF.[25] As in most of natural language processing methods, we first pre-process our documents by removing very common words (called stop words) and applying a stemming algorithm so as to keep only the stem of the words. We also apply a multiplicative factor of five to the words of the title as they are supposed to describe well the event, resulting in an overweigth in the global vector describing the document.

2. The documents are then clustered in a bottom-up fashion to form the events based on their semantic similarity. The similarity between two documents is given by the distance between their two semantic vectors. As these vectors lie in a very high dimensional space, it is well known that the angle between the vector is a good mesure to asses similarity. We thus use the cosine similarity measure (on TF-IDF/cosine, see the seminal paper by Salton et al. (1975)).

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

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

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

Finally, to ensure consistency, we keep only the events with documents from at least two different media outlets, and with more than 10 documents.

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

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

[25] Term frequency-inverse document frequency, a numerical statistic intended to reflect how important a word is to a document in a corpus. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus. We describe more formally the TF-IDF weight in the online Appendix A.2.

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

We test the quality of the algorithm by running it on a standard benchmark dataset: the Topic Detection and Tracking (TDT) Pilot Study Corpus.[27] The TDT dataset contains events that have been created "manually" by humans. The goal is to compare the performance of the algorithm with the one of humans. To test the performance of our algorithm on the English corpus, we slightly adapt it.[28] First, we use an English stop-word list and an English stemming algorithm. Second, the time frame of the test corpus being wider than ours, the one-day window used to close clusters is not adapted. When testing our algorithm we thus follow the literature (Allan et al., 2005) and close a cluster if 2,500 documents have been treated by the algorithm and none of them has been added to the cluster. We find that the performance of our algorithm is as good as the one of the state-of-the-art algorithms. Furthermore, we find that the main parameter of our implementation, the distance threshold, is the same for this English test corpus and our corpus of French news article. This is very reassuring as to the quality of our algorithm.

As an additional robustness check, we compare our events to those obtained by the Europe Media Monitor (EMM) NewsExplorer.[29] The EMM NewsExplorer provides on a daily basis the top 19 stories of the day. With our event detection algorithm, we match 92% of the stories in their sample.

3.3 News events

We obtain a total number of $24,985$ news events. Events can last more than one day; on average, they last 42 hours (we provide below more details on the length of the events; note that what we define here as the length of an event is the length of the event *coverage* – the time interval between the first and the last article covering the event – not the length of the actual event). The average number of documents per event is 34, and on average, 15 media outlets refer to an event. There are 183 events per day on average, of which 68 new events beginning every day. These events are roughly equally distributed during the year. Figure 4 plots the total number of events per day, as well as the number of new events.

[FIGURE 4 HERE]

Out of the $2,548,634$ documents in the dataset, $850,139$ (33%) are classified in an event (for a daily plot of this ratio, see Figure D.1 in the online Appendix). The remaining 67% of the documents are not classified in events. Note however that the classified documents represent, in terms of characters, 40.4% of the total content produced in 2013. They are indeed longer on average (online Appendix Table C.1).

[27] The goal of the TDT initiative is to investigate the state of the art in finding and following events in a stream of news stories (see e.g. Allan et al., 1998).

[28] There is no similar test corpus in French.

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

An important share of the unclassified documents are from local daily newspapers (while local newspapers represent 57% of the documents in our dataset, they account for 68% of the unclassified documents). Local newspapers indeed cover very local "events" that are not covered either by other local outlets (whose market differs) nor by national outlets.[30] E.g. on January 31th, 2013, a 1,421 character-long article in the local daily newspaper *Nice Matin* about the closure of the dilapidated Plan-de-Grasse – a small city in the Alpes-Maritimes department – city hall. While more than 50% of the documents published by national newspapers, radio, TV and the AFP are classified in events, less than 27% of those published by local newspapers are (see online Appendix Figure D.2 for a plot of these ratios).

Other unclassified documents either correspond to one-off reports or to what Schudson (2015) call "contextual reporting".[31] These articles tend to be relatively long. E.g. on November 20th, 2013, the national daily newspaper *Le Monde* published a 14,968 character-long article revisiting the Rey-Maupin affair (*"Retour sur l'affaire Rey-Maupin : les tueurs de la Nation"*). Florence Rey and Audry Maupin were involved in a shoot-out in Paris in October 1994 following a high speed car chase, causing the deaths of five people. Obviously, in 2013, i.e. 9 years later, this affair is no longer a news event; but lengthy articles revisiting affairs are a good illustration of what contextual reporting is. While on average unclassified documents are smaller than classified documents, the variance of the distribution of their size is higher because they tend to be either very short articles covering local events or much longer leading articles.

In this paper, given that what we are interested in is the propagation of news stories online and the importance of copying, we focus mainly on the $850,139$ articles classified in our $24,985$ events.[32] In Cagé et al. (2016), when estimating the production and demand for news, we take into account both the classified and unclassified documents.

Length of the events On average events, last 41 hours . Figure 5 plots the distribution of this length. Around 10.5% of the events in our sample last less than six hours. These short events are mainly minor news items (e.g. on February 27th, 2013 the death of a woman in Paris poisoned by her leaking boiler, event which was first covered by the AFP at 6:49pm and last-mentioned by the free newspaper *Metro* at 8:02pm.)

Some longer events, lasting over multiple days, are also minor news items, often first

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

[31] *"The journalist's work is less to record the views of key actors in political events and more to analyze and explain them with a voice of his or her own."* Fink and Schudson (2014) classify news articles into five possible categories: investigative, contextual, conventional (conventional stories focus on one-time activites or actions that have occured or will occur within 24 hours), social empathy (social empathy stories describe a person or group of people not often covered in news stories) and other. In earlier work, Tuchman (1980) defined five categories of news: hard, soft, spot, developing, and continuing.

[32] We will nonetheless control for unclassified content when studying the determinants of online audience in Section 6.

covered by a local newspaper before generating buzz at the national level. E.g. at 6:05am on January 1st, 2013, *Le Dauphiné Libéré* wrote the story of robbers returning to the jewelry store the loot they stole together with... a chocolate box! Only two other media outlets (*Le Parisien* and *Direct Matin*) covered the story on January 1st, while it got increasing coverage on January 2nd after an article published on the website of the radio RTL. Finally, this event is last-mentioned on January 3rd by the website of France Télévision. We discuss in Section 5.1 the different profiles of news events.

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

Hence we cannot infer the nature of an event directly from its length. While the length of the coverage sometimes reflects the actual length of the event, it may also simply stem from editorial choices of media outlets. In Section 3.5, we resort to manual coding to improve our understanding of the specific nature of the different events in our sample.

[FIGURE 5 HERE]

Topic of the events We classify the events according to their topic. In order to do so, we rely on the metadata associated with AFP dispatches included in the event. There is at least one AFP dispatch in 93% of our events. (We do not define the topic of the remaining events.)

AFP uses the 17 IPTC classes to classify its dispatches.[33] These top-level media 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. Moreover, in 95% of the events it covers, the AFP also provides information on sub-categories (e.g. "crime" is a sub-category of "Crime, law and justice", and "agriculture" a sub-category of "Economy, business and finance"). An event can be associated with more than one top-level media topic ($8,428$ events in our dataset are). E.g. when on January 1, 2013, the US Senate passed a compromise bill to eliminate the fiscal cliff, this event was classified by the AFP both in the "Economy, business and finance" category (with a sub-category "macroeconomy") and in the "Politics" category (with a sub-category "government").

[33] 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.5).

Figure 6 plots the share of events associated with each media topic. (Given that some events are associated with more than one topic, the sum of the shares is higher than 100%.) The vast majority of events are about "Politics" (nearly 35%), "Economy, business and finance" (26%) and "Crime, law and justice" (around 22% of the events). "Sport" comes fourth, appearing in 11% of the events. The other topics like "Weather", "Education" or "Science and technology" have much less importance. This does not mean that there is no article related to these topics, but that these topics are not associated with any *events*.

[FIGURE 6 HERE]

3.4 Timeline and plagiarism detection

Timeline 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 that breaks out the story, and then rank the other outlets. Using the publication time, we also document how long it takes the media to cover the story.

It is not because a media outlet is talking about a story that it 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 original content in each document compared to the content of all the documents published earlier in the event.

Plagiarism detection algorithm Consider a document n. To compute the original content of this document we proceed as follows. We develop a plagiarism detection algorithm to track efficiently Identical Portions (IP) of text between documents. $\forall n' \in [1, n[, IP(n, n') = \mid n \cap n' \mid$.[34] We then determine for each document the portions of text that are identical to content previously published by all the documents out earlier in the event, and isolate the original content in the document. The originality rate of a document is defined as the share of the document's content (in number of characters) which is original.

For all the documents classified in events, we also compute the copy rate of the document with respect to each document previously published within the event in a bilateral way (note

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

that if a media outlet only reproduces the content of a single document previously published in the event, the copy rate will be equal to 1 - the originality rate). We do so to identify, if any, the document previously published within the event whose share present in the document is the highest.

Finally, we trace back each portion of text to its first occurence in the event. It allows us to determine for each document (in particular breaking news), the number of times it is copied and the share of the (original) story which is ultimately copied.

3.5 Exclusive vs. non-exclusive news events

In the case of a government press release giving rise to a number of articles, the first media covering the story cannot be considered as a news breaker providing exclusive news. We may also overestimate verbatim copying by attributing the release to the first media and counting as copy the reproduction of the release by other outlets. To deal with this issue, we code manually all the news stories in our sample to isolate the stories that are the results of a piece of original reporting by one outlet.[35] We call these stories exclusive news events. The remaining stories are either non-exclusive news events or short news items with multiple witnesses.

To distinguish between these three types of news events, we investigate the nature of the information issuer. We first define six categories of non-exclusive news events, i.e. news events where the original information can be considered to be in the public domain, and was not produced by the media itself. This includes news events where the information issuer is the government, the police, companies or non-governmental organizations, as well as cultural and sport events:

- Government and politicians: e.g. government's communiqués and releases; press conferences; political interviews.
- Police and justice: e.g. police department press releases; court rulings.
- Companies: e.g. financial reporting releases; sales figures releases.
- Non-governmental organizations: e.g. official statistics publications; scientific research reports; publication of stock prices.
- Cultural events and celebrities: e.g. celibrity press conferences; awards ceremony.
- Sports events: e.g. soccer games; Olympics.

[35] To ensure consistency, all the stories have been coded twice, by two different Research Assistants. The classification of the stories for which the Research Assistants disagree to begin with has then been discussed at lenght by the authors. As of today, we have been able to code 10,973 out of the 24,985 news stories, and are working on finishing the coding. Hence the results we present using the nature of the news events should be treated cautiously.

We then define two categories of exclusive news events:

- Investigative stories: stories for which the news originate from *"the revelation of new facts"* that someone wants to keep secret by the media outlet (Hamilton, 2016)[36]; this involves substantial in-depth reporting, whereby media outlets are playing watchdog.
- (Non-investigative) Reporting stories: stories for which the news originate from the exposition of facts by the media outlet, but with limited in-depth reporting; typically these are facts that nobody tries to hide and which the media decides to present to the public.

What distinguishes both types of stories is the amount of in-depth expository reporting involved. E.g. the NSA spying scandal revealed by *Le Monde* on October 21, 2013 and described in Section 4 is an example of what an investigative story is. On the contrary, we classify as (non-investigative) reporting story a series of articles on Chinese divorcing to skip property tax or an event dealing with a Japanese man vacations on Syrian front lines.

Finally we define short news items with multiple witnesses as our ninth category. These are news events for which there are multiple witnesses: e.g. public protest; murder in public space; terrorist attack; plane crash.[37] The journalists may not be the first to report the story – e.g. due to breaking news alerts on social media – but they are the first to provide "reliable" information on the story.

We find that non-exclusive news events represent around 80% of the events, as illustrated in Figure 7. The most important information issuer is the government and politicians (nearly one third of the events), followed by the police and the justice. 10% of the events are sport events where the original information can be considered to be in the public domain.

Short news items with multiple witnesses account for 10.7% of the events, and exclusive news events for around 6%. Only 1.3% of the events in our sample can be considered as investigative stories. While this number may seem low, it is in fact in line with previous findings in the literature. E.g. examining a sample of front-page stories at the *Milwaukee Journal Sentinel*, the *New York Times*, and the *Washington Post*, Fink and Schudson (2014) find that investigative reports only represent 1% of the stories in 2003.[38] In a content analysis of over 33,000 stories aired between 1998 and 2001 on 154 local television stations, Rosenstiel et al. (2007) show that station-initiated investigations accounted for .62% of all political stories and 1.10% of nonpolitical stories.[39]

[FIGURE 7 HERE]

[36] *"Investigative reporting involves original work, about substantive issues, that someone wants to keep secret."*

[37] We include weather-related events in this category (but they only represent .6% of the events).

[38] And even less before: 0% in 1955, 0% in 1967, 1% in 1979, and 3% in 1991.

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

3.6 Citation detection

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

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

We obtain a total of 388, 035 citations in 2013, 57% of which being references to the AFP. We study citation patterns in Section 5.3.

4 The propagation of information: an 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. *Le Monde* published almost simultaneously (at 06:01:23 am) a second article on the topic, entitled *"L'ampleur de l'espionnage mondial par la NSA"* ("Inside the NSA's web of surveillance"). The first article is 5,379 characters long, the second 5,768, and there is no internal copy between the two articles (both are 100% original). Finally, thirty minutes later (at 06:32:52am) *Le Monde* published a third article covering the story, also entirely original, and entering into the details of the surveillance (*"Les services secrets américains très intéressés par Wanadoo et Alcatel-Lucent"*).

30 seconds after the publication of the 1st article by *Le Monde*, at 06:01:43 am, the AFP published a dispatch on the same topic (*"La NSA a recolté des millions de données en France"*). (The very high reactivity of the AFP in this specific case comes from the fact that *Le Monde* gave the news to the AFP but with an embargo.) The AFP classified the story using 3 different IPTC topics: (i) media (included in "Arts, culture and entertainment"), (ii) computing and information technology (included in "Economy, business and finance"), and (iii) "Politics".

The AFP dispatch was very short (494 characters) and 40% of its content was copied-and-pasted from *Le Monde*'s original article, as it appears in Figure 8a which illustrates our plagiarism detection algorithm. At 06:01:48 am (35 seconds after the publication of the first article by *Le Monde*), the AFP published a second and longer dispatch (3,177 characters). 75% of the content of this dispatch was copied-and-pasted from *Le Monde*'s article (Figure 8b). In both cases, the AFP refers to *Le Monde* a number of times as the source of the information (*"révèle lundi le quotidien Le Monde"*, *"indique Le Monde"*, *"d'après Le Monde"*,...).

Half an hour later, the first non-news agency media outlet to report online on this extensive electronic eaversdropping 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 the AFP (Figure 8c). Both media outlets refer to *Le Monde* as the source of the information.

[FIGURE 8 HERE]

Overall, the story broken out by *Le Monde*, gave rise directly to 119 articles (115 excluding *Le Monde*) by 52 media outlets. Just within three hours after the publication of the first article by *Le Monde*, 53 articles related to the event had already been published. In the online Appendix Figure D.3 we illustrate the propagation of the information during the first two hours following the publication of the scoop by *Le Monde*. Beginning at 6:30am, competing media outlets react quickly to the breaking news and cover it on their website. However, the articles they publish tend not to be original, to the exception of *Le Figaro* (sub-Figure D.3a). But if *Le Figaro*'s article is original, it is also very short (680 characters) (sub-Figure D.3b). In other words, while competing media outlets are fast to react, they do not provide additional original information.

In this specific example, copy goes with acknowledment. Out of the 115 articles written on the story by media outlets other than *Le Monde*, 100 (nearly 87%) refer to *Le Monde* as the source of the information. The AFP also receives some credit for the story, in part because a number of articles have been written with the AFP (*"avec AFP"*). Finally, at 8am, in a rapid reaction, Interior Minister Manuel Valls spoke out againt US spying on the radio station Europe 1; thanks to Valls' reaction, the radio station also receives a number of citations. Figure 9 illustrates our reference detection algorithm.

[FIGURE 9 HERE]

This example also illustrates how our event detection algorithm organizes the news articles into coherent clusters. It indeed generates on the same day two other news events linked to but distinct from the one we just discussed. The first distinct news event covers the French foreign minister, Laurent Fabius, summoning the US ambassador to the foreign office; the second one the Spanish Prime Minister, Mariano Rajoy, summoning in this turn the US ambassador.

5 Empirical analysis

5.1 The speed of news dissemination

In this Section, we study the speed of news dissemination online.[40] We construct the typical timeline of a news story. More precisely, we investigate how fast news would be delivered to readers of different media outlets after having been published first on the website of the news breaker.[41]

Studying the speed of news dissemination is of interest because, as we highlighted in Section 2, the "hot news" doctrine relies on the idea that a news media retains exclusive use of its product so long as it has a *commercial value*. 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 171 minutes for an information published by a media oulet to be published on the website of another oulet. But this average masks a lot of heterogeneity. In half of the cases, it takes less than 22 minutes, of which less than 230 seconds in 25% of the cases and less than 4 seconds in 10% of the cases. (In our companion paper (Cagé et al., 2016), we use daily readership data to evaluate the extent to which readers value learning about news stories sooner.)

Table 1 reports the average reaction time depending of the offline format of the news breaker. If the news agency (AFP) is the first media outlet to publish an information (which is the case for half of the events), then the reaction time is shorter. When the AFP is the news breaker, we find that the second media outlet covers it in 116 minutes on average, but in 10 minutes in half of the cases and in 1 second or less in 10% of the cases. This rapidity comes from the fact that media outlets receive the news directly from the AFP; they don't have to monitor it the way they monitor what is published on their competitors' website. Furthermore, a number of media outlets have automatized the posting of prepackaged AFP content. In other words, AFP content of their choice is automatically integrated into their website.

We find that the reaction time is the highest when the news breaker is a pure online media. Even if demonstrating it is beyond the goal of this article, a possible explanation is that pure online media may suffer from a lower reputation. Hence legacy media may want to wait for multiple sources before covering an event broken by these new media.

[TABLE 1 HERE]

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

[41] Unfortunately, we don't have information on when the actual news event takes place; the only information we have is the exact time at which the event is reported for the first time by a media outlet in our sample, and then we know the exact publication time of all the articles related to this event until the last media outlet reports about it.

We also investigate how the reaction time varies depending on the nature of the news events. We show that the reaction time is lower for short news items with multiple witnesses than for non-exclusive and exclusive news events.[42] However, the difference are not statistically significant.

Profile of news events Finally, we study the distribution of the documents within events. To do so, we split all the events into 25 bins of similar length (hence the length of the bins is higher for longer events) and we compute the share of the number of documents in each bin. Figure 10 plots this distribution. The highest share of documents related to an event is published in the first bin and then this number is decreasing through time (it is divided by more than two between the first and the second bin). It increases again in the last three bins, especially the penultimate and the 25th one, where on average there are more documents published than in the third bin.

[FIGURE 10 HERE]

Interestingly, this pattern of document publication reflects two different kinds of events. On the one hand, a lot of events are characterized by immediacy in news dissemination: as soon as a story has been published by a media outlet, other media outlets cover it quickly. From which the high share of documents in the first three bins – especially in the very first one – and the flat right tail. On the other hand, other events present the exact opposite profile: a flat left tail and the majority of the documents published in the last bin. Those are events which need a "gatekeeper"[43]: one media outlet covers a news story in the first place, but for it to be covered more broadly and disseminated to the public, it needs to be covered by a high-reputation outlet, very often a national one. The robbers returning their plunder with chocolate box story we described above instantiates this news profile. We illustrate these different patterns of event profile in the online Appendix Figure D.4 where we cluster the events depending on the share of the documents in each bin.

5.2 The importance of copying online

We now turn to an estimation of the originality of the articles published online in 2013. This is a key question because the high reactivity of the media we just discussed may actually come from the use of plagiarism.

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

[43] The gatekeeping theory was first proposed by Kurt Lewin (a German psychologist) in 1943. Gatekeeping is the information managing process by media for selecting information to be published.

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

On average, the originality rate of the documents classified in events is equal to 36%.[44] In Figure 11a, 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 19% of the documents) and another peak for the 100%-original articles (21% of the documents). The median is 14%. In other words, to the exception of the documents which are entirely original, the articles published within events consist mainly of verbatim copying: 56% of the articles classified in events have less than 20% originality.

We study how the originality varies with the nature of the news event. Figure 11b plots the Kernel density estimates. We find that articles published in non-exclusive news events tend to have a lower originality rate. This is not surprising (and reassuring as to our manual coding of the events): non-exclusive news events are indeed events derived from information that is in the public domain (e.g. a government press release) and media outlets tend to reproduce this information as it is.

[**FIGURE 11 HERE**]

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

We find that on average documents include content from 3.7 documents previously published in the event (the median is 2).

Internal vs. external copying Verbatim copying can be either "internal" or "external". A media outlet can indeed copy-and-paste content from documents it has itself previously

[44] Given that documents are of different lengths, we also compute the ratio of original content in the dataset over the total content. We find that the share of original content is equal to 32%. 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.

published (in particular when it is updating previous versions of the same article, for example adding new elements). Conditional on publishing at least one document related to the event, half of the media outlets publish at least 3 documents in the event, 2 when we exclude AFP. (The AFP publication strategy is characteristic of the work of news agencies which consists in publishing first short dispatches and then by supplementing them with more details during the day. Hence the median number of dispatches published by the AFP in an event – conditional on the AFP covering the event – is 12.)

We find that out of the $850,139$ documents classified in events, $601,130$ (70.7%) present at least some external copy. On average, documents include content from 3.1 documents previously published in the event by competing media outlets. If we sum up the external copied content, we obtain an external copy rate of 51.3% (84.1% conditional on copying).

Excluding the AFP When considering copyright law, one needs to distinguish between content copied from the AFP (the news agency) and content copied from other media outlets. All the media outlets that are clients of the AFP are indeed allowed to reproduce the AFP content in its entirety. $457,168$ documents present at least some external verbating copying from a media outlet other than the AFP, and on average, documents include content from 1.96 documents published by competing media outlets other than the AFP. If we exclude content copied from the AFP, we find that the average external copy rate is 24.9% (41.3% conditional on copying).

The news agency AFP also tends to reproduce the content published by other media outlets in its dispatches so that to circulate the information broken by others. If we exclude the documents published by the AFP from our analysis, we find that the average external copy rate is 57.2%, i.e. slightly higher to what we obtained including the AFP.

Share of the original story that is copied Does this external verbating copying fall within the bounds of copyright law? To violate the exclusive right of reproduction, copying should be "substantial". Substantiality can be defined either from the point of view of the copying media, or from the one of the copied media. We compute the share of each document which is copied. On average, we find that each document is copied by 3.7 documents published later in the event, 3.1 if we exclude internal verbating copying (the media being 0). If we focus on external verbating copying and sum up the portions of the documents that are at least reproduced by one external media outlet, we find that on average the share of a document that is copied is equal to 9.15%.

The majority of the documents are not copied, however. If we restrict our analysis to copied documents, we obtain that the share of a document that is copied by at least one external media outlet is 21.2% on average. If, as before, we exclude documents published by the AFP (and that are thus not protected by copyright law) from our sample, we find that this share is equal to 12.5% on average.

This share varies strongly depending on the publication rank of the document. Online Appendix Figure D.6 plots the average share of a document that is copied by at least one external media outlet depending on the publication rank of the document. We find that for breaking news documents (documents that are first published within the event), this share is above 60%, 30% when we exclude documents published by the AFP. It then decreases to nearly 25% (15%) for the second document and converge rapidly to around 5%.

Finally, we investigate whether the share of the original story that is copied varies depending on the nature of the news events. We show that the share of the breaking news document that is copied is higher for exclusive news events, at 66.6%.

As we noted in Section 2, substantiality is not enough to decide whether an article is protected by copyright law. We ultimately need to analyze whether media outlets refer to the media outlets they copy.

5.3 Credit and citations patterns

In France, under certain conditions, media outlets are allowed to reproduce content originally published by their competitors, but the "right to quote" (*"exception de courte citation"*) is subject to the mention of the source. In this Section, we study the extent to which the occurences of verbatim copying we identified above come with acknowledment. In other words, we analyze whether media outlets tend to name the outlets they copy.

We perform this analysis at the event level. For each document presenting at least some external verbatim copying, we investigate whether it refers to the media oultet(s) it copies as the source of the information. We find that in only 8.66% of the cases, the document mentions the competing media outlet it copies as the source of the information. Moreover, if we exclude verbatim copying from documents published by the AFP, we obtain that the probability of crediting decreases to 1.66%.

When a media outlet reproduces content from multiple documents, it may choose only to refer to the competitor whose document it copies the most. We study the extent to which it is the case, and find that documents refer to media oultets whose document they copy the most in 16.77% of the cases. However, if we exclude all the cases where the most copied media is the AFP, we show that this probability drop to 2.47%.

Rather than refering to the outlets they copy as the source of the information, media outlets may choose simply to credit the breaking news outlet. We find that they do so in only 18.6% of the cases, 7.2% when the breaking news outlet is not the AFP. To which extent does it vary with the nature of the news event? We show the media refer more to the breaking news outlet when the news event is exclusive: 27.8% of the outlets refer to the breaking news outlet as the source of the information in that case, 14.03% if the breaking news outlet is not the AFP. Moreover, this propensity is higher if we focus on investigative news stories: 38.98% of the documents mention in this case the news breaker, and still 36.2% when the news breaker is another outlet than the AFP.

Wrap-up We note above that to measure the number of verbatim copying occurences that fall outside the bounds of copyright law, four criteria should be considered: substantiality, originality, speed of news dissemination, and acknowledgment. If we should insist on the fact that decisions on whether copyright is violated are inevitably ad hoc, we can nonetheless conclude from this section that copyright is frequently violated in France. Not only news are disseminated within the click of a mouse (in less than 230 seconds in 25% of the cases), but high reactivity comes with a substantial use of plagiarism: 70% of the documents in our sample present at least some external copy. Moreover, conditional on copying, the average external copy rate of the copying media is 84%, and conditional on being copied, 21% of the content of the original story is copied. Finally, copying media hardly ever refer (only in 8% of the cases, and in less than 2% when the copied media is not the AFP) to the media copied as the source of the information.

6 Breaking news and online audience

Do breaking news outlets nonetheless benefit from their investment in newsgathering? To answer this question, we investigate the relationship between the production of original information and the audience of the websites. In particular, we study the share of total views captured by breaking news media on a given day.

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 each website, we have information on the number of unique visitors, the number of visits and the number of page views.[45] This information is available at the daily level.[46] To the extent of our knowledge, we are the first to compute and use such detailed audience data. The average daily number of unique visitors is $250,000$. Table 2 provides summary statistics for these variables. Data sources are described in details in the online Appendix.

[TABLE 2 HERE]

[45] We have audience data for 58 out of the 85 media outlets in our sample (the AFP being based on a Business-to-Business model, it does not deliver news to individual consumers on its website). Websites whose audience is very small are indeed not monitored by the OJD.

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

Cross-sectional estimation The average number of unique visitors varies strongly from one media to the other, e.g. from less than 20,000 daily unique visitors on average for the website of a small local daily newspaper like *La République du Centre*, to more than one million for national newspapers like *Le Monde* or *Le Figaro.* Figure 12 shows the raw correlation between the total number of news broken by a media outlet in 2013 and its average number of unique visitors. This correlation is positive.

[FIGURE 12 HERE]

We first perform a cross-sectional estimation by computing the average quantity of information produced daily by the media outlets in 2013. Equation 1 describes our identification equation:

$$\text{average daily online audience}_n = \alpha + Z_n^{'}\beta + \gamma_{\text{media}} + \epsilon_n \tag{1}$$

where n index the media outlets and the dependent and explanatory variables are in log. The outcome of interest, average online audience$_n$, is the average daily number of unique visitors in 2013. The vector of explanatory variables $Z_n^{'}$ includes (i) the content not classified in events (measured using the number of characters); (ii) the content classified in events that we separate between the original content and the non-original content; and (v) the number of news stories broken. γ_{media} are media category (newspapers, television, radio, pure online media) fixed effects. Table 3 presents the results. We find that online audience is positively correlated with all our different measures of information production when taken separately, but that once they are included all together, only the production of *original* content is positively and statistically associated with audience. A 1% increase in the production of original content leads to a 0.96% increase in the number of unique visitors. Moreover, we find a negative and statistically significant effect of the production of non-original content.

This positive correlation may be driven by the fact that media outlets whose audience is higher may also be more likely to invest in newsgathering, e.g. thanks to higher advertising revenues. To identify a causal effect, we thus exploit daily-level variations.

[TABLE 3 HERE]

Daily-level analysis We use the daily audience data to investigate how online audience varies with the number of breaking news, and in particular the extent to which the number of unique viewers is higher for the breaking news media outlets.[47] Equation 2 describes our prefered identification equation (the observations are at the media outlet-day level):

[47] In our companion paper (Cagé et al., 2016), we estimate structurally the demand function for news and discuss a number of alternative factors that may affect demand, in particular low swtiching costs online and political preferences.

$$\text{unique visitors}_{dn} = \alpha + \mathbf{Z}'_{\mathbf{dn}}\beta + \gamma_n + \delta_d + \epsilon_{dn} \quad (2)$$

where n index the media and d the date and we use the log of the variables. $\mathbf{Z}'_{\mathbf{dn}}$ is a vector that includes measures of the production of information of media outlet n on day d. γ_n and δ_d denote fixed effects for media outlets and day, respectively. Standard errors are clustered by media.

Table 4 presents the results. When considering the content produced, we find that the only statistically significant determinant of the audience share is the original content (classified in events). An increase by one percent on a given day in the production of original content leads to a .018% increase in the number of unique visitors. This effect is statistically significant but economically small. We find no effect of the copied content classified in events, nor of the non-classified content.

When we turn to breaking news, we find that the number of news broken during the day has no statistically significant effect on the number of unique visitors. However, if rather than considering the number of breaking news we study the effect of an indicator variable taking on the value 1 if the outlet has broken at least one news story during the day and 0 otherwise, we obtain a positive and statistically significant effect. Breaking at least one news story during the day increases the number of unique visitors by $1,4\%$.

[TABLE 4 HERE]

Estimates are robust to fixed effects Poisson regressions to control for the count data nature of our breaking news variables. Online Appendix Table B.1 presents the results with standard errors that are robust to serial correlation across media.

7 Conclusion

The combination of technological progress in digitization and computer networking raises concern about the economic viability of traditional forms of news production (Varian, 2005). In 2010, the Federal Trade Commission (FTC) in the United States issued a discussion paper outlining the enactment of "Federal Hot News Legislation" as a proposal aimed at reinventing journalism and adressing newspapers' revenue problems. But now that digital information is very easy to copy and distribute, copyright laws may become almost impossible to enforce.

This paper is a first attempt at quantifying the extent of copyright violations in the online world. It uses a unique dataset covering the online production of information of the universe of French news media during the year 2013. We investigate the speed of news dissemination and distinguish between original information production and copy-and-paste. We also investigate whether copying media refer to the media copied as the source of the information. We find that copyright is frequently violated in France. Moreover, using daily-level variations, we show that producing original information does not affect the audience share.

Should copyright laws be strengthen to give the news industry more protection? This paper does not aim at answering this extremely complicated question. More research is still needed; in particular, it would be of interest to extend the analysis to news aggregators (e.g. Google or Facebook), search engines or other "parasitic" distribution mechanisms that are capturing an increasing share of the audience with the growth of the Internet. Moreover, copyright is a second-best solution to intellectual property provision. As highlighted by Hamilton (2004), *"once the data have been created (...) the tension remains that allowing someone to charge more than zero for the information will exclude some consumers who value the information more than its marginal cost of distribution."* We hope this paper will inform the debate on the level of news property online, however. This sounds all the more important that while the number of journalists is collapsing – in the United States, the number of daily newspaper journalists decreases by more than 20,000 between 2007 and 2015 (see e.g. Angelucci and Cagé, 2016) – journalists are a key input for the production of original information in the online world (Cagé et al., 2016).

References

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

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

Angelucci, Charles and Julia Cagé, “Newspapers in Times of Low Advertising Revenues,” CEPR Discussion Paper 11414, CEPR 2016.

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

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

Balganesh, Shyamkrishna, ““Hot News”: The Enduring Myth of Property in News,” *Columbia Law Review*, 2011, *111* (3), 419–497.

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

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

Boldrin, Michele and David Levine, “The Case against Intellectual Property,” *The American Economic Review*, 2002, *92* (2), pp. 209–212.

_ and _ , “Perfectly competitive innovation,” *Journal of Monetary Economics*, 2008, *55* (3), 435–453.

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

_ , Nicolas Hervé, and Marie-Luce Viaud, “Estimating the Production and Demand for Online News: Micro-Level Evidence from the Universe of French News Media,” Working Paper 2016.

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

Chiou, Lesley and Catherine Tucker, "Content Aggregation by Platforms: The Case of the News Media," Working Paper 21404, National Bureau of Economic Research 2015.

Dellarocas, Chrysanthos, Zsolt Katona, and William Rand, "Media, Aggregators, and the Link Economy: Strategic Hyperlink Formation in Content Networks," *Management Science*, 2013, *59* (10), 2360–2379.

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

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*, jun 2007, *97* (3), 713–744.

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

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

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

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

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

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

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

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

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

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.

Jeon, Doh-Shin and Nikrooz Nasr Esfahani, "News Aggregators and Competition Among Newspapers in the Internet," Working Papers 12-20, NET Institute 2012.

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

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

Nagaraj, Abhishek, "Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project," Working Paper 2016.

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

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

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

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

Rosenstiel, T, M Just, T Belt, A Pertilla, W Dean, and D Chinni, *We Interrupt This Newscast: How to Improve Local News and Win Ratings, Too*, Cambridge University Press, 2007.

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

Salton, G, A Wong, and C S Yang, "A Vector Space Model for Automatic Indexing," *Commun. ACM*, 1975, *18* (11), 613–620.

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

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

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

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

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

Varian, Hal R, "Copying and Copyright," *The Journal of Economic Perspectives*, 2005, *19* (2), 121–138.

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

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

Table 1: Reaction time

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

	Mean	sd	Median	Min	Max	Obs
Reaction time (in minutes)	171	395	22	0	25,797	24,959
If news breaker is						
Print media	224	474	58	0	25,797	8,892
Television	232	398	53	0	2,222	1,205
Radio	235	393	70	0	2,402	1,016
Pure internet player	400	492	178	0	2,164	514
News agency	116	316	10	0	2,779	13,332

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

				Differences		
	Non-exclusive	Exclusive	Mult wit	Non-exc vs. Exc	Non-exc vs. Mult	Exc vs. Mult
	mean/sd	mean/sd	mean/sd	b/t	b/t	b/t
	165.1	181.6	155.8	16.5	-9.3	25.9
	(444.4)	(383.1)	(337.3)	(1.0)	(-0.7)	(1.5)
Obs	9,104	698	1,171	9,802	10,275	1,869

Notes: The Table gives summary statistics for the reaction time (in minutes).

Table 2: Summary statistics: Media outlets

	Mean	Median	sd	Min	Max
Online audience					
Number of unique visitors	247,349	107,856	382,574	3,689	2,031,580
Number of visits	339,097	156,735	542,470	4,650	2,945,172
Number of page views	1,589,726	626,841	2,938,612	0	15,203,845
Audience share	1.66	0.72	2.57	0.02	13.65
Content (nb of characters)					
Total content not classified	40,859,417	16,464,426	138,526,968	1,301,931	1,066,161,792
Total content classified	25,261,949	16,661,134	27,688,881	51,377	114,474,544
Total original content	8,355,501	4,778,750	8,375,425	51,377	32,068,514
Total non-original content	16,906,449	8,187,566	23,161,788	0	103,577,880
Number of breaking news	176	83	231	0	1,075
Observations	58				

Notes: The Table gives summary statistics. Year is 2013. Variables are values for the media outlets for which we have audience data. The observations are at the media outlet/year level.

Table 3: Cross-sectional analysis: Number of unique visitors and breaking news

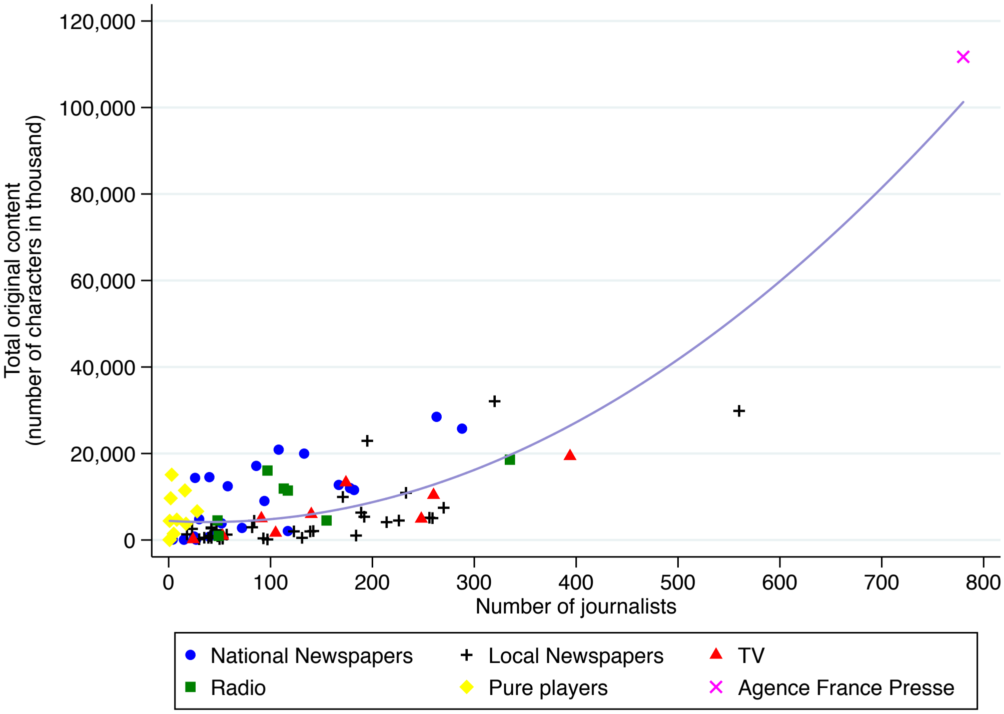
	(1)	(2)	(3)	(4)	(5)	(6)
Content not classified (log)	0.593***					0.029
	(0.161)					(0.121)
Content classified (log)		0.441***				
		(0.091)				
Original content (log)			0.672***			0.964***
			(0.109)			(0.195)
Non-original content (log)				0.240***		-0.228***
				(0.064)		(0.083)
Number of breaking news (log)					2.279***	0.036
					(0.425)	(0.676)
Category FE	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.29	0.33	0.52	0.20	0.33	0.58
Observations	58	58	58	58	58	58

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependant variable is the log of the average daily number of unique visitors. Robust standard errors in parentheses. Models are estimated using OLS. The unit of observation is a media outlet. Variables are described in more details in the text.

Table 4: Daily-level analysis: Unique visitors and breaking news (log-log estimation)

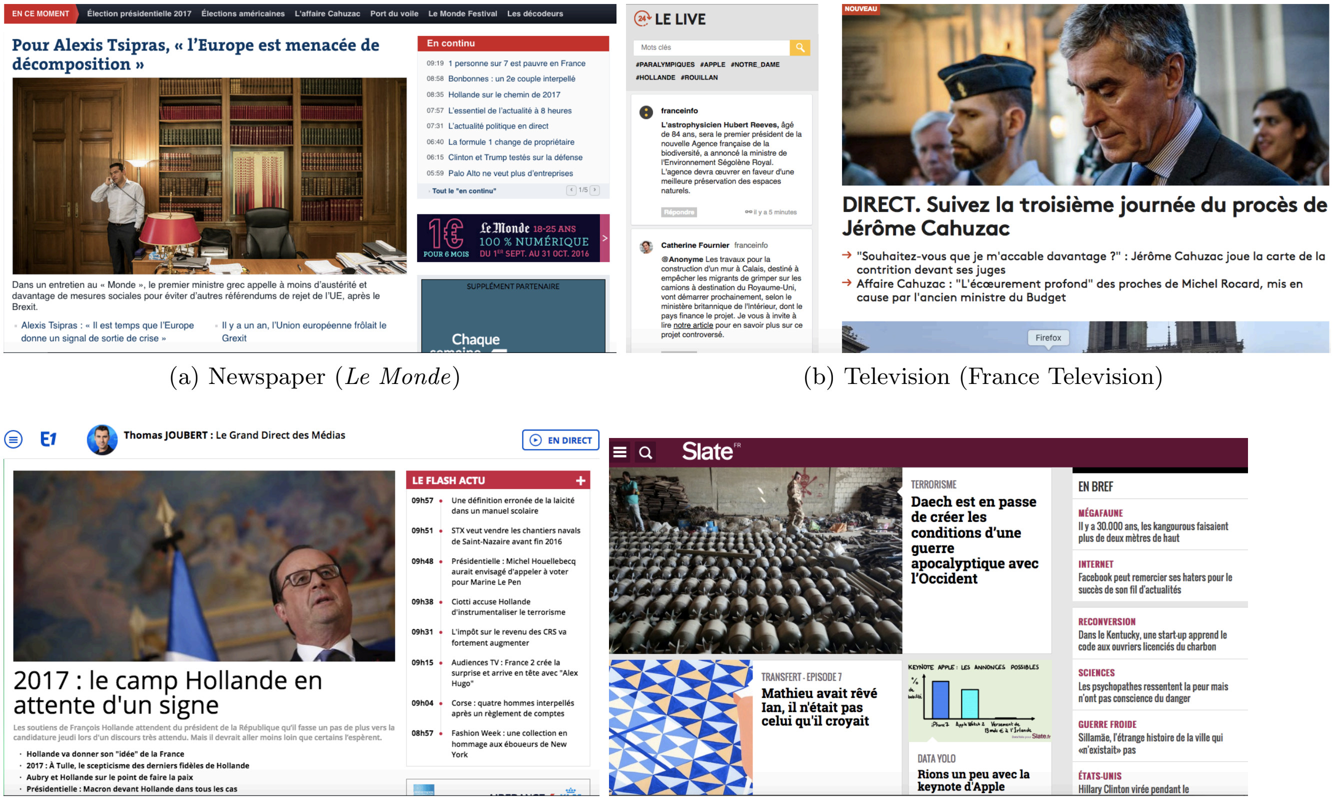
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Content not classified (log)	0.010						0.008	0.008
	(0.006)						(0.005)	(0.005)
Content classified (log)		0.022***						
		(0.005)						
Original content (log)			0.020***				0.018***	0.018***
			(0.005)				(0.005)	(0.005)
Non-original content (log)				0.002			-0.000	-0.000
				(0.002)			(0.002)	(0.002)
Number of breaking news (log)					0.020		0.011	
					(0.012)		(0.011)	
Dummy breaking news						0.022***		0.014**
						(0.007)		(0.007)
Media outlets FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
R-sq	0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17
Observations	16,390	16,390	16,390	16,390	16,390	16,390	16,390	16,390
Clusters (media)	58	58	58	58	58	58	58	58

Notes: * p<0.10, ** p<0.05, *** p<0.01. The dependant variable is the log of the number of unique visitors. Standard errors in parentheses are clustered by media. Models are estimated using OLS estimations. The unit of observation is a media outlet-day. All the estimations include media outlets and date fixed effects. Variables are described in more details in the text.



Notes: The Figure shows the correlation between the total original content produced by each media outlet in 2013 and the number of journalists. Source: Cagé et al. (2016).

Figure 1: Production of original content and number of journalists



Pour Alexis Tsipras, « l'Europe est menacée de décomposition »

Dans un entretien au « Monde », le premier ministre grec appelle à moins d'austérité et davantage de mesures sociales pour éviter d'autres référendums de rejet de l'UE, après le Brexit.

- Alexis Tsipras : « Il est temps que l'Europe donne un signal de sortie de crise »
- Il y a un an, l'Union européenne frôlait le Grexit

En continu

09:19 1 personne sur 7 est pauvre en France
08:58 Bonbonnes : un 2e couple interpellé
08:35 Hollande sur le chemin de 2017
07:57 L'essentiel de l'actualité à 8 heures
07:31 L'actualité politique en direct
06:40 La formule 1 change de propriétaire
06:15 Clinton et Trump testés sur la défense
05:59 Palo Alto ne veut plus d'entreprises

Tout le "en continu"

(a) Newspaper (*Le Monde*)

franceinfo

L'astrophysicien Hubert Reeves, âgé de 84 ans, sera le premier président de la nouvelle Agence française de la biodiversité, a annoncé la ministre de l'Environnement Ségolène Royal. L'agence devra œuvrer en faveur d'une meilleure préservation des espaces naturels.

Catherine Fournier franceinfo

@Anonyme Les travaux pour la construction d'un mur à Calais, destiné à empêcher les migrants de grimper sur les camions à destination du Royaume-Uni, vont démarrer prochainement, selon le ministère britannique de l'Intérieur, dont le pays finance le projet. Je vous à invite à lire notre article pour en savoir plus sur ce projet controversé.

DIRECT. Suivez la troisième journée du procès de Jérôme Cahuzac

→ "Souhaitez-vous que je m'accable davantage ?" : Jérôme Cahuzac joue la carte de la contrition devant ses juges
→ Affaire Cahuzac : "L'écœurement profond" des proches de Michel Rocard, mis en cause par l'ancien ministre du Budget

(b) Television (France Television)

Thomas JOUBERT : Le Grand Direct des Médias

2017 : le camp Hollande en attente d'un signe

Les soutiens de François Hollande attendent du président de la République qu'il fasse un pas de plus vers la candidature jeudi lors d'un discours très attendu. Mais il devrait aller moins loin que certains l'espèrent.

- Hollande va donner son "idée" de la France
- 2017 : À Tulle, le scepticisme des derniers fidèles de Hollande
- Aubry et Hollande sur le point de faire la paix
- Présidentielle : Macron devant Hollande dans tous les cas

LE FLASH ACTU

09h57 Une définition erronée de la laicité dans un manuel scolaire
09h51 STX veut vendre les chantiers navals de Saint-Nazaire avant fin 2016
09h48 Présidentielle : Michel Houellebecq aurait envisagé d'appeler à voter pour Marine Le Pen
09h38 Ciotti accuse Hollande d'instrumentaliser le terrorisme
09h31 L'impôt sur le revenu des CRS va fortement augmenter
09h15 Audiences TV : France 2 crée la surprise et arrive en tête avec "Alex Hugo"
09h04 Corse : quatre hommes interpellés après un règlement de comptes
08h57 Fashion Week : une collection en hommage aux éboueurs de New York

(c) Radio (Europe1)

TERRORISME
Daech est en passe de créer les conditions d'une guerre apocalyptique avec l'Occident

TRANSFERT - EPISODE 7
Mathieu avait rêvé Ian, il n'était pas celui qu'il croyait

KEYNOTE APPLE : LES ANNONCES POSSIBLES
%
de
babilité
iPhone 7
Apple Watch 2
Versement de 13 mds € à l'Irlande
Slate.fr

DATA YOLO
Rions un peu avec la keynote d'Apple

EN BREF

MÉGAFAUNE
Il y a 30.000 ans, les kangourous faisaient plus de deux mètres de haut

INTERNET
Facebook peut remercier ses haters pour le succès de son fil d'actualités

RECONVERSION
Dans le Kentucky, une start-up apprend le code aux ouvriers licenciés du charbon

SCIENCES
Les psychopathes ressentent la peur mais n'ont pas conscience du danger

GUERRE FROIDE
Sillamäe, l'étrange histoire de la ville qui «n'existait» pas

ÉTATS-UNIS
Hillary Clinton virée pendant le

(d) Pure online media (Slate)

Figure 2: Screenshots of the websites of different media, depending on their offline format

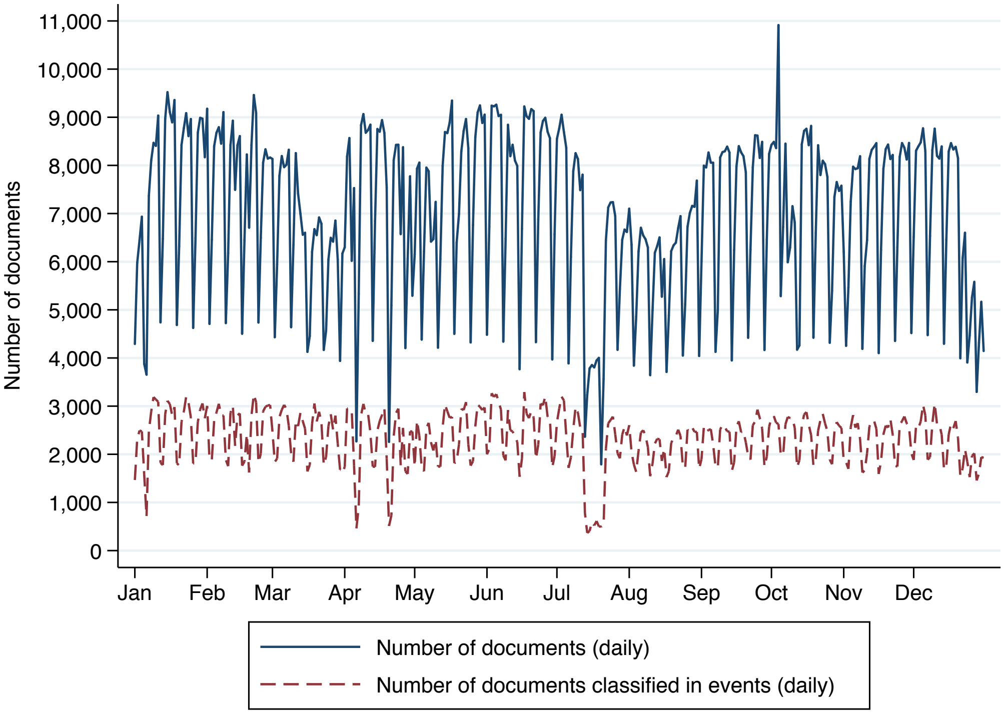


Figure 3: Daily distribution of the number of documents and of the number of documents classified in events in the dataset

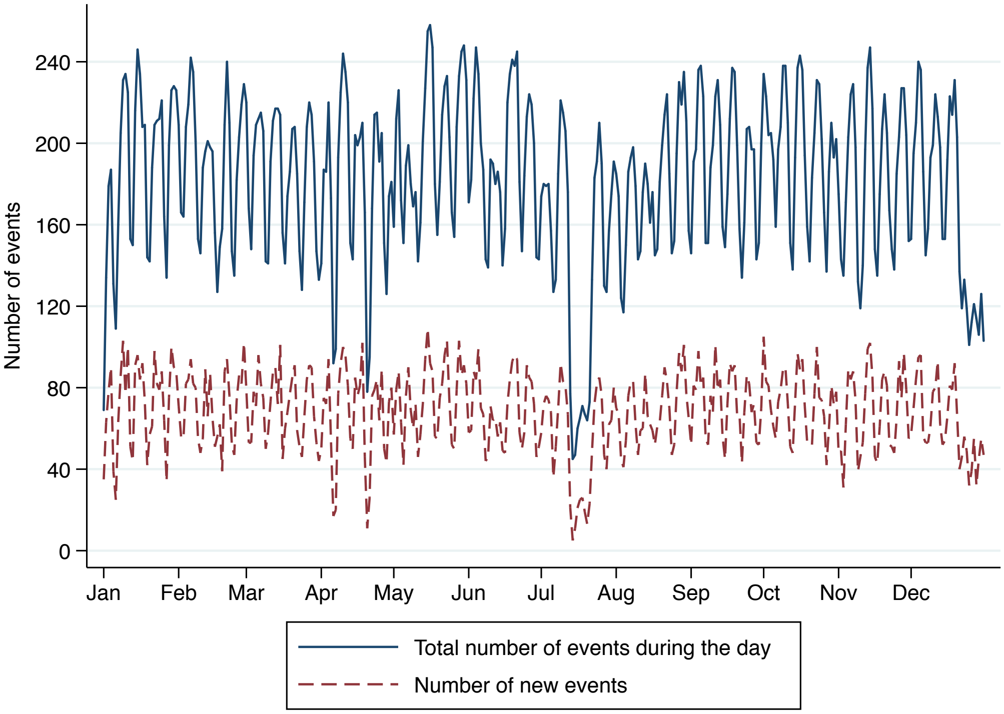
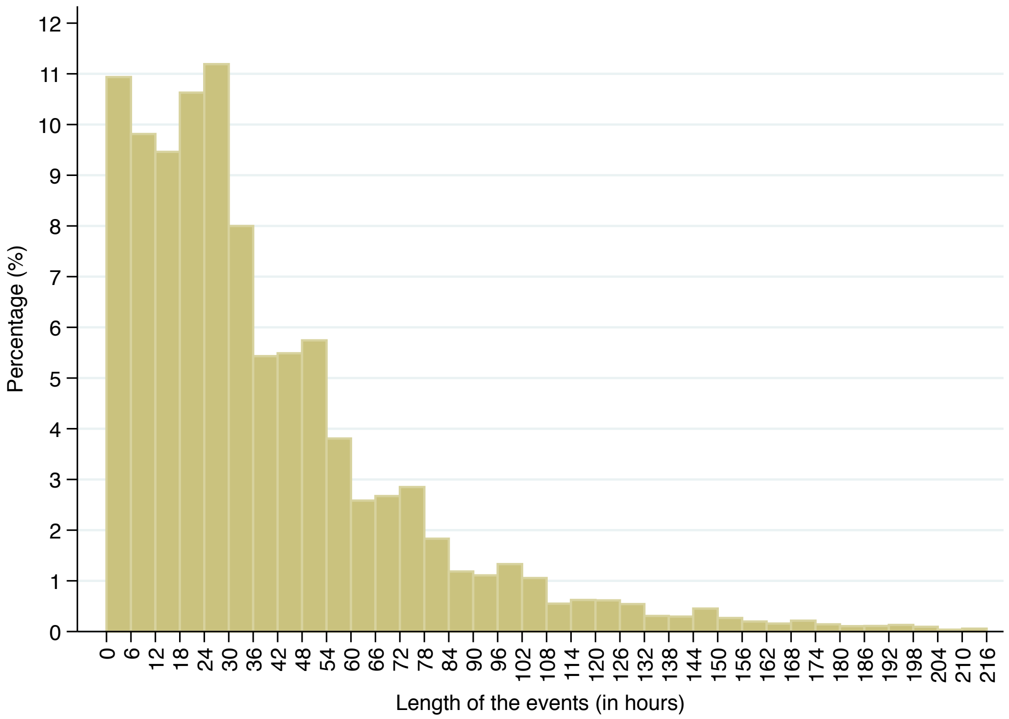
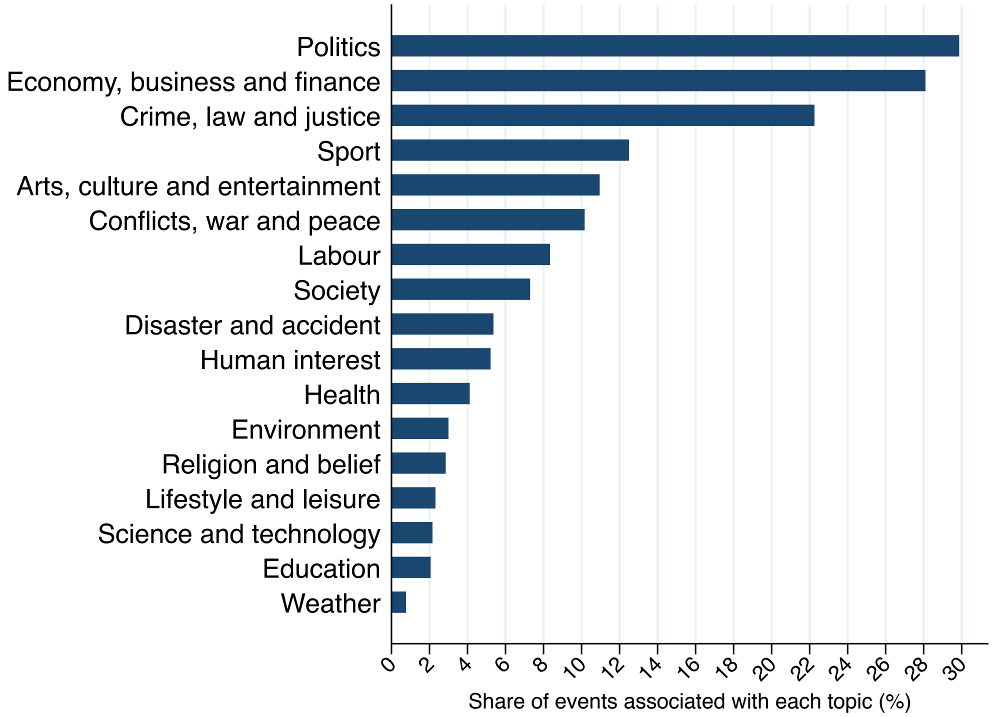


Figure 4: Daily distribution of the number of events



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

Figure 5: Distribution of the length of the events (in hours)



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

Figure 6: Share of events associated with each media topic

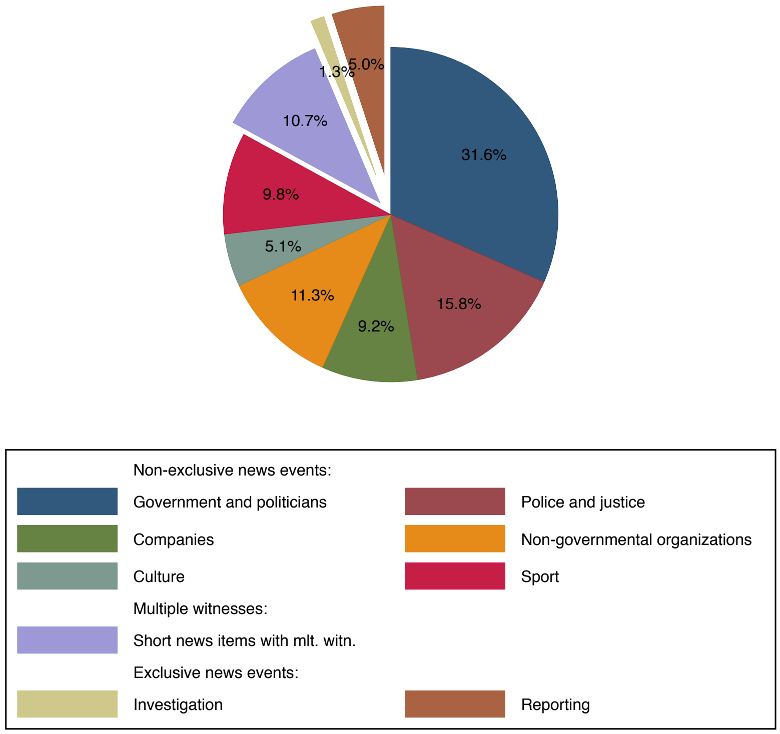
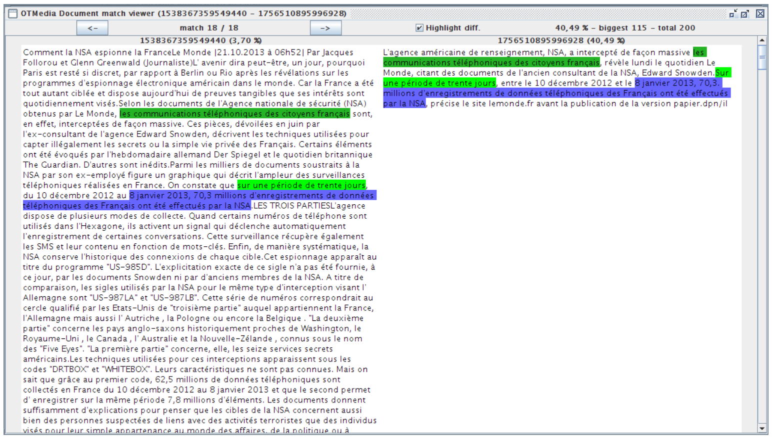


Figure 7: Share of the events depending on the information issuer


OTMedia Document match viewer (1538367359549440 – 1756510895996928)
<-
match 18 / 18
->
☑ Highlight diff.
40,49 % – biggest 115 – total 200

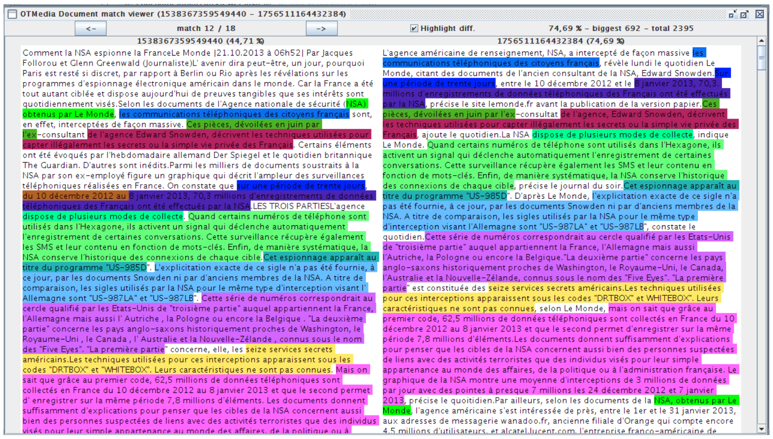
1538367359549440 (3,70 %)

Comment la NSA espionne la FranceLe Monde |21.10.2013 à 06h52| Par Jacques Follorou et Glenn Greenwald (Journaliste)L' avenir dira peut-être, un jour, pourquoi Paris est resté si discret, par rapport à Berlin ou Rio après les révélations sur les programmes d'espionnage électronique américain dans le monde. Car la France a été tout autant ciblée et dispose aujourd'hui de preuves tangibles que ses intérêts sont quotidiennement visés.Selon les documents de l'Agence nationale de sécurité (NSA) obtenus par Le Monde, les communications téléphoniques des citoyens français sont, en effet, interceptées de façon massive. Ces pièces, dévoilées en juin par l'ex-consultant de l'agence Edward Snowden, décrivent les techniques utilisées pour capter illégalement les secrets ou la simple vie privée des Français. Certains éléments ont été évoqués par l'hebdomadaire allemand Der Spiegel et le quotidien britannique The Guardian. D'autres sont inédits.Parmi les milliers de documents soustraits à la NSA par son ex-employé figure un graphique qui décrit l'ampleur des surveillances téléphoniques réalisées en France. On constate que sur une période de trente jours, du 10 décembre 2012 au 8 janvier 2013, 70,3 millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA.LES TROIS PARTIESL'agence dispose de plusieurs modes de collecte. Quand certains numéros de téléphone sont utilisés dans l'Hexagone, ils activent un signal qui déclenche automatiquement l'enregistrement de certaines conversations. Cette surveillance récupère également les SMS et leur contenu en fonction de mots-clés. Enfin, de manière systématique, la NSA conserve l'historique des connexions de chaque cible.Cet espionnage apparaît au titre du programme "US-985D". L'explicitation exacte de ce sigle n'a pas été fournie, à ce jour, par les documents Snowden ni par d'anciens membres de la NSA. A titre de comparaison, les sigles utilisés par la NSA pour le même type d'interception visant l' Allemagne sont "US-987LA" et "US-987LB". Cette série de numéros correspondrait au cercle qualifié par les Etats-Unis de "troisième partie" auquel appartiennent la France, l'Allemagne mais aussi l' Autriche , la Pologne ou encore la Belgique . "La deuxième partie" concerne les pays anglo-saxons historiquement proches de Washington, le Royaume-Uni , le Canada , l' Australie et la Nouvelle-Zélande , connus sous le nom des "Five Eyes". "La première partie" concerne, elle, les seize services secrets américains.Les techniques utilisées pour ces interceptions apparaissent sous les codes "DRTBOX" et "WHITEBOX". Leurs caractéristiques ne sont pas connues. Mais on sait que grâce au premier code, 62,5 millions de données téléphoniques sont collectés en France du 10 décembre 2012 au 8 janvier 2013 et que le second permet d' enregistrer sur la même période 7,8 millions d'éléments. Les documents donnent suffisamment d'explications pour penser que les cibles de la NSA concernent aussi bien des personnes suspectées de liens avec des activités terroristes que des individus visés pour leur simple appartenance au monde des affaires, de la politique ou à

1756510895996928 (40,49 %)

L'agence américaine de renseignement, NSA, a intercepté de façon massive les communications téléphoniques des citoyens français, révèle lundi le quotidien Le Monde, citant des documents de l'ancien consultant de la NSA, Edward Snowden.Sur une période de trente jours, entre le 10 décembre 2012 et le 8 janvier 2013, 70,3. millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA, précise le site lemonde.fr avant la publication de la version papier.dpn/il

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


OTMedia Document match viewer (1538367359549440 – 1756511164432384)
<-
match 12 / 18
->
☑ Highlight diff.
74,69 % – biggest 692 – total 2395

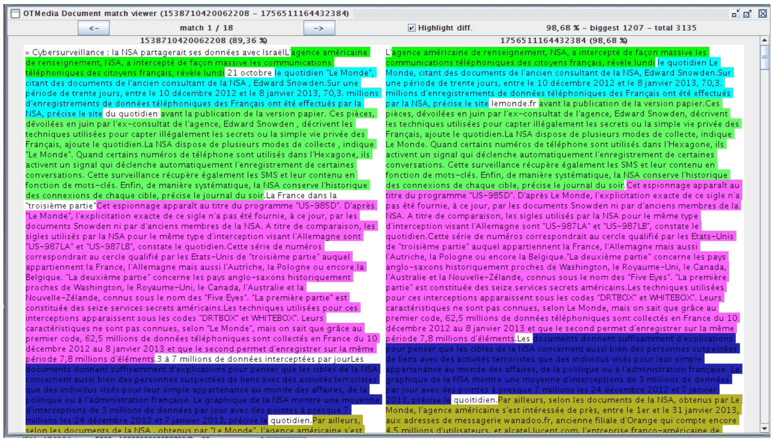
1538367359549440 (44,71 %)

Comment la NSA espionne la FranceLe Monde |21.10.2013 à 06h52| Par Jacques Follorou et Glenn Greenwald (Journaliste)L' avenir dira peut-être, un jour, pourquoi Paris est resté si discret, par rapport à Berlin ou Rio après les révélations sur les programmes d'espionnage électronique américain dans le monde. Car la France a été tout autant ciblée et dispose aujourd'hui de preuves tangibles que ses intérêts sont quotidiennement visés.Selon les documents de l'Agence nationale de sécurité (NSA) obtenus par Le Monde, les communications téléphoniques des citoyens français sont, en effet, interceptées de façon massive. Ces pièces, dévoilées en juin par l'ex-consultant de l'agence Edward Snowden, décrivent les techniques utilisées pour capter illégalement les secrets ou la simple vie privée des Français. Certains éléments ont été évoqués par l'hebdomadaire allemand Der Spiegel et le quotidien britannique The Guardian. D'autres sont inédits.Parmi les milliers de documents soustraits à la NSA par son ex-employé figure un graphique qui décrit l'ampleur des surveillances téléphoniques réalisées en France. On constate que sur une période de trente jours, du 10 décembre 2012 au 8 janvier 2013, 70,3 millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA.LES TROIS PARTIESL'agence dispose de plusieurs modes de collecte. Quand certains numéros de téléphone sont utilisés dans l'Hexagone, ils activent un signal qui déclenche automatiquement l'enregistrement de certaines conversations. Cette surveillance récupère également les SMS et leur contenu en fonction de mots-clés. Enfin, de manière systématique, la NSA conserve l'historique des connexions de chaque cible.Cet espionnage apparaît au titre du programme "US-985D". L'explicitation exacte de ce sigle n'a pas été fournie, à ce jour, par les documents Snowden ni par d'anciens membres de la NSA. A titre de comparaison, les sigles utilisés par la NSA pour le même type d'interception visant l' Allemagne sont "US-987LA" et "US-987LB". Cette série de numéros correspondrait au cercle qualifié par les Etats-Unis de "troisième partie" auquel appartiennent la France, l'Allemagne mais aussi l' Autriche , la Pologne ou encore la Belgique . "La deuxième partie" concerne les pays anglo-saxons historiquement proches de Washington, le Royaume-Uni , le Canada , l' Australie et la Nouvelle-Zélande , connus sous le nom des "Five Eyes". "La première partie" concerne, elle, les seize services secrets américains.Les techniques utilisées pour ces interceptions apparaissent sous les codes "DRTBOX" et "WHITEBOX". Leurs caractéristiques ne sont pas connues. Mais on sait que grâce au premier code, 62,5 millions de données téléphoniques sont collectés en France du 10 décembre 2012 au 8 janvier 2013 et que le second permet d' enregistrer sur la même période 7,8 millions d'éléments. Les documents donnent suffisamment d'explications pour penser que les cibles de la NSA concernent aussi bien des personnes suspectées de liens avec des activités terroristes que des individus visés pour leur simple appartenance au monde des affaires, de la politique ou à

1756511164432384 (74,69 %)

L'agence américaine de renseignement, NSA, a intercepté de façon massive les communications téléphoniques des citoyens français, révèle lundi le quotidien Le Monde, citant des documents de l'ancien consultant de la NSA, Edward Snowden.Sur une période de trente jours, entre le 10 décembre 2012 et le 8 janvier 2013, 70,3. millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA, précise le site lemonde.fr avant la publication de la version papier.Ces pièces, dévoilées en juin par l'ex-consultat de l'agence, Edward Snowden, décrivent les techniques utilisées pour capter illégalement les secrets ou la simple vie privée des Français, ajoute le quotidien.La NSA dispose de plusieurs modes de collecte, indique Le Monde. Quand certains numéros de téléphone sont utilisés dans l'Hexagone, ils activent un signal qui déclenche automatiquement l'enregistrement de certaines conversations. Cette surveillance récupère également les SMS et leur contenu en fonction de mots-clés. Enfin, de manière systématique, la NSA conserve l'historique des connexions de chaque cible, précise le journal du soir.Cet espionnage apparaît au titre du programme "US-985D". D'après Le Monde, l'explicitation exacte de ce sigle n'a pas été fournie, à ce jour, par les documents Snowden ni par d'anciens membres de la NSA. A titre de comparaison, les sigles utilisés par la NSA pour le même type d'interception visant l'Allemagne sont "US-987LA" et "US-987LB", constate le quotidien.Cette série de numéros correspondrait au cercle qualifié par les Etats-Unis de "troisième partie" auquel appartiennent la France, l'Allemagne mais aussi l'Autriche, la Pologne ou encore la Belgique."La deuxième partie" concerne les pays anglo-saxons historiquement proches de Washington, le Royaume-Uni, le Canada, l'Australie et la Nouvelle-Zélande, connus sous le nom des "Five Eyes". "La première partie" est constituée des seize services secrets américains.Les techniques utilisées pour ces interceptions apparaissent sous les codes "DRTBOX" et WHITEBOX". Leurs caractéristiques ne sont pas connues, selon Le Monde, mais on sait que grâce au premier code, 62,5 millions de données téléphoniques sont collectés en France du 10 décembre 2012 au 8 janvier 2013 et que le second permet d'enregistrer sur la même période 7,8 millions d'éléments.Les documents donnent suffisamment d'explications pour penser que les cibles de la NSA concernent aussi bien des personnes suspectées de liens avec des activités terroristes que des individus visés pour leur simple appartenance au monde des affaires, de la politique ou à l'administration française. Le graphique de la NSA montre une moyenne d'interceptions de 3 millions de données par jour avec des pointes à presque 7 millions les 24 décembre 2012 et 7 janvier 2013, précise le quoitidien.Par ailleurs, selon les documents de la NSA, obtenus par Le Monde, l'agence américaine s'est intéressée de près, entre le 1er et le 31 janvier 2013, aux adresses de messagerie wanadoo.fr, ancienne filiale d'Orange qui compte encore 4,5 millions d'utilisateurs, et alcatel.lucent.com, l'entreprise franco-américaine de

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


OTMedia Document match viewer (1538710420062208 – 1756511164432384)
<-
match 1 / 18
->
☑ Highlight diff.
98,68 % – biggest 1207 – total 3135

1538710420062208 (89,36 %)

» Cybersurveillance : la NSA partagerait ses données avec IsraëlL'agence américaine de renseignement, NSA, a intercepté de façon massive les communications téléphoniques des citoyens français, révèle lundi 21 octobre le quotidien "Le Monde", citant des documents de l'ancien consultant de la NSA , Edward Snowden.Sur une période de trente jours, entre le 10 décembre 2012 et le 8 janvier 2013, 70,3. millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA, précise le site du quotidien avant la publication de la version papier. Ces pièces, dévoilées en juin par l'ex-consultat de l'agence, Edward Snowden , décrivent les techniques utilisées pour capter illégalement les secrets ou la simple vie privée des Français, ajoute le quotidien.La NSA dispose de plusieurs modes de collecte , indique "Le Monde". Quand certains numéros de téléphone sont utilisés dans l'Hexagone, ils activent un signal qui déclenche automatiquement l'enregistrement de certaines conversations. Cette surveillance récupère également les SMS et leur contenu en fonction de mots-clés. Enfin, de manière systématique, la NSA conserve l'historique des connexions de chaque cible, précise le journal du soir.La France dans la "troisième partie"Cet espionnage apparaît au titre du programme "US-985D". D'après "Le Monde", l'explicitation exacte de ce sigle n'a pas été fournie, à ce jour, par les documents Snowden ni par d'anciens membres de la NSA. A titre de comparaison, les sigles utilisés par la NSA pour le même type d'interception visant l'Allemagne sont "US-987LA" et "US-987LB", constate le quotidien.Cette série de numéros correspondrait au cercle qualifié par les Etats-Unis de "troisième partie" auquel appartiennent la France, l'Allemagne mais aussi l'Autriche, la Pologne ou encore la Belgique. "La deuxième partie" concerne les pays anglo-saxons historiquement proches de Washington, le Royaume-Uni, le Canada, l'Australie et la Nouvelle-Zélande, connus sous le nom des "Five Eyes". "La première partie" est constituée des seize services secrets américains.Les techniques utilisées pour ces interceptions apparaissent sous les codes "DRTBOX" et WHITEBOX". Leurs caractéristiques ne sont pas connues, selon "Le Monde", mais on sait que grâce au premier code, 62,5 millions de données téléphoniques sont collectés en France du 10 décembre 2012 au 8 janvier 2013 et que le second permet d'enregistrer sur la même période 7,8 millions d'éléments.3 à 7 millions de données interceptées par jourLes documents donnent suffisamment d'explications pour penser que les cibles de la NSA concernent aussi bien des personnes suspectées de liens avec des activités terroristes que des individus visés pour leur simple appartenance au monde des affaires, de la politique ou à l'administration française. Le graphique de la NSA montre une moyenne d'interceptions de 3 millions de données par jour avec des pointes à presque 7 millions les 24 décembre 2012 et 7 janvier 2013, précise le quotidien.Par ailleurs, selon les documents de la NSA , obtenus par "Le Monde", l'agence américaine s'est

1756511164432384 (98,68 %)

L'agence américaine de renseignement, NSA, a intercepté de façon massive les communications téléphoniques des citoyens français, révèle lundi le quotidien Le Monde, citant des documents de l'ancien consultant de la NSA, Edward Snowden.Sur une période de trente jours, entre le 10 décembre 2012 et le 8 janvier 2013, 70,3. millions d'enregistrements de données téléphoniques des Français ont été effectués par la NSA, précise le site lemonde.fr avant la publication de la version papier.Ces pièces, dévoilées en juin par l'ex-consultat de l'agence, Edward Snowden, décrivent les techniques utilisées pour capter illégalement les secrets ou la simple vie privée des Français, ajoute le quotidien.La NSA dispose de plusieurs modes de collecte, indique Le Monde. Quand certains numéros de téléphone sont utilisés dans l'Hexagone, ils activent un signal qui déclenche automatiquement l'enregistrement de certaines conversations. Cette surveillance récupère également les SMS et leur contenu en fonction de mots-clés. Enfin, de manière systématique, la NSA conserve l'historique des connexions de chaque cible, précise le journal du soir.Cet espionnage apparaît au titre du programme "US-985D". D'après Le Monde, l'explicitation exacte de ce sigle n'a pas été fournie, à ce jour, par les documents Snowden ni par d'anciens membres de la NSA. A titre de comparaison, les sigles utilisés par la NSA pour le même type d'interception visant l'Allemagne sont "US-987LA" et "US-987LB", constate le quotidien.Cette série de numéros correspondrait au cercle qualifié par les Etats-Unis de "troisième partie" auquel appartiennent la France, l'Allemagne mais aussi l'Autriche, la Pologne ou encore la Belgique."La deuxième partie" concerne les pays anglo-saxons historiquement proches de Washington, le Royaume-Uni, le Canada, l'Australie et la Nouvelle-Zélande, connus sous le nom des "Five Eyes". "La première partie" est constituée des seize services secrets américains.Les techniques utilisées pour ces interceptions apparaissent sous les codes "DRTBOX" et WHITEBOX". Leurs caractéristiques ne sont pas connues, selon Le Monde, mais on sait que grâce au premier code, 62,5 millions de données téléphoniques sont collectés en France du 10 décembre 2012 au 8 janvier 2013 et que le second permet d'enregistrer sur la même période 7,8 millions d'éléments.Les documents donnent suffisamment d'explications pour penser que les cibles de la NSA concernent aussi bien des personnes suspectées de liens avec des activités terroristes que des individus visés pour leur simple appartenance au monde des affaires, de la politique ou à l'administration française. Le graphique de la NSA montre une moyenne d'interceptions de 3 millions de données par jour avec des pointes à presque 7 millions les 24 décembre 2012 et 7 janvier 2013, précise le quoitidien.Par ailleurs, selon les documents de la NSA, obtenus par Le Monde, l'agence américaine s'est intéressée de près, entre le 1er et le 31 janvier 2013, aux adresses de messagerie wanadoo.fr, ancienne filiale d'Orange qui compte encore 4,5 millions d'utilisateurs, et alcatel.lucent.com, l'entreprise franco-américaine de

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

Figure 8: Illustration of the plagiarism detection algorithm: the NSA spying scandal on October 21, 2013

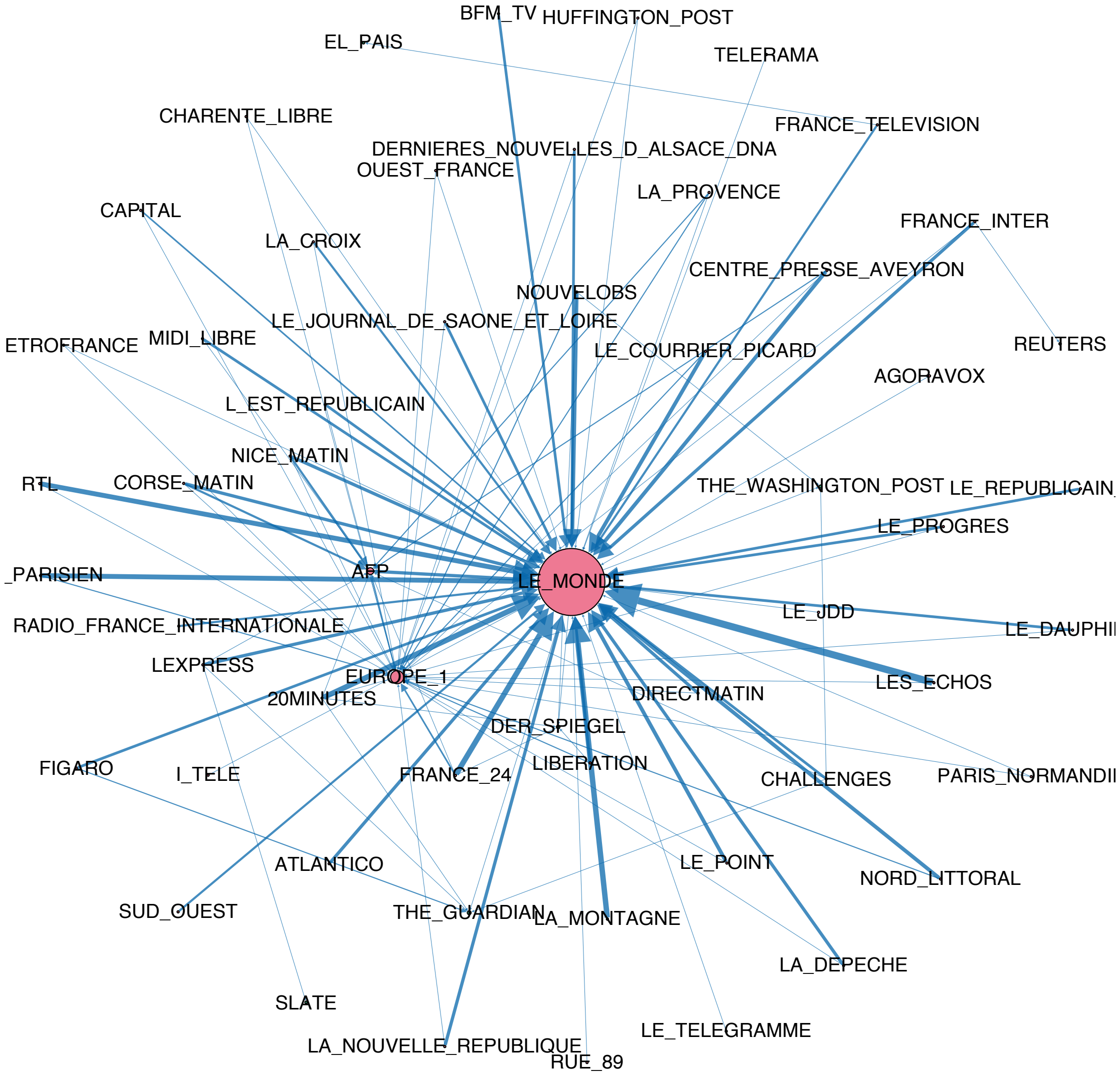
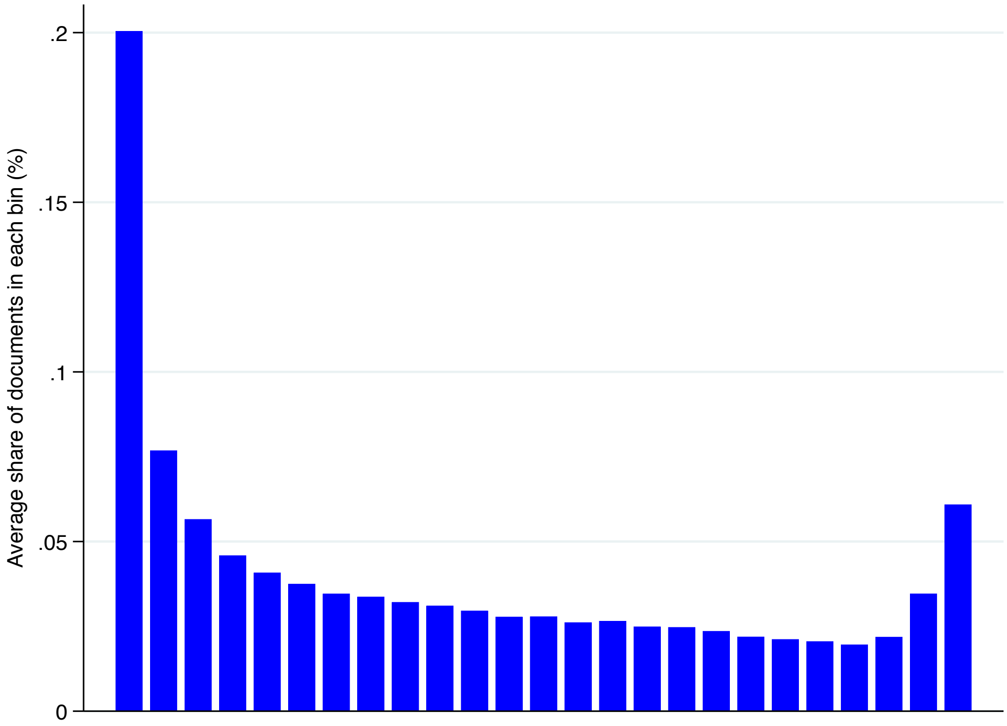
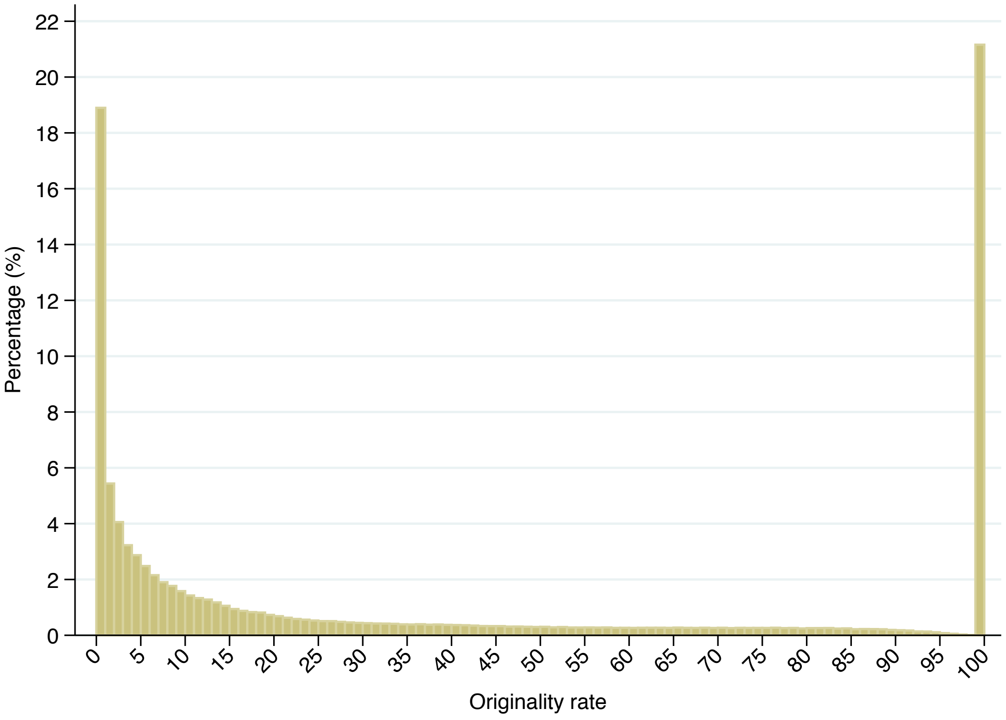


Figure 9: Illustration of the citation detection algorithm: the NSA spying scandal on October 21, 2013

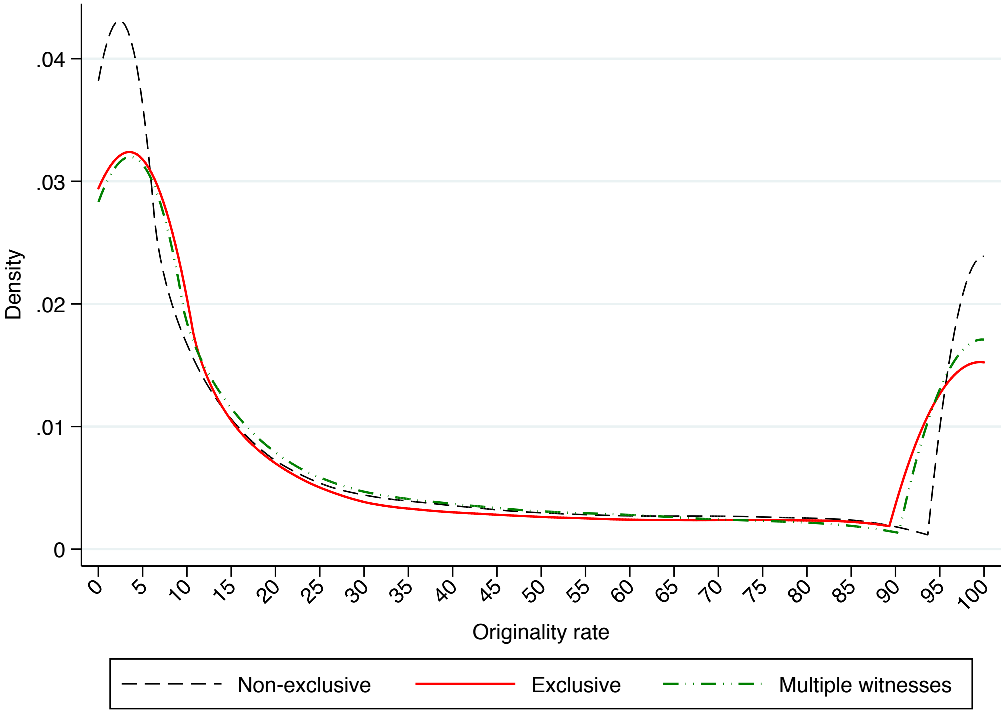


Notes: The Figure plots the distribution of the documents within the events. Each event is split into 25 bins of equal length.

Figure 10: Distribution of the documents within the events



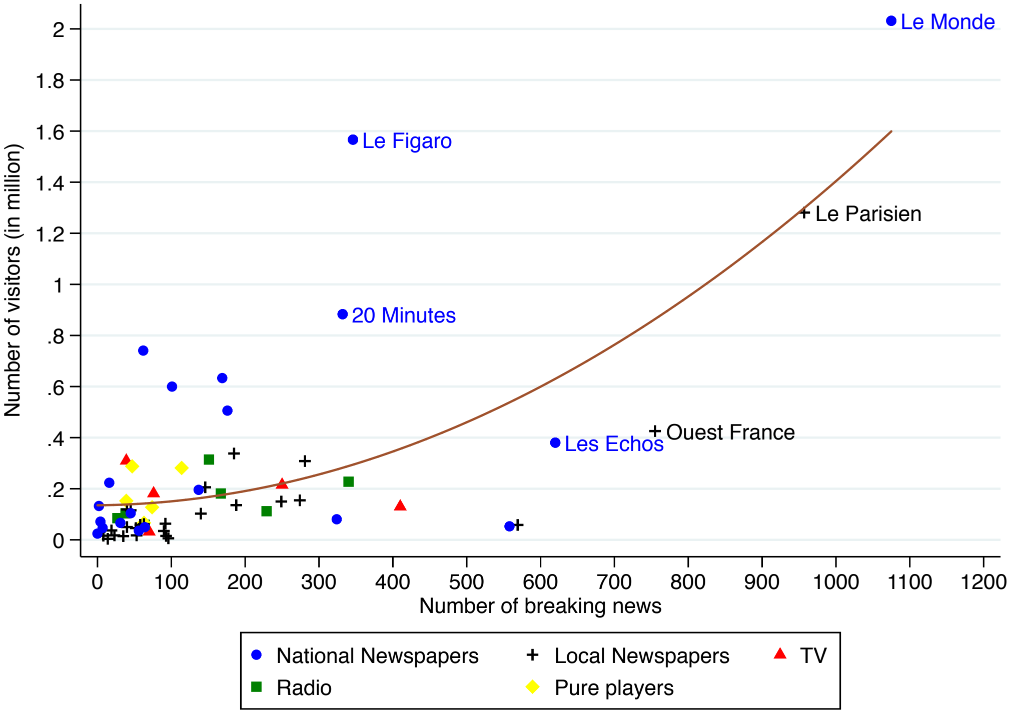
(a) Distribution (all documents included in events)



(b) Kernel density estimate

Notes: The upper Figure plots the distribution of the originality rate (with bins equal to one percent). The bottom Figure plots the Kernel density estimates depending on the nature of the news event.

Figure 11: Originality rate



Notes: The Figure shows the correlation between the total number of news broken by each media outlet in 2013 and its average daily number of unique visitors.

Figure 12: Number of breaking news and online audience