Who Broke the News? An Analysis on First Reports of News Events

Matthias Gallé
Xerox Research Centre
Europe

Jean-Michel Renders
Xerox Research Centre
Europe

Eric Karstens
European Journalism Centre

ABSTRACT

We present a data-driven study on which sources were the first to report on news events. For this, we implemented a news-aggregator that included a large number of established news sources and covered one year of data. We present a novel framework that is able to retrieve a large number of events and not only the most salient ones, while at the same time making sure that they are not exclusively of local impact.

Our analysis then focuses on different aspects of the news cycle. In particular we analyze which are the sources to break most of the news. By looking when certain events become *bursty*, we are able to perform a finer analysis on those events and the associated sources that dominate the global news-attention. Finally we study the time it takes news outlet to report on these events and how this reflects different strategies of which news to report.

A general finding of our study is that big news agencies remain an important threshold to cross to bring global attention to particular news, but it also shows the importance of focused (by region or topic) outlets.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data Mining*

Keywords

breaking news; topic detection and tracking; data journalism

1. INTRODUCTION

Being the first to report on an event and to *break news* is the daily holy grail of many journalists and news agencies. A long-standing promise of social media is that the reporting of these news would become more and more the realms of social media. However, recent studies do not seem to support this: blogs for instance have been proven very useful for an a-posteriori analysis of the news [12], but only a small fraction of events originate from them [11].

However, who exactly was the first of the traditional media to report on a given event has not been studied quantitatively to our knowledge. This papers aims to bridge this gap. In particular we try to answer two related questions regarding an eventual existence of a group of "gatekeepers" in the news sphere. The first often-heard opinion is that only this group creates and reports news, while the remaining outlets only look at and build upon what this group reports. Another related opinion is that, although many different outlets act as news generator, only those picked up by this group reach the attention of the general public.

To investigate to which extent these opinions are exact, we crawled and analyzed 820 000 articles coming from 60 sources, covering the whole planet over a period of one year (see the whole list of sources in Table 2). We did not restrict our attention exclusively to the most noteworthy events, but purposely looked for smaller events – provided that their repercussions were big enough to be considered international news. While we focused on English-speaking articles, the selected sources guarantee that besides America and Europe, also Africa, Asia and Oceania were covered.

It used to be that the news-sphere was dominated by the Big Three [13], the group of largest global news agencies (Thomson Reuters, Associated Press, Agence France Press). Hester for instance reports that in 1971 "half of all the daily papers [of the United States] use only one of the major wire services – the Associated Press". Of course many things have changed in the last 40 years. However, the reports of social news as taking over the role of breaking news may have been greatly exaggerated, for now. With respect to the blogosphere, a famous study tracking *memes* [11] reported that only 3.5% of these phrases originated outside traditional media. In fact, most of the research on the influence of blogs in news reporting has turned around using the repercussions in blogs to rank the news presented to a user (see in particular the Top stories identification task of the TREC-2009 Blog track [12]). Microblogging platforms, most notably Twitter, have more potential to take on this role thanks to their informal, real-time and quick creation characteristics. However, separating the wheat from the chaff is a difficult problem due to the sheer volume of these tweets. Petrović et. al [14] considered first story detection on Twitter, and presented some of the typical challenges faced when processing tweet streams. An interesting finding there is that the type of events detected is biased towards events involving celebrities. Whether other general events (of higher geo-political or economic impact) also get detected first on Twitter is not so clear, even if anectodal evidence from the Arab Spring for example may let think so. In any case, a very recent (late 2012) survey of the New York Times reports that, no matter from which source people heard about a story, 60% of the people turned to an established outlet to confirm it [4]. Therefore, despite the growing importance of social media

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

(and without denying it), the question of which traditional news media was the first to report on a given event is still of interest.

Most studies on event detection have considered a focused geography, few sources, a short time period or a very narrow definition of event [11, 17, 15]. Note in particular that the number of articles we worked with is an order of magnitude bigger than the standard TDT (*Topic Detection and Tracking*) collections [8]. The challenges due to the variety and volume of the articles make it hard not to fall into one of two extremes: either to report only the most salient events, or get drowned by the many local events which are not of interest to wider regions. Addressing these challenges needs alternative solutions to traditional news event detection algorithms.

In this paper we present a methodology to recover a general concept of event from a rich variety of sources while not being biased by the number of local events which are not interesting at a global level (Sect. 2). Once having defined and retrieved these events, we can then analyze them to see which source was the first to report on each one of them. Moreover, by using algorithms for automatic detection of burstiness, we can analyze when an event becomes "hot news", getting most of the attention of the news-sphere, and which were the sources that initiated this burst (Sect. 3). Note the difference between "the first to report" and "the first to initiate the burst", as the second kind of sources is supposed to constitute some gateway or obligatory marker for an event to gain significantly in popularity over a general audience.

2. EVENT CREATION

Finding events from a continuous stream of incoming articles originated by different sources is the problem of study of the Topic Detection and Tracking (TDT) community [2]. We provide a short overview of our algorithms, emphasizing how they diverge from more standard approaches.

Our definition of event is the one given by the TDT community [8]: "an event is a particular thing that happens at a specific time and place". This emphasizes the precise nature of an event, as opposed to bigger (in time or number of actors) *stories*. The events we are interested in have generally a life-span of less than a week (in the news-sphere). Note that an emphasis of this study is to not focus only on the biggest news events, but to inspect also those of lesser importance. At the same time this has to be balanced with the fact that we are interested in *global* news, whose interest exceeds a particular country or region. To address the balance between coverage and global significance, we implemented a two-stage framework. In the first phase we selected a list of *primary sources* (the sources that are well known to offer factual, concise description of events) and extract from the articles of these sources the key – and hopefully unique – subject it treats. These are then clustered and if the resulting clusters pass a threshold criterion on diversity and quantity, these clusters become events. The remaining articles (including those of the close past) are then compared to these events and may become attached to one or more of them. Let us examine each step in more detail.

In order to retrieve news articles, we first selected the list of sources to consider (see Table 2). They were then crawled every hour to retrieve any new article that was published since the last inspection. Any given article may discuss more than one event. In order to permit such a behavior when assigning articles to events, we actually considered the article at a segment level, where a segment is a syntactic unit (a sentence typically) as given by a parser [1].

In the first phase, we selected a list of *primary* news sources, whose articles build the basic, core structure of potential events. Because one article may talk about more than one event, we furthermore refined this by only considering the segments containing the first 100 words (including the whole segment containing the 100th word), which is known to be very a good baseline for automatic summarization [9]. These *main segments* constituted the scaffolds of (potential) events and were given as input to the clustering algorithm.

To cluster the articles, we designed the Star-EM algorithm [6], a clustering algorithm that takes as input parameter a similarity threshold that should be respected by the main segments of articles inside a cluster. Using such a similarity threshold as parameter, rather than the number of clusters, gives the system more flexibility to adapt to the evolving nature of the news cycle. Additionally, this algorithm easily permits to handle the mini-batch setting associated to the periodic crawling of the articles. It performs better than other variants (including the popular incremental clustering algorithm [16] commonly used for this task), while able to scale up to the large amount of data we had to process. To enforce that events should be rather concentrated within a short space of time, we marked them as inactive (and archived them) once the average timestamp difference of articles belonging to an event exceeded 3 days. A further hard constraint is that a found cluster is only considered as event if it is reported by at least two different sources (diversity) and contains at least three articles (quantity).

Once the clustering is completed, the second phases initiates. In this phase, called *excerpt extraction* we consider all remaining segments (this is, segments of the primary sources not part of the main segments, and articles from non-primary sources). For each crawl and its associated active events, we looked at all articles from the last 48 hours, and assigned their segments to zero, one or more of the existing events. The main goal is to find consecutive segments which are semantically coherent and which could be related to a pre-identified event. In other words, when associating segments to events, the decision of association is not done independently for each segment, but taking into account the context of the neighboring segments, as it is likely that consecutive segments deal with the same event. We developed and compared three models to formalize this.

2.1 FullHMM

In our first approach, we model this task with a Hidden Markov Model (HMM), where each node corresponds to one event. For each segment, the distribution over the output tokens is proportional to the similarities between the segment and the event (the centroid of the cluster corresponding to this event). We fix the transition probabilities to stay in the same node ($\propto \beta$), or to switch ($\propto 1-\beta$). We add one node for the unknown event, denoted by u. This special node has the meaning of capturing all segments related to events not modelled by our current model of the news-sphere. We fix the similarity between the unknown event and any segment with the fixed parameter t. This parameter can be inter-

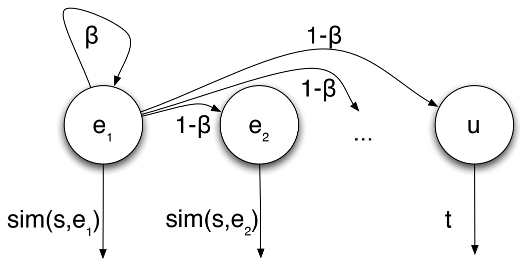

Figure 1: FullHMM

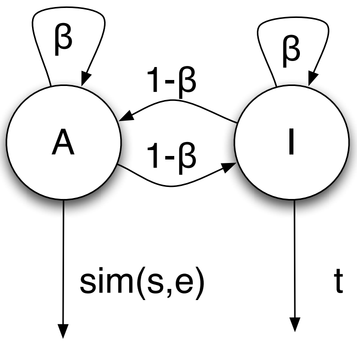

Figure 2: IndependentHMM

preted as a threshold to belong to one of the existing event (similar to the threshold used in the clustering module). The additional parameter β is the weight of staying in the same state, a stickiness parameter. The problem of finding the most probable path for the given sequence of segments is a classical dynamic programming problem. A graphical representation is given in Fig. 1. The states $e_1 \dots e_n$ correspond to the currently active events. Given one article, we then compute the most probable paths over the segments. Under the given assumptions this path corresponds then to an assignment of each segment to one event (or u).

2.2 IndependentHMM

In this approach, the model again is represented by an HMM, but this time we use one independent HMM for each event. Each HMM consists of only two states, Active and Inactive, meaning that the considered event has a relationship to the current segment or not. Due to this explicit independence assumption, a given segment may belong to more than one event (even if this happens rarely in practice). In this model, the most probable path for each event gives the segments of the article that deals with this event. A graphical representation on one HMM is given in Fig 2, with the same interpretation of the two parameters (t and β) as before. This is a similar model to the previous one, with the difference that one independent model is instantiated per event.

2.3 Segmentation

The last approach diverges from the previous ones and tries to find boundaries in the given sequence of segments.

	Segmentation		FullHMM		IndependHMM	
	Closed	Open	Closed	Open	Closed	Open
micro P	0.693	0.687	0.565	0.679	0.702	0.627
micro R	0.891	0.896	0.822	0.782	0.853	0.708
micro F_1	**0.779**	**0.778**	**0.670**	**0.727**	**0.770**	**0.665**
macro P	0.695	0.685	0.691	0.694	0.684	0.638
macro R	0.875	0.870	0.812	0.774	0.836	0.707
macro F_1	**0.741**	**0.732**	**0.718**	**0.695**	**0.726**	**0.629**

Table 1: Results of the Excerpt Extraction phase

The problem is cast as a segmentation problem, which again can be resolved by a dynamic programming algorithm. Resolving simultaneously a classification and a segmentation problem is a well-studied problem [7, 3]. Here we take a traditional approach inspired by [5]. The formal definition of the maximization problem we want to solve is as follows:

Given articles $d = s_1 \dots s_n$, find (meta-)segments $T_1, \dots, T_k$, such that $d = T_1 \dots T_k$, $T_i = s_j \dots s_{j+\ell}$ for some j and $\ell > 0$ and

$$\frac{1}{k}\left(\sum_{i=1}^{k} \max_{e \in E \cup \{u\}} sim(T_i, e)\right) - \beta \times k$$

is maximised.

Here, β regulates the number of changes (transitions) of events from one segment to another, and again we use a fixed threshold t for the similarity between any meta-segment and the unknown event u ($sim(T_i, u) = t$). Note that the interpretation of t and β is similar to those of the previous models.

2.4 Evaluation

Among the crawled articles of one particular month, 429 (185 from primary sources) articles were annotated with a total of 47 events and were used as ground truth to evaluate the clustering and excerpt extraction procedures. Each article could be annotated with one or more events. We added to this dataset 274 articles from another month to act as *noise* and imitate an open-world behavior where some articles may never get attached to an event. In all cases, we report micro-precision, micro-recall and micro-F_1, considering the clusters as predictors of the groundthruth events and the alignment between cluster labels and groundtruth labels that maximized the F_1 score.

An extensive evaluation of the clustering algorithm on the TDT collection is given in [6] and we only report here the results on our collection (the results are slightly higher than those on TDT): we obtained a F_1 of 0.8505.

With respect to the excerpt extraction, we report more exhaustive results in Table 1. As can be appreciated, the last model (segmentation by applying dynamic programming on a suitable cost functional) performs slightly better than the HMM models. Probably more important is the fact (which cannot be appreciated in Table 1) that the Segmentation approach is more robust with respect to the choice of thresholds. We therefore used this method in our experiments.

In practice, a bit over half of the crawled articles were assigned to an event. The remainder were either filtered as spam (this is, a list of pointers to other articles), duplicate or reporting a very local event not selected by the algorithm due to the hard constraints of diversity and quantity, as explained above. While this number seems very low, the variety of sources (see Table 2) explains that most of the articles will be of a local interest only.

3. RESULTS & ANALYSIS

Now that we have created our event structure, we can turn to the questions we posed in the Introduction. We analyzed 820 000 articles over one year (June 2011 – May 2012) coming from 59 sources. They are detailed in Table 2, highlighting in italics those that were considered as primary sources. In all cases we used APIs provided by these sites to access their feed of English articles. The region specified in Table 2 refers to the main geographical reporting target, which of course is not exclusive. Our news-aggregator retrieved a total of 10 752 events during this time.

3.1 First to arrive

Recall that our first question was: who is the first to report on any given event? Considering the events we captured to be a snapshot of all events happening globally, we answer this question in Table 3 where we report for each source the events it reported first on (as a percentage over the total number of events). Note the surprising high rank of `allafrica`: this makes sense considering that it is actually a site aggregating news from "over 130 African news organizations" (`allafrica.com`).

However, the top ranked sources happen to be also some of the most prolific sources of articles. Without any other prior knowledge, the likelihood of being the first one is of course higher for a source that publishes more than another. In Table 4 we show the *percentage* of breaking-news articles, over the total number of articles published by any given source that are attached to some event (remember that about half of the published articles are not related to any "event" due to our way of identifying and constructing events). When this normalization takes place, the ranking changes dramatically (see Table 4). The first places here belong to sources which cover a very specific area, and have a very low rate of (English) articles. Because we normalize by the number of articles attached to an event (and not the total one published), this can be interpreted as follows: if a source is reporting about an event happening in a particular region or on a specific topic *that will be of global interest*, then these sources are good candidates to be those that break the news. Note also that, when we filter out the sources that report less than a thousand of news articles, `AP` appears second, behind `BBC`. Compare this to the rather low rank in Table 3.

Source Name	Coverage region
ABC News	US
Al Jazeera	Arabic World
All Africa	Africa
ANSA	Italy
Antara News	Indonesia
AOL news	Global
AP	Global
BBC	UK
Boston Globe	US
Budapeast Business Journal	Hungary
Businessweek	Global
CBS News	US
China News Service	China
Chosun	South Korea
CNN	US
Cyprus Mail	Cyprus
Daily Mail	UK
Daily Mirror	UK
Der Spiegel	Germany
DW-World	Germany
EHealthNews	Europe
EUbusiness	Europe
EUobserver	Europe
EurActiv	Europe
Euronews	Europe
EuropeanAgenda	Europe
EuroTopics	Europe
Fox News	US
France24	France
FT	Global
Helsinki Times	Finland
Kyodo News	Japan
The Wall Street Journal	US
Irish Examiner	Ireland
Le Monde diplomatique	France
Mercopress	Latin America
Moscov News	Russia
MSNBC	Global
New Europe	Europe
New Scientist	Global
North Africa Journal	North Africa
Novaya Gazeta	Russia
Novinite	Bulgaria
NPR	US
NY Post	US
NY Times	US
Reuters	Global
RFERL	Asia, M East
RIAN	Russia
The Australian	Australia
The Globe and Mail	Canada
The Guardian	UK
The Herald (Glasgow)	Scotland
The Star	Malaysia
The Sun	UK
The Telegraph	UK
Times of India	India
Times of Malta	Malta
Voice of America	US

Table 2: List of crawled sources, highlighting in italics primary ones.

3.2 Bursty periods

Our second question was: in order for an event to attract the global attention, is it necessary to get reported by one of the few "gatekeeper" sources? To answer this question, one has to better understand the dynamics of the news cycle. For any given event, the distribution of articles over time is of course not uniform. In Fig. 3 we show a snapshot of the evolution of 100 events. Each dotted, horizontal line represents one event, and a dot at (x, y) indicates the arrival at time x of an article discussing event y. As can been seen, the dynamics of different events can vary a lot, but in general there is a trend towards a (i) slow build-up of an event, followed by a (ii) dense zone and (iii) finished by a slow decay, which sporadic reports some time after the dense part of the event ended. We are interested in (ii), the dense

source	percentage
Reuters	12.96%
All Africa	11.68%
France24	10.41%
The Globe and Mail	5.47%
BBC	4.91%
CNN	4.89%
Businessweek	3.01%
RIAN	2.67%
Daily Mirror	2.59%
CBS News	2.32%
Daily Mail	2.21%
The Telegraph	2.21%
NY Post	2.19%
Kyodo	2.07%
NY Times	2.06%
Fox News	1.78%
The Sun	1.71%
DW	1.67%
NPR	1.64%
Times of India	1.45%
AP	1.42%
Al Jazeera	1.36%
RFERL	1.34%
Chosun	1.31%
Novinite	1.25%

Table 3: Sources reporting first on an event (percentage of the events for which a given source was the first to report).

source	percentage	#articles
North Africa Journal	25.00%	28
EHealthNews	18.18%	22
The Herald (Glasgow)	17.39%	69
Helsinki Times	10.98%	255
Le Monde Diplomatique	10.87%	46
Voice of America	9.38%	32
New Scientist	6.90%	377
BBC	6.56%	7 754
Budapest Business Journal	6.54%	734
AP	6.26%	2 347
New Europe	6.26%	991
Chosun	6.03%	2 254
France24	5.53%	19 495
NY Times	5.40%	3 962
Times of India	5.39%	2 783
Moscow News	5.06%	692
Reuters	4.85%	27 678
Cyprus Mail	4.27%	562
Times of Malta	4.25%	2 422
Novinite	4.19%	3 103
The Sun	4.14%	4 271
NY Post	4.08%	5 557
The Star	4.06%	2 068
Daily Mirror	3.99%	6 713

Table 4: Percentage of breaking news articles, over the total number of published articles per source that are linked to events (last column)

zone where this event is being *hot news*. This dense area is not exactly in the middle of the lifespan of an event.

Note that the overall amount of articles published is roughly constant (actually periodic, with low volumes during the weekends). This means that the total bandwith of the news-sphere does not change radically: a dense area in Fig. 3 indicates therefore a particular emphasis of the attention of the global news-sphere and the general public audience on this particular event. Such a dense period does not necessarily appear for an event; in some way, this event failed to become "hot topic" by capturing sufficient "market shares" in the news-sphere.

In order to detect these dense zones, we need a way to identify an increase in the emission of articles on one event over a reduced timeperiod. Kleinberg's *bursty detection* algorithm [10] addresses exactly situations where the goal is to "detect features that occur with high density over a limited time period". We used the two-state approach described there which models the emission of articles with an HMM of two states: a normal state, and a bursty one. Bursts of arrivals are then detected tracking the most probable emission path in this model. This also permits to associate a weight to each burst which is proportional to the prominence and the duration of the burst, which corresponds to the likelihood gain associated in using the bursty state over using the normal one. Note that, for some events, the optimal path consists in remaining in the normal state; this corresponds to events that do not break the news.

When applied to each of the events, this algorithm detects a total of 6 880 bursts. This corresponds to roughly two-thirds of all events. For each burst period, we then look at the source that created the first article in the burst, which may in general be different from the source that reports the event for the absolute first time. Adding the weight of the score to the source of this burst-generator, we then have a prominence score for each source. With such ranking, the top 25 sources are given in Table 5, together with their total score (normalized by the highest value).

There are several different possible interpretations of this prominence score: on the one hand, it is natural that news agencies rank high because it is their purpose to actively push news to third parties for re-dissemination and to do so at global scale. On the other hand, news outlets may look at local sources to filter news coming from the regions covered by them. The high rank of some local sources (like those covering Russia, the Arab world and Canada) may then be explained as an approval stamp of the news-sphere regarding their authenticity and (non-)bias. Finally, it may just be that the non-agencies sources just have a good journalistic "nose", being able to distinguish those news that will become trendy and publishing on them at the right time.

3.3 Lag Time

We finally performed a finer grained analysis of the time it took each source to report on an event. Even if a news outlet was not the first to break a news, it is of interest to measure how much time it takes it before publishing an article about this event.

We consider an event to be interesting for a given source, if the latter eventually publishes an article on it. The question we are analyzing here is: how much time does it take for a source to report on an event that interests it? For each event *e* a source reports on (without being the one breaking the

source	total burst score
Reuters	100.0
The Globe and Mail	83.9
CNN	72.7
Al Jazeera	58.0
France24	53.1
RIAN	47.0
The Star	45.6
CBS News	43.8
MSNBC	42.4
NPR	38.6
The Sun	37.5
DW	34.7
The Guardian	32.1
BBC	30.9
Businessweek	26.8
All Africa	22.1
AP	21.6
Kyodo	21.0
Novinite	19.5
NY Ttimes	18.5
RFER	17.8
The Telegraph	17.7
Chosun	16.3
Fox News	16.3
Daily Mirror	14.7

Table 5: Total burstiness score of sources

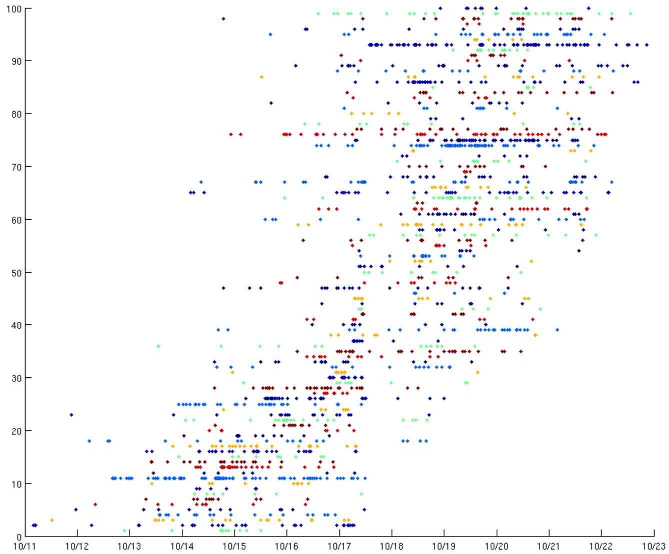

Figure 3: Progression of events over time. A dot at (x, y) indicates the arrival of an article concerning event y at time x

source	hours
France24	20.85
Reuters	20.87
BBC	21.41
Antara News	21.78
All Africa	22.69
Kyodo	23.47
Fox News	24.74
Al Jazeera	24.86
ANSA	24.95
CNN	24.96
RIAN	25.38
RFERL	26.29
The Telegraph	26.53
Daily Mirror	26.70
Euronews	27.40
The Globe and Mail	27.45
NPR	27.94
The Sun	27.95
Novinite	28.16
CBS News	28.52
DW	28.73
NY Post	28.79
AP	28.99
Daily Mail	29.45
MSNBC	29.67

Table 6: Median lag to report on an event per source (in hours)

news), we measured the time difference of the first report of this source and the timestamp of the first article reporting on it. In Table 6 we report the fastest 25 sources, measured by the median value of this lag.

It seems surprising how small the difference in the lag is in general. While there is a considerable difference between the first and the last in our list (50 hours), the difference between consecutive positions is almost negligible. This may indicate an existence of a basic time necessary for journalists to hear from a story, investigate it, write it and get the authorization to publish it.

The lag-time per event seems to follow a power law in general, underlying the obvious fact that most events are reported very fast, while others few take considerably more time. While the median (which we report) is more robust to these outliers than the average for instance, the numbers in Table 6 do not say anything about eventual strategies of a news outlet to give higher priority to certain events. We therefore analyzed more finely the fastest sources, to see which percentage of events they are interested in get reported quickly. For each source, we sorted the events it reported on with respect to their specific lag time, and analyzed increasing percentages of this sorted list. This is, we looked at the median lag time for the first x% events this source reported on the fastest. The curves in Fig. 4 reveal more subtle differences between the news sources: from the five sources displayed there, three (`france24`, `reuters`, `bbc`) seem to be clearly governed by two lines with different slope. The first line increases more slowly and controls the behavior for the fastest 50–60% events, which get detected and reported in the first 5–6 hours, while the remaining events are reported at a slower pace. Note that, for the other two sources, the growth seems to be more uniform.

It appears that, despite the “online first” policy that many news outlets have adopted, the classic news cycle of around

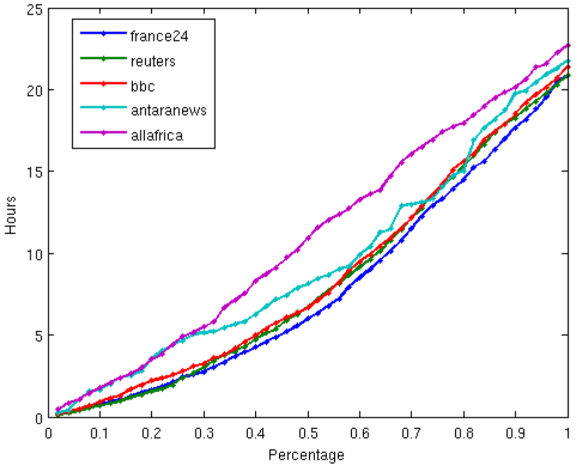

Figure 4: Median lag evolution for events

one day between two separate issues of a newspaper or main news broadcast remains intact. Of course, this delay is not any more a fixed deadline but it has been replaced by what might be called a "standard delay": a period waiting for confirmation and assessing news before picking it up from the original source. For many sources, about half of that time could just be explained by nighttime (with the exception of a few 24/7 operations such as global news agencies, CNN and Al Jazeera). This may be interpreted as an indicator of a lingering conservatism of news producers and news consumers – even news that are broken in social media or are exclusive by individual news organisations keep being vetted by trusted sources and only then reach the critical mass (burst) to become widely disseminated.

4. CONCLUSIONS

In this paper we made a extensive and quantitative analysis on which news sources were the first to break news. For this we developed a novel framework to cluster the news articles, particularly targeted towards analyzing most of the global events in opposition to just recover the biggest one, or all local events.

In a broad sense our data-driven study shows that big news agencies remain an important threshold to cross to bring global attention to any news (but apparently less so to be actually the first to report on them). Local news outlets, with a focused geographical target, also appear to be of primordial importance as they may act as a filter for the global community.

The analysis we used here can also be used as an indicator for the productivity and quality of news agencies and other outlets reporting first on any events. A potential customer of a news agency could look at these data to determine which to subscribe to (or whether just to make a deal with a local newspaper of his interest, for that matter). A news consumer could narrow down her news sources accordingly, going for the most salient yet least redundant selection of sources.

Finally, this study shows that there is a promising gap that may be filled with the rise of citizen journalism and social media. The considerable time it takes a news outlet to report on a given event shows that there is enough space for a crowd of citizen journalist to report and extend the news, as is already happening. Of course, to build trust in the accuracy of these alternative journalists will face other challenges.

Acknowledgments This work was partly funded by ICT-Content & Knowledge Program of the EC (SYNC3, FP7-231854)

5. REFERENCES

[1] S. Aït-Mokhtar, J.-P. Chanod, and C. Roux. Robustness beyond shallowness: incremental deep parsing. *Nat. Lang. Eng.*, 8:121–144, June 2002.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, pages 37–45, New York, New York, USA, 1998. ACM Press.

[3] T. Brants, F. Chen, and A. Farahat. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 330–337, New York, NY, USA, 2003. ACM.

[4] B. Brett. 2012 News eco-system study by the New York Times, October 2012. `http://www.poynter.org/latest-news/top-stories/190586/new-data-show-shifting-patterns-as-people-seek-news-across` last accessed April 2013.

[5] P. Fragkou, V. Petridis, and A. Kehagias. A dynamic programming algorithm for linear text segmentation. *J. Intell. Inf. Syst.*, 23(2):179–197, Sept. 2004.

[6] M. Gallé and J.-M. Renders. Full and semi-batch clustering of news articles with Star-EM. In *ECIR*. Springer, 2012.

[7] M. A. Hearst. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64, Mar. 1997.

[8] http://www.ldc.upenn.edu/ProjectsTDT2004. TDT: Annotation manual - version 1.2.

[9] I. Kastner and C. Monz. Automatic single-document key fact extraction from newswire articles. In *EACL*, pages 415–423, Athens, Greece, March 2009. Association for Computational Linguistics.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 91, New York, New York, USA, 2002. ACM Press.

[11] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–506. ACM, 2009.

[12] C. Macdonald, I. Ounis, and I. Soboroff. Overview of trec-2009 blog track. In *In Proceedings of TREC 2009*, 2009.

[13] T. L. McPhail. *Global Communication: Theories, Stakeholders, and Trends*. John Wiley and Sons, 2010.

[14] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10,

pages 181–189, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[15] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In *In Proceedings of the International Conference on Weblogs and Social Media (ICWSM 2009). AAAI*, 2009.

[16] C. van Rijsbergen. *Information Retrieval.* Butterworths, 1979.

[17] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR*, pages 28–36. ACM, 1998.