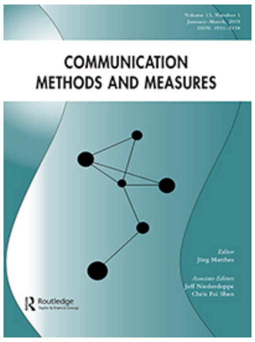

Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: https://www.tandfonline.com/loi/hcms20

Understanding News Story Chains using Information Retrieval and Network Clustering Techniques

Tom Nicholls & Jonathan Bright

To cite this article: Tom Nicholls & Jonathan Bright (2019) Understanding News Story Chains using Information Retrieval and Network Clustering Techniques, Communication Methods and Measures, 13:1, 43-59, DOI: 10.1080/19312458.2018.1536972

To link to this article: https://doi.org/10.1080/19312458.2018.1536972

Published online: 07 Nov 2018.

COMMUNICATION METHODS AND MEASURES
2019, VOL. 13, NO. 1, 43–59
https://doi.org/10.1080/19312458.2018.1536972

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Understanding News Story Chains using Information Retrieval and Network Clustering Techniques

Tom Nicholls[a] and Jonathan Bright[b]

[a]Reuters Institute for the Study of Journalism, University of Oxford, Oxford, United Kingdom; [b]Oxford Internet Institute, University of Oxford, Oxford, United Kingdom

ABSTRACT
Content analysis of news stories is a cornerstone of the communication studies field. However, much research is conducted at the level of individual news articles, despite the fact that news events are frequently presented as "stories" by news outlets: chains of connected articles offering follow up reporting as new facts emerge or covering the same event from different angles. These stories are theoretically highly important; they also create measurement issues for general quantitative studies of news output. Yet, thus far, the field has lacked an efficient method for detecting groups of articles which form stories in a way that enables their analysis. In this work, we present a novel, automated method for identifying news stories from within a corpus of articles, which makes use of techniques drawn from the fields of information retrieval and network analysis. We demonstrate the application of the method to a corpus of almost 40,000 news articles, and show that it can effectively identify valid story chains within the corpus. We use the results to make observations about the prevalence and dynamics of stories within the UK news media, showing that more than 50% of news production takes place within the form of story chains.

Content analysis of published news is one of the most common techniques in studies of mass communication and journalism. This analysis, which is frequently supported by large-scale manual coding efforts (and more recently by automated techniques), has underpinned investigations into many of the core theories in the field such as news values (Harcup & O'Neill, 2001), news agendas and agenda setting (Iyengar & Adam, 1993), news diffusion and sharing (Bright, 2016), gatekeeping and editorial decision making processes (Bright & Nicholls, 2014), and news readership dynamics (Graber, 1988), to give but a few examples.

A common simplification that is, to our knowledge, found in the vast majority of these studies is the coding of content at the level of the individual news article. For example, Harcup and O'Neill's widely cited work on news values is based on a coding of the most prominent article on every page of three British newspapers in March 1999 (Harcup & O'Neill, 2001, pp. 266–267). While article level work is of course valuable, it nevertheless marginalizes a second potential level of observation, which we call here news "story chains": events or single issues which receive repeated coverage in the news media through a series of initial articles and follow-up pieces. Story chains are narrower and more temporally restricted than news "topics," which are broad areas that also attract repeated news coverage. For example, "accidents and disasters" is a news topic often found in news media codebooks, while repeated coverage of one individual train crash would constitute a news story chain within the "accident and disaster" topic.

These story chains are theoretically significant in their own right because they have more potential impact than individual news articles, and indeed *prima facie* evidence suggests that the

CONTACT Jonathan Bright jonathan.bright@oii.ox.ac.uk Oxford Internet Institute, University of Oxford
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hcms.

news media itself reserves them for what are perceived to be the most important or significant events of the day. They also create measurement issues for quantitative studies of news in general. However, research using story chains has been limited thus far largely because of the practical difficulties of observing and measuring them. Manual identification of whether articles are linked is a simple task for a human coder, but also one that is prohibitively time expensive unless only a small amount are considered (as their accurate identification involves comparing all possible pairs of articles within a dataset). Current automatic content analysis tools, meanwhile, are not yet oriented toward the particular task of story chain detection.

In this article, we present a novel method for the automatic detection of story chains within news article datasets that seeks to resolve these problems. The method draws from two distinct fields: information retrieval for measuring textual similarity between articles, and network analysis for clustering articles into story chains. These fields have yet to be fully incorporated into the burgeoning work on automatic content analysis in communication studies, yet (as we will argue below) they offer considerable advantages for this particular task when compared with other more common approaches such as topic modeling. We also employ a moving window to reduce the computational complexity of the operation so that the method itself can be applied to a corpus of articles of any size without recourse to a high performance computing infrastructure. Application of the method would allow researchers specifically interested in story dynamics to select and study them without resorting to purposive sampling. It would also allow researchers interested in quantitative studies of news more generally to work with stories as a unit of analysis.

The rest of this article is structured in the following way. In Section 2, we discuss the concept of news story chains in more detail, highlighting the limited amount of existing literature on the subject, and showcasing why stories are theoretically highly important even if in practice they remain understudied. We also explore in more detail the practical difficulties involved in identifying story clusters with manual content analysis techniques. In Section 3, we look at existing automatic content analysis techniques within communication studies, and explain why they are also unsuitable to the task of story chain detection. In Section 4, we introduce our method, and describe the information retrieval and network analysis approaches on which it is based. Finally, in Section 5, we apply the method to a news article dataset containing almost 40,000 articles, and perform a variety of different validation tasks. The results show that the method performs well, and they also allow us to draw out some first order descriptive insights about the prevalence of stories within the UK news media.

Story Chains in the News Media: Definition, Theory and Potential Impact

In this article, we define news "story chains" as events or single issues which receive repeated coverage in the news media through a series of initial articles and followup pieces. These follow-ups could simply involve reporting of new facts which have emerged about an incident. For example, a violent crime might attract an initial news article simply reporting on its commission; there may then be a follow up as the police apprehend a suspect. Repeated coverage could also take the form of more in-depth analysis: for example, news of a film winning an Oscar might be followed up by an in-depth profile of the successful director. Follow-ups might also involve opinion pieces, or they may look at the same issue through a different "frame" which places emphasis on different aspects of the story (Entman, 1993). News stories are inherently time bound in nature: they begin in response to the emergence of a single issue and end when that issue ceases to be relevant. Indeed, we expect most stories to last just a few days as public interest declines and the news media moves on, a claim supported by empirical work on the "news cycle" (Patterson, 1998). Of course some stories will not fit this pattern: certain highly important events may receive repeated coverage over days, weeks, or even longer (see, e.g., Boydstun, Hardy, & Walgrave, 2014): but we expect these to be the exception rather than the norm.

News stories are conceptually distinct from news "topics," which we define as thematic news areas which also receive repeated coverage but which naturally encompass multiple events, and whose time span is much longer (indeed many news topics are essentially permanently recurring features of news coverage). For example, "health" would be an example of a news topic which is typically present in most news code books for content analysis (see, e.g., the codebook used in John, Bertelli, Jennings, & Bevan, 2013), whereas a high-profile medical malpractice lawsuit would be an example of a news story chain within the health topic area. Another example would be a group of articles discussing the results of the 2018 London Marathon: we would consider these a story chain which also falls within the wider topic of "sports."

The fact that the news media structures much of its reporting work around ongoing and follow up reporting is of course common knowledge. However, this structure is largely ignored in quantitative studies of the news media, where the level of observation is almost always the individual article. A few examples selected from recently published work will illustrate the point: Leupold et al.'s study of journalistic depictions of social cohesion is based on 1,300 individual articles selected from seven German local newspapers (Leupold, Klinger, & Jarren, 2018); Damstra and Vliegenthart's work on the framing of the economic crisis is based on articles whose headline made some reference to this crisis taken from three Dutch newspapers (Damstra & Vliegenthart, 2018); Nygren et al.'s work on coverage of the war in Ukraine uses the top five articles published in newspapers in four countries (Nygren et al., 2018); and Jóhannsdóttir's study of the extent of commercialization in the Icelandic press uses articles published in two print and two online outlets in three weeks selected from a four-year period (Jóhannsdóttir, 2018). Extending this list would be a straightforward task: for research which wishes to make a quantitative claim about news (for example, about the extent of coverage of a particular issue, or the type of frames through which an issue is presented), large-scale coding of individual news articles is essentially the default methodological choice.

The lack of focus on the story chain level of observation has several potential consequences. First and most obviously, it creates measurement issues for those wishing to make quantitative claims about the news. In work studying (for example) the distribution of topic coverage in news outlets, such as the Jóhannsdóttir piece above, five linked articles tackling a major political event will receive the same "weight" in the dataset as five separate articles about sports or celebrity gossip. However, even though the actual volume of articles is the same, one might argue that the production of one large, linked story is more consequential (or at least deserves to be measured in a different way) than five individual pieces. If researchers interested in quantitative news dynamics were able to sample observations at the story level, then this concern could be alleviated.

Second, it also means that researchers cannot study story production itself, which is theoretically interesting for a number of reasons. In research on news values and journalism, the potential for follow-up reporting is sometimes listed as a motivating factor for publishing an initial article. For example, spectacular crimes have been known to lend themselves to repeated coverage (Peelo, 2006), something that could be a motivating factor in publication of initial articles about them. Which events are suitable for follow-up is something that can evolve over time: research has highlighted, for example, that particular types of crime coverage have served as "prototypes" that are then repeated in later stories (Brosius & Eps, 1995). However, inability to work with articles at the story level has meant that research on which types of topic lend themselves most to follow-up reporting is in its infancy.

Once an initial article has been published, follow-up pieces also seem to become more likely, even if they could not be foreseen at the time of initial publication. Indeed, Harcup and O'Neill define "follow-up" as a news value in and of itself (Harcup & O'Neill, 2001), while Vasterman has claimed that the news "threshold" for follow-up articles may be lower than that for initial articles (Vasterman, 2005, p. 514). Research has also highlighted that those working in the news media may actively "manufacture" fresh angles for follow up stories on events they consider particularly worthy of coverage (Chadwick, 2011, p. 7). But, again, the possibility of studying this manufacturing behavior in a quantitative setting is limited by our inability to detect stories.

Also of importance in this regard is the small body of empirical work in communication studies that has focussed on the idea that news production appears to have two different "modes" (Boydstun et al., 2014): a mode of normal production, and a mode characterized by intense focus on a single issue. These moments of focus have been called, variously, "media hypes," "news waves," and "media storms" (Boydstun et al., 2014; Vasterman, 2005; Waldherr, 2014; Wien & Elmelund-Præstekær, 2009), but all of these terms capture the same basic premise: in response to certain events or incidents, the news media as a whole essentially abandon (or at least marginalize) normal routines of coverage to dedicate themselves to an exclusive focus on a particular current event, with multiple follow up pieces and different angles explored (in our terminology, these news "storms" would constitute an especially large type of news story chain, but not all story chains would be considered news storms). Major terrorist attacks (Entman, 2003), catastrophes such as the Challenger shuttle explosion (Riffe & Stovall, 1989), or political scandals (Chadwick, 2011) present examples of such media storms. These storms have significant potential consequences. Pieces of news which become media storms are more likely to reach widespread public attention as repeated publishing increases the likelihood that people are exposed to news, while those people who read multiple articles on the same event are likely to receive a signal about its importance. Furthermore, even in the age of online journalism, news production is still largely a "zero sum" game (Zhu, 1992), whereby increase in attention to one topic or event must imply decrease in attention to another. Boydstun et al. find that just over 11% of news coverage is, on average, attributable to very large "mega stories", that last for around 15 days on average (Boydstun et al., 2014, p. 520). Most seriously, perhaps, Vasterman has argued forcefully that media storms can often inflate a given news event beyond any objective measure of its actual importance and significance, as the news media start to involve themselves in a self-referential cycle which is detached from other ongoing events, such that "even the most trivial details [about the event] can become the most important news fact of that day" (Vasterman, 2005, p. 509). These moments are often when the news media is perceived as having the most influence on things like the political agenda (Van Aelst & Walgrave, 2011, p. 303). However our ability to study "news storm" behavior is again limited by our inability to detect the occurrence of such storms in the first place.

Existing Approaches to Automatic Content Analysis in Communication Research

Despite the theoretical importance of news story chains, in practice they have attracted relatively little empirical research. The main reason for this, we believe, is that the identification of groups of linked news stories would be prohibitively expensive in terms of researcher time (and is also not facilitated by current standard datasets for news research such as LexisNexis). Determining whether two articles address the same topic requires a researcher to perform a pairwise comparison between the two articles. The number of pairwise comparisons required to exhaustively evaluate a given dataset of $_n$ articles is $\frac{n(n-1)}{2}$. Even for a small dataset of 100 news articles (less than the amount produced on a typical news website in a typical day), fully 4,950 separate comparisons would have to be performed to detect all possible stories. More formally, this type of pairwise comparison can be said to take "quadratic time": i.e., the length of time required grows with the square of the number of input articles (often represented as $O(n^2)$ in computer science literature). This is because, as each new article is added to the dataset, it must be compared with all previous articles.

This difficulty of producing wide ranging and systematic story chain datasets is reflected by the methodological sacrifices made by research work up until now which has focused specifically on story chains. Of the work on news storms referenced above, Waldherr (2014) uses simulated data, while both Vasterman (2005) and Wien and Elmelund-Præstekær (2009) purposively sample news stories which are previously known to be important, which is a reasonable strategy for making initial observations but undermines potential generalizability. Boydstun et al. (2014) are the closest to being able to execute a fully quantitative approach, however their methodology relies both on a hand coded dataset of over 50,000 articles separated into more than 230 categories (a coding effort that would be extremely time consuming to reproduce) and also a heuristic method for detecting stories based on a sudden increase in the coverage of particular topics (a method that appears well suited to capturing some stories but that is very unlikely to capture all of them).

The method we introduce in this article is a form of automatic content analysis which seeks to resolve the above problems. The aim of the method is to automatically identify news story chains from within a "corpus" of news articles, and to distinguish these chain relationships from articles that are on similar thematic topics but relate to different events, thus enabling researchers to work with news story data. Automatic content analysis is of course a growth area in communications research (and indeed the social sciences more generally), inspired by advances in the fields of both corpus linguistics and machine learning. As such, automatic techniques themselves are nothing new. Before introducing our technique in more detail, it is therefore worth reviewing existing approaches to automatic content analysis, and highlighting why they are unsuitable for the task of detecting stories in large news datasets. In their overview article on the subject Grimmer and Stewart (2013) identify three main types of automatic content analysis technique which have so far been applied: dictionary methods, supervised methods, and automated (unsupervised) clustering. We will look at each of these in turn here.

The dictionary approach is probably the most widely used current approach to automatic content analysis, and also in some senses the simplest. Dictionary approaches involve manually developing lists of key words which relate to particular topics of interest: for example, the word "doctor" might be associated with a health topic, while the word "budget" might be associated with economic topics. Classification decisions are based on the appearance of keywords in a given document, typically weighted by the frequency of appearance. Examples of this approach can be found in work by Funk and McCombs, who study of the relationship between agenda setting and community structure effects on news production, using a system which detected concepts in news articles (such as "aggression") based on lists of keywords (Funk & McCombs, 2017); it can also be seen in the work of van Dalen et al. who use dictionaries to study tone and visibility of economic news (Van Dalen, de Vreese, & Albæk, 2017); a similar approach was also used by Soroka (2012) in his quantitative measurement of the impact of gatekeeping. Such a dictionary approach could, of course, be applied to the task of finding all articles related to a known individual news story of interest. However, it would be difficult to adapt them to the task of detecting story chains in general, because separate keyword lists would have to developed for each news story, which would require that the number and nature of stories were known in advance.

Supervised machine learning methods are similar in style to dictionary approaches. However, rather than requiring manual development of keyword lists, supervised machine learning works by asking human coders to separate a training set of documents into pre-existing categories of interest. One of a variety of algorithms can then be used to both extract "features" from documents in each category (which are typically words but could also be phrases, punctuation, document length, or anything else for that matter) and make classifications on the basis of these features. Supervised machine learning is increasingly popular because it is systematic, easy to validate, and a number of effective off-the-shelf methods are available. For example, Theocharis et al. use supervised classification to understand patterns of incivility in political discussions on social media (Theocharis, Barberá, Fazekas, Popa, & Parnet, 2016), while Colleoni et al. use supervised classification to detect different types of political orientation on Twitter (Colleoni, Rozza, & Arvidsson, 2014), and Bright and Gledhill use it to detect different types of journal article in a study of academic citation patterns (Bright & Gledhill, 2018). However, again, supervised machine learning techniques require definition of categories of interest before classification can be performed: in particular, to allow a training set of documents to be coded. This makes it unsuitable for detecting news stories which are unknown *a priori*.

Unsupervised methods are an alternative approach to classification (Guo, Vargo, Pan, Ding, & Ishwar, 2016; Roberts et al., 2014) that are also becoming increasingly popular in communication studies (see Maier et al., 2018), and which are conceptually quite similar to our method. Rather than defining categories of interest before the classification commences, these approaches aim to extract structure purely from the observation of the data. Two main families of approach exist: those based on "clustering," which involves the measurement of some kind of distance between documents (Manning, Raghavan, & Schütze, 2008, p. 321), and then the separation of documents into groups of clusters such that intra-cluster distance is minimized while inter-cluster distance is maximized; and those often referred to as "topic models" which attempt to derive a hidden set of topics from the observed distribution of words in documents. A good example of this type of approach can be found in the work of Quinn et al., who use a topic model to analyse the evolution of political attention in the U.S. Senate (Quinn, Monroe, Colaresi, Crespin, & Radev, 2010).

While these approaches satisfy the requirement of working without prior knowledge of the nature of the topics or clusters being developed, they nevertheless have limitations for the particular use case of identifying news story chains. Two difficulties are particularly worth highlighting. First, they typically require the researcher to specify the number of categories (k) to be used in advance. In practice researchers typically fit a number of models across a short range of values of $_k$ and then select between them by making at least some reference to the actual human interpretability of the model (see Maier et al., 2018, p. 10). This approach makes sense, but only if k is small enough such that the interpretation task is itself feasible. Our expectation is that, in a corpus containing tens of thousands of documents, $_k$ would have to be in the thousands to have a hope of capturing all stories, which would make manual interpretation an almost impossible task, or at least one which would be hugely time consuming. Methods for automatically selecting an appropriate value of k are developing (see, e.g., Greene, O'Callaghan, & Cunningham, 2014; Kanagal & Sindhwani, 2010), however they are yet to be perfected.

A second problem is that unsupervised models often make little use of very infrequent words. Typically, some cap is imposed on the number of words used (with the least frequently appearing being discarded) to limit the computational complexity of the model, under the assumption that very infrequent words will have little value in identifying broad topics. This is a valid simplification for the topic approach: words must appear with at least some frequency to be useful in defining topics. However, in detection of news story chains we expect that it is exactly the co-occurrence of rare words that will be the strongest marker of relatedness. For example, an unusual place name might appear in just two articles in the dataset: this would be a strong indicator that these two articles are part of the same story (but the word would probably be discarded in a topic model).

Automatically Detecting News Story Chains[1]

In this article, we therefore offer a new unsupervised approach to the detection of news story chains in corpora of news articles, based on techniques drawn from the field of information retrieval and network analysis. The approach is designed to work on a collection of articles of any size, although of course it is oriented toward large collections where manual coding would be completely infeasible (below we test it on a corpus of almost 40,000 articles). It is designed to identify connections both within and across outlets, and thus can address both repeated coverage in one outlet and in the media as a whole. Furthermore, and perhaps most importantly, it requires no *a priori* assumption about the number of news story chains within the corpus, which is a key improvement on existing unsupervised methods. The one assumption it does make is that stories are temporally regular rather than sporadic (i.e., coverage appears relatively continuously after publication of an initial article), and therefore the textual similarity of articles is only computed if they are within a certain time window of each other (we set the window to three days in our particular application). This assumption is required to allow the method to operate on a standard desktop or laptop computer within a reasonable period of time (although could potentially be relaxed if there was the possibility of access to a high-performance computing infrastructure).

The method involves two steps. First, we use information retrieval methods to measure the pairwise similarity between different news articles in the corpus: making use in particular of keyword

[1] All data and code used in the article are available from: http://dx.doi.org/10.5287/bodleian:R5qdeJxYA.

scoring and the BM25F algorithm (described in Section 4.1). As we will show below, these methods (applied in search engines such as Google) offer considerable conceptual advantages for our particular task, because they are oriented toward the aim of identifying very small amounts of potentially relevant content from within a large corpus. Second, we conceptualize these similarity measures as connections in a network of articles, and then use a clustering technique from network analysis (the Infomap algorithm) to detect related stories (as described in Section 4.2). Again, this network clustering method offers an important advantage for our purposes because it does not require the definition of the number of clusters (stories) in advance, and is agnostic as to the size of a cluster (hence a "cluster" may be composed of a single article, or 100 articles, or anything else).

Of course, it is important to note that the overall method (of using information retrieval methods for pairwise similarity and network clustering techniques for detecting related articles) is modular and other implementations would be possible. For example, it would be possible to replace the two approaches for calculating pairwise similarity described in Section 4.1 with any other similarity metrics: the important feature is that a set of pairwise linkages are made. Likewise, there are a large number of methods in the computer science and social networks literature for partitioning networks which could be applied to our similarity networks in place of the approach described in Section 4.2 (see Lancichinetti & Fortunato, 2009). Hence, although the approaches set out here are theoretically suitable and performed well on this story partitioning problem, other choices are possible.

Our approach is implemented in Python, and code is available from: http://dx.doi.org/10.5287/bodleian:R5qdeJxYA.

Calculating pairwise similarity with information retrieval approaches

Information Retrieval [IR] is a set of approaches, drawing on the tools of computer science, information science, and corpus linguistics, for accurately locating small amounts of relevant information in large data sources (Manning et al., 2008). Its most prominent modern application is in search engines such as Google, which are the archetypal IR system: from a given query, they need to identify and retrieve the most appropriate documents from a vast range of possibilities. Some of the tools used in IR overlap with the more general approaches already used by social scientists using large-scale text analysis. However, IR contrasts with more general classification approaches in that the task is to extract a small set of relevant documents from a much larger whole, rather than to partition all the available documents into a known number of groups. These techniques at heart look at the content and structure of underlying documents, indexing the information contained therein to allow answers to be given to arbitrary queries.

IR approaches are useful in our context because they provide a number of ways of thinking about the extent to which two documents are "similar." This similarity is at the heart of what it means to be a chain of articles all related to the same overall news event. They are also useful because they are designed for situations where the majority of comparisons will be negative (i.e., the majority of article pairs in a large corpus will not be related to each other).

We employ two IR techniques to develop our pairwise similarity measures.[2] First, we identify and score the most distinctive words in each article compared to the corpus as a whole by relative frequency, allowing documents to be labeled with key distinctive terms. Each term t is scored based on its frequency f in the document d and in the corpus of all text:

$$kwscore(t) = \frac{f_{t,d}}{f_t} \qquad (1)$$

[2] We also experimented with various other metrics here, including straightforward vectorspace cosine similarity and approaches based on co-occurrence of particular people and places identified with Named Entity Recognition. Results were substantially similar.

In our implementation, the words in each article are scored using Equation (1) and ranked, with the top 100 most informative terms for each article recorded.[3] This allows us to calculate a keyword similarity score for each pair of articles, which is simply the proportion of keywords which are common to both articles' top 100 list.

This scoring method selects strongly for the most unusual words in a given document. Although this is fairly naïve in IR terms, it is theoretically very suitable for news clustering. The intuition here is that news stories are about something concrete: a place, a person, or an event. By finding the most unusual terms in each article compared to the full output of the parent news source, it is possible to extract with some specificity the most distinctive words in each article. If articles share keywords, they are presumptively about the same subject. For this reason, we do not discard rare stopwords, as is conventional—here they are central to the theoretical justification for the method (see, more generally, Grimmer & Stewart, 2013, p. 273, for the importance of the research question in choice of approach).

For the second part of the pairwise similarity measurement, we use the BM25F scoring algorithm to select related articles. BM25F is an example of the class of *scoring rules*, which are used in information retrieval applications to identify which documents among many candidates are most relevant. BM25F is a standard general purpose scoring algorithm and we selected it because it is widely applied in the field of information retrieval and has been found to be successful in a variety of different tasks (Manning et al., 2008, p. 234). It is a development of Okapi BM25, which handles (as in our case) documents with multiple separate fields (body and title). Unlike the keyword approach identified above, which simply selects the most important terms in an article, BM25F scores documents in relation to a query. In a search engine context, this would be the text entered by the user; when attempting to find similar documents, it uses the content of the document being matched against as the query key.[4]

[5]The following BM25F equations are drawn from Pérez-Iglesias, Pérez-Agüera, Fresno, and Feinstein (2009):

$$BM25Fscore\,(q,d) = \sum_{t\,in\,q} \log\left(\frac{N - df(t) + 0.5}{df(t) + 0.5}\right) \cdot \frac{w(t,d)}{k_1 + w(t,d)}$$

where q is a given query and d a document, N is the number of documents in the collection and df is the number of documents in which the term t appears. The accumulated weight of a term over all fields $w(t,d)$ is calculated as follows:

$$w(t,d) = \sum_{c\,in\,d} \frac{occurs^{d}_{t,c} \cdot boost_c}{\left(\left(1 - b_c + b_c \cdot \frac{c^l}{avl_c}\right)\right)}$$

where l_c and avl_c is the average length for the field c, and $boost_c$ is the boost factor applied to field ${}_c$[5]. $k1$ and b are free parameters (with b free to vary between fields), which can be empirically selected to best improve the results of the subsequent objective function, or left at reasonable default values (such as $k1 \approx 1.2$, $b \approx 1$).

Our pairwise similarity measure is the mean of these these two measures: the proportion of keywords in common, and the BM25F score between the two articles. As we have remarked above, pairwise similarity calculations take polynomial time to produce (often denoted as $O(n^2)$). Although the automatic nature of the calculations means that time taken is less important than it would be for manual operations, it nevertheless can be a significant impediment to research work if run-time starts to be calculated in days or weeks.[6] A simple way of reducing this complexity is to only conduct

[3] The list is limited to terms with a $_{kwscore}$ of greater than 100, to avoid short articles generating spurious non-keywords.

[4] This is a simplification and elides the intermediate step of query expansion, but this can be automatically handled; for this work we used Bo1, one of the standard Bose-Einstein query expansion models, to create the final BM25F query from the text of the article being matched.

[5] We use $boost_{title} = 2$ and $boost_{body} = 1$ to give a modest increase in the importance of words in the title.

[6] The simple Python implementation of BM25F scoring used by the authors requires a few seconds of CPU time per comparison, which would quickly become infeasible on large datasets.

pairwise calculations within a moving window. News story chains are, we assume, intrinsically time-bound entities, with articles being released in close proximity. Hence by restricting the time window within which comparisons are made, we can restrict the overall run time of the method.

For this analysis, we have used a sliding window of three days. As the window is a continuously moving one, it does not prevent longer stories being grouped: the network partitioning method used below can group articles into larger blocks as long as the articles are indirectly connected via other articles which are within the window. The disadvantage of this approach is that it will, obviously, not identify stories whose publication arc takes place outside the time window in question. For example, a crime might lead to a prosecution weeks after its original development, and then a court case which takes place months or even years after that. However, we expect such sporadic story chains to be the exception rather than the norm.

In cases where this time window is perceived to be a problem, one potential response could simply be to increase the computational power deployed by, for example, implementing the approach on a framework such as Hadoop which allows a large number of processors to work simultaneously. Alternative techniques are also available for general-purpose complexity reduction in IR which require less subject knowledge but a stronger view of the initial query. One is use of the Boolean IR model, which simply uses unscored Boolean query matching/not matching (via computationally cheap lookup techniques such as hash tables of features) to select candidate texts to analyze with more sophisticated and expensive IR scoring methods (see Manning et al., 2008, Ch. 1).

Construction and partitioning of a similarity network

Having constructed some kind of metric-based way of scoring documents in response to a query, it is necessary to identify which documents are considered "matching." The two main approaches taken in IR are rank ordering and using a score cutoff. The first is familiar from search engine use: those pages the system has identified as "most relevant" are shown first to the user who entered the search query, and further pages of responses can be fetched until the user is satisfied with one of the pages retrieved. The second is more common in partitioning problems: having identified that certain documents are partially similar (or, using more probabilistic methods, have a given probability of being in the same group) then those under a certain threshold can be discarded, and those above presented as part of a set to the user. This threshold approach works well in a pairwise context (are these two items part of the same group or not?) but less well where there are multiple groups into which documents can be placed, and it is undesirable to allow cutoff scoring to potentially place documents into either zero or two plus groups each.

We hence use an alternative approach, of converting a matrix of pairwise similarity metrics into a similarity network, then applying network partitioning tools to supply the boundaries. Networks inherently arise in the context of object-by-object comparison, as these similarity judgments have the natural interpretation of relating how far apart two documents are in some sense. As these comparisons are made for increasing numbers of document pairs, the table of scores becomes equivalent to a list of weighted edges between nodes (where nodes are articles themselves), which is then tractable by standard methods for analyzing and partitioning networks.

The problem of taking a network-based representation of data and assigning the network nodes into groups is called community detection. There are a large number of community detection algorithms available (see Lancichinetti & Fortunato, 2009, for a discussion). We make use in particular of the Infomap method (Rosvall, Axelsson, & Bergstrom, 2009). This method aims to detect communities by modeling a random walk on the network, and by optimizing a quality function based on compression of the information contained within the network by minimizing the description length of the random walk (Lancichinetti & Fortunato, 2009, p. 4). We make use of this algorithm in particular because it appropriately handles network links which are both weighted and directed, because the partitions completed are fully hierarchical, and because it was the best performing method in the series of tests carried out by Lancichinetti and Fortunato on different network partitioning techniques (Lancichinetti & Fortunato, 2009; see also Fortunato, 2010, pp. 80–81). Additionally, no prior selection of the number of groups is required, which as we have described above is also a vital requirement for our application; Infomap will continue to create sub-clusters as long as the links between part of a cluster are stronger than those with the rest of it. Consequently, the optimum number of groupings is extracted from the data rather than being decided in advance by the researcher. However, as with the pairwise similarity method described above, the use of Infomap as the community detection algorithm is a choice, not a hard requirement, and another clustering method could equally be swapped in by a researcher with a different set of requirements.

The output of our method is a hierarchical clustering of articles into stories, with high-level groupings repeatedly split into smaller groups of stories and the final level being individual article nodes. The level of clustering varies—some top-level stories will be large, some will be small, and some will be a single article not detected as part of any given story. This highlights another major advantage of hierarchical clustering, which is that the researcher is able to select the level of abstraction that best suits the research problem tackled. In particular, if coarser granularity is desired then this can be done by selecting higher-level nodes at the analysis stage.

Validation of the method and descriptive results

We demonstrate the applicability of our methods on a corpus of 39,558 news articles collected over a three month period from a variety of different UK online news sources.[7] Articles were collected using an adapted version of the Heritrix web crawler (Internet Archive, 2013) as HTML documents directly from the websites in question. Post-processing was then performed on the HTML pages to extract only the article title and body text, removing images, videos, sidebars and other miscellaneous content. Following this process, we excluded articles which had less than 200 words of body text (these were typically image galleries or video content which had little accompanying text). We then conducted pairwise similarity calculations on all articles in the dataset within a three day moving window, and then separated these articles into story chains using Infomap as described above. Within the dataset, our method identified 5,753 story chains of a range of sizes, from a minimum of just two linked articles to a maximum of more than 100 (which was a story relating to three kidnappings in Ohio in the U.S.).

Validation of the results of our method is not a straightforward task as we have no reliable "ground truth" dataset of pre-existing, classified story chains to draw upon. Hence, we performed a variety of different validation tests, each one checking a particular aspect of the method. First, we look at the statistical performance of the matching algorithm from the perspective of two independently dual coded validation datasets, and also conduct a detailed qualitative examination of the areas where the classifier disagreed with the coders. Second, we compare overall descriptive results from the output of our method with some expectations generated by previous work in the field.

Complete datasets for both our validation and for replicating our descriptive results are available from: http://dx.doi.org/10.5287/bodleian:R5qdeJxYA.

Performance of the matching algorithm

We will begin by formally assessing the performance of the matching algorithm. The standard method for evaluating a classifier in the machine learning literature consists of assessing its performance on hand-coded "ground-truth" data. Hence, we constructed a validation dataset by hand-coding a series of 20,706 article pairs, which represent all possible pairwise comparisons of 204

[7] Articles were taken from from BBC News, The Mail Online, The Express, The Guardian, and The Mirror.

articles that were selected randomly from within a 3-day window. Each article pair was marked "related" or "not related" depending on the researcher's judgment about whether the articles were part of the same story chain, as defined in section 2. The full set of article pairs was independently dual coded by both authors in order to allow calculation of an inter-rater reliability metric. There was in fact perfect concordance between the researchers on the coding (i.e. there was 100% agreement, leading to a Krippendorff's alpha of 1.0). However, it was also the case that, out of the 20,706 article pairs, only 59 were found to be related to each other (i.e., part of the same story chain). This highlights again the difficulty described above in manually detecting story-chains in the news media: huge volumes of pairwise comparisons are required to find relatively small amounts of story chains. It also means that our validation dataset is quite unbalanced, and of more use for checking for the extent of "false positives" (i.e., the extent to which the classifier incorrectly connects articles together) than it is for "false negatives" (i.e., the extent to which the classifier misses articles which should have been connected).

The performance of the classifier on our test dataset was assessed using standard machine learning classification metrics, presented in Table 1. Alongside overall accuracy (the percentage of classification decisions made correctly), we also report the precision of the classifier (the proportion of items matched which are truly matching), recall (proportion of all truly matching items which are identified), and the F_1 statistic which is the harmonic mean of precision and recall. We also break down the precise numbers of false positives and negatives.

Overall, the pair-wise precision, recall, and F_1 results show that the classifier performs well. Only four false positives were generated, meaning that the precision of the classifier was high. There were ten false negatives, meaning that recall was lower than precision although still reasonable in relative terms. In other words, the technique seems to be quite reliable in finding story chains even among large volumes of unrelated content, and perhaps more importantly it produces very few false positives even in this highly unbalanced test where most articles are unrelated.

There are, however, some weaknesses with this approach. The absolute number of articles is quite small, even if the number of article pairs is very large. This reflects the difficulty of manually generating story chain data to which we have already referred. Related to this, the validation dataset contains relatively few "true positives" within it.

In response to these weaknesses we conducted a post-classification validation of the results of our method when applied to the full 39,558 article dataset. We randomly sampled 25 of the story chains generated by our method, then coded each article within each story chain according to whether it matched the rest of the articles in the chain. This provides a means of assessing the precision of each story chain: i.e. whether these articles truly belong together, and therefore gives us more of an idea of how many false positives are typically mixed in with true positives.

In total, 300 articles were assigned to these story chains, which were of a variety of sizes (ranging from several containing just a few articles to one large story chain which collected 57 articles together). Each article was independently checked by both authors and coded as either "belonging" to the cluster or not. There was a 93% agreement between the authors on this coding task, however

Table 1. Validation results for the classifier.

Accuracy	0.999
Precision	0.925
Recall	0.831
F_1	0.875
True Positive	49
True Negative	20,642
False Positive	4
False Negative	10
$N_{articles}$	204
$N_{articlepairs}$	20,706

Krippendorff's alpha was only 0.50. In particular, 265 of the 300 articles were coded as "true positives" by both authors, 13 of the 300 articles were coded as "false positives" by both authors, and the remaining 22 were coded as a "false positive" by at least one of the two authors. Considering this disagreement, we decided to take a restrictive approach for the purposes of evaluating the overall accuracy of our story chains, and therefore labeled all articles which at least one of the authors thought to be a false positive as an actual false positive. This is a strong test: the classifier is only judged to have performed correctly if both coders independently agree that it did not make an error. This leads us to estimate the precision of the data contained within our story chains as 88% (i.e., 265 out of 300). This is slightly lower than the precision metric reported in the previous section, and indeed we would consider this 88% to be a more reliable measure as it is based on a greater quantity of true and false positives.

Finally, we decided to look in detail at the articles which caused the classifier to make an error in the above two analyses. We also examined areas where the two coders disagreed in the second analysis. There were two main causes of error.

First, errors were witnessed when unrelated stories had unusual words in common. As we highlight above, the method is oriented toward matching on these unusual words, and will therefore likely match together articles which coincidentally share some features. For example, pairs of articles were matched together where the main protagonist had the same surname, or when a certain model of car featured prominently in the article. Another example concerned a group of four articles about kidney disease and kidney transplants, although only two of the four referred to the same story (of someone donating his kidney to a friend). Finally, we found (remarkably) a pair of articles about the uncovering of two different medieval skeletons from underneath two different car parks (with neither article referring directly to the other case).

Second, and more commonly, errors occurred where articles tackled the same broad topic or subject, but addressed different sub-stories. In these cases, the line between related and unrelated articles becomes somewhat blurry, depending on just how narrow a view we take of the idea of a story chain. One example of this was a set of articles discussing the announcement of the British Lions touring rugby team. The human coders considered each of these pairs to be related, as part of a "team announced for rugby tour" story. However, while some of the articles in the dataset took a general view and discussed the team as a whole, others focused on the human interest angle of individual players being selected and their back stories, meaning that the key words in each article were quite different. Hence, the classifier itself missed the connection. Another example concerns a pair of articles on separate aspects of the TV show "The Apprentice," which the classifier labeled as related but the human coders did not. While for the human coders the fact that the articles were separate was obvious, many of the key terms (such as the main protagonists in the show) were the same, hence the classifier made an error.

It is also worth referring to the reasons for the disagreements between coders in our analysis of 25 story chains above, and are hence the reason for the relatively low Krippendorff's alpha score. Again, borderline cases were the problem here. For example, several articles discussed film star Angelina Jolie's experiences of mastectomy, and several more discussed her husband (and also film star) Brad Pitt's new film, some of which made passing reference to the issue of mastectomy. One coder regarded these as related, while one did not. Another example here concerns English football club Wigan, which during the observation window won a major cup competition but also achieved several bad results in a league competition. One coder regarded these articles as related, while the other did not.

Overall, the analysis in this section demonstrates two things. First, it shows of course that the classifier is not perfect, and it does make errors. Researchers should be conscious that story chains detected using this method are likely to have incorrectly labeled articles within them. However, it also shows that these erroneously labeled articles are quite likely to be a small fraction of the overall number. Second, it also shows that there is inevitably some ambiguity in the definition of news story chains, and that researchers themselves will likely disagree about the precise boundaries of a story and how some stories may be distinguished from topics. However, again, this disagreement is not extensive: we believe that in general stories and topics are concepts which are quite clearly defined, even if there are inevitably some borderline cases.

Descriptive results from the full dataset

As a final validation of our technique, we compare some of the results we produce with descriptive findings obtained in the limited set of existing articles on the subject of news story chains. In this section, we make use of our classifier to find news story chains only within individual outlets (i.e., articles which are published and then receive a follow up in the same article). Although the method is also capable of finding connections between articles in different papers, the majority of the literature has focused on these intra-outlet chains. If our classifier is working well, we would expect it to produce descriptive statistics which are similar to the results in these pieces of research.

Existing works have made three basic descriptive claims about news story chains that can be tested by our data. First, in terms of prevalence of story chains, Harcup and O'Neill found around 30% of news articles are follow up pieces to an original in their study of a sample of articles from three British newspapers in March 1999 (Harcup & O'Neill, 2001, p. 1477). Although they do not specify how many original articles attracted these follow-ups, the fact that there are so many follow-ups suggests that perhaps 50% of the news publication agenda in any given outlet might be devoted to articles which form some part of a news story chain. In terms of duration, Boydstun et al. find that just over 11% of news coverage is, on average, attributable to very large "mega stories", which they find last for around 15 days (Boydstun et al., 2014, p. 520). Their study was based on articles published in one U.S. and one Belgian newspaper over a 10-year period. As our data covers all types of story chain, not just the very big ones, we would expect our average duration of stories to be lower: but we would nevertheless expect to find a significant quantity lasting for around 15 days. Finally, it has been argued that the temporal distribution of stories should be heavily right-skewed, with the majority of articles on the topic published shortly after the story breaks, while the volume of coverage then decays exponentially over time Vasterman (2005, p. 524). However, within this overall pattern, new smaller peaks may be observed as further developments emerge in the story (Wien & Elmelund-Præstekær, 2009, p. 197). Both of these articles were, it should be mentioned, based on purposively sampled news stories from within a news corpus: it remains to be seen, therefore, whether the pattern holds true for stories more generally.

In total there are 39,558 articles in our dataset. The classifier assigned 21,390 of these articles (54%) to a story chain within their own outlet (i.e., they were associated with at least one other article in the venue in which they are published). This figure is quite similar to the one suggested above by Harcup and O'Neill. The distribution of the size of these story chains (see Table 2) is heavy tailed, with the majority (69%) having between 2 and 5 articles. The larger story chains (with 11 articles or more) account in total for around 11% of the dataset, which is very similar to Boydstun et al.'s finding that 11% of news coverage is attributable to large stories.

The average story in our dataset lasts 1.5 days. Larger stories are unsurprisingly longer ones as well, though once stories go above 10 articles in size there is not a great deal of difference between them: lasting on average around 4–4.5 days. Only one story in the dataset lasted for more than 10 days, and no story lasts for the average of 15 days identified by Boydstun et al. (a point to which we will return below). The average evolution of large stories is shown in Figure 1, which graphs the amount of time after an initial publication it took for new articles to appear. The right skew and periodicity predicted by Vasterman and Wien and Elmelund-Præstekær is clearly observable, with new peaks roughly corresponding to daily news cycles.

In other words, in a variety of respects our findings seem to confirm other descriptive results produced on the subject. The only area of disagreement concerns the length of the stories, with our stories typically being shorter than "mega stories" identified in other research. Although we expected our stories to be shorter on average, it is nevertheless significant that none of the stories identified by

Table 2. Story chain sizes.

Story size	N	Articles in these stories	As % of totalarticles	Average duration (days)
2–5	5,620	14,795	69%	1.2
6–10	563	4,165	19%	3.1
11–20	145	1,990	9%	4.1
21–30	11	271	1%	4.5
30+	4	169	1%	4.4
Total	6,343	21,390	100%	1.5

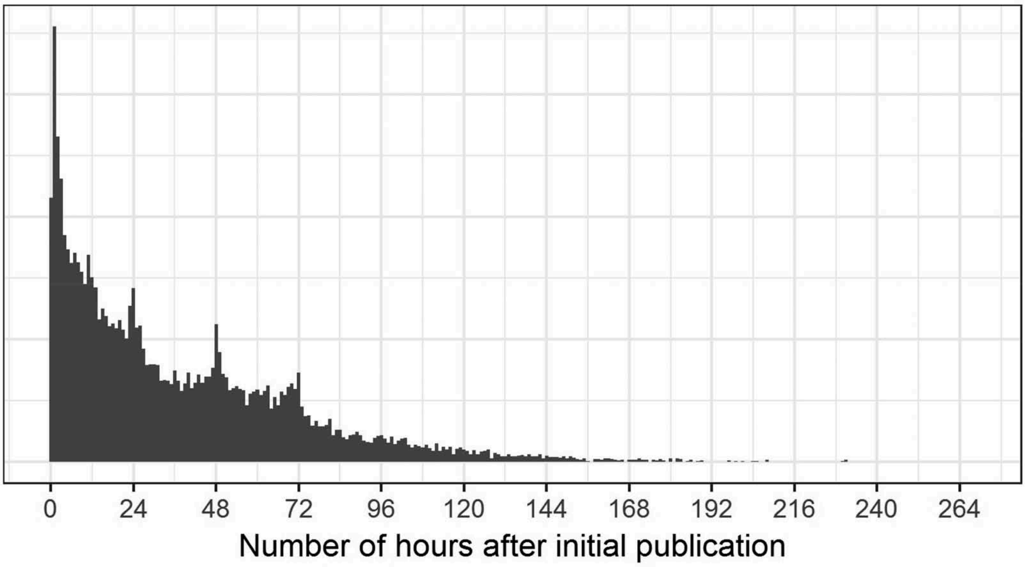

Figure 1. Time between initial publication and follow-up articles.

our technique were as long as the ones identified by Boydstun et al., suggesting that the method has a problem with identifying very large stories.

In order to understand why, we conducted some further manual inspection of the dataset and the resulting story clusters, focusing in particular on news events which we suspected would have generated at least 15 days of coverage. During the period of data collection, there were two such stories: the bombing of the Boston marathon and the terrorism related murder of Lee Rigby, a British soldier in London. When we examined the hierarchical clustering of news articles related to these topics, we found that the method had separated out these large stories into distinct "sub stories," each one relating to a different facet of the incidents themselves. This explains why the size of the stories themselves was smaller than hypothesized.

The reason for this behavior is that the links within the smaller groups that make up the overall the story are stronger than those between random articles from within the story as a whole. The Infomap algorithm, using information theoretic principles to extract structure, will repeatedly subdivide larger stories as long as it can identify suitable sub-stories. This is not purely a methodological point, of course: indeed it raises theoretical questions about the proper scope of a story. In the Boston case, the articles assigned to the sub-stories identified really do have more in common with each other than with the Boston story as a general whole (for example, coverage of the immediate aftermath of the bombing is quite distinct from coverage of the arraignment). Whether "Boston" is the desired unit of analysis, or the more granular parts of it are, is a matter for the researcher. Interestingly however, the use of network methods to create our original clustering may also offer the opportunity to partially resolve this situation, by indicating not only relations between articles but also relations between story chains. By allowing the researcher to select from the hierarchical clustering on the basis of substantive concerns and the desired level of abstraction, the approach hence provides flexibility as well as rigor. However, it is certainly the case that some manual work would be required on the part of the researcher to both inspect these hierarchical clusterings and to decide the appropriate level of aggregation. Hence, it would only be feasible to tackle this in the case of individual large stories (rather than across the dataset as a whole).

Discussion

The grouping of articles into chains of related stories by computational methods is an interesting open research problem, of significant value to communications researchers studying news output at scale. This paper has introduced a new method, based on information retrieval and network analysis tools, which gives researchers the opportunity to perform quantitative work which uses the news story as a unit of analysis in addition to the article. The obvious application of this work would be to enable studies of news stories themselves, i.e., when and why news outlets choose to engage in repeated coverage; but there are many other questions which would benefit from the ability to handle grouped article data. For example, studies of the dynamics of issue frame competition (see, e.g., Gamson & Modigliani, 1989; Guggenheim, Jang, Bae, & Neuman, 2015) would benefit from the ability to group together articles covering the same event, as would studies of "churnalism," that is, duplicated bits of text which are continually repackaged as novel articles (see, e.g., Jackson & Moloney, 2016; Nicholls, 2018). Indeed, we suggest that for any quantitative study of news production, the ability to work at the story level would be a useful complement to existing article focussed work, and would allow novel and potentially more realistic measures of the extent of coverage of certain issues to be developed.

It is also worth highlighting that the technique could be fruitfully combined with a post-classification manual analysis. If a particular piece of research is aided most by maximizing precision at the expense of recall (because you need all identified items to be similar, but don't need to identify all similar items) then a simple manual analysis of the identified pairwise connections could increase precision to arbitrarily high levels. In this case, the information retrieval tools would be acting as an automated filter, running through the huge number of possible links and identifying a tiny percentage for manual analysis. Although not fully automated, this approach is perfectly valid in many situations.[8]

This method is in no sense the final word on the subject. It achieves good results in terms of validation, and its subjective results are largely consistent with previous work in the area. But it's also clear that there remains much scope for further work: a large number of possible approaches could be taken to the pairwise matching step, and there are several network partitioning approaches that could be applied. Future research could usefully investigate the extent to which these approaches might offer improved precision in the story chains identified, or perform better when linking large stories together. It would also be useful to experiment with alternative approaches to validation: as we highlight above, our ground-truth validation dataset is far from perfect. As research continues in these areas, it will enable more research on the subject of news story chains in the field of communication research.

Acknowledgments

We thank colleagues from the Political Methodology Specialist Group of the Political Studies Association who commented on an earlier presentation of this method at their 2017 meeting, and to the anonymous reviewers for Communication Methods and Measures for their insightful feedback.

[8] It also reflects how search engines use information retrieval techniques in the real world. Google needs to provide a great result on the front page, but if only 9 out of 10 of its links are to suitable sources then the human user will simply ignore those which are obviously wrong.

Funding

The research was partly supported by supported by the Economic and Social Research Council [grant number 1219081], and by Google UK as part of the Google News Initiative.

ORCID

Tom Nicholls http://orcid.org/0000-0002-6971-8614

References

Boydstun, A. E., Hardy, A., & Walgrave, S. (2014). Two faces of media attention: Media storm versus non-storm coverage. *Political Communication*, *31*(4), 509–531. doi:10.1080/10584609.2013.875967

Bright, J. (2016). The social news gap: How news reading and news sharing diverge. *Journal of Communication*, *66*(3), 343–365. doi:10.1111/jcom.12232

Bright, J., & Gledhill, J. (2018). A divided discipline? Mapping peace and conflict studies. *International Studies Perspectives*, *19*(2), 128–147. doi:10.1093/isp/ekx009

Bright, J., & Nicholls, T. (2014). The life and death of political news: Measuring the impact of the audience agenda using online data. *Social Science Computer Review*, *32*(2), 170–181. doi:10.1177/0894439313506845

Brosius, H.-B., & Eps, P. (1995). Prototyping through key events. *European Journal of Communication*, *10*(3), 391–412. doi:10.1177/0267323195010003005

Chadwick, A. (2011). The political information cycle in a hybrid news system: The British prime minister and the "bullygate" affair. *The International Journal of Press/Politics*, *16*(1), 3–29. doi:10.1177/1940161210384730

Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication*, *64*(2), 317–332. doi:10.1111/jcom.12084

Damstra, A., & Vliegenthart, R. (2018). (Un)covering the economic crisis? *Journalism Studies*, *19*(7), 983–1003. doi:10.1080/1461670X.2016.1246377

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. doi:10.1111/j.1460-2466.1993.tb01304.x

Entman, R. M. (2003). Cascading activation: Contesting the white house's frame after 9/11. *Political Communication*, *20*(4), 415–432. doi:10.1080/10584600390244176

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3–5), 75–174. doi:10.1016/j.physrep.2009.11.002

Funk, M. J., & McCombs, M. (2017). Strangers on a theoretical train. *Journalism Studies*, *18*(7), 845–865. doi:10.1080/1461670X.2015.1099460

Gamson, W. A., & Modigliani, A. (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, *95*(1), 1–37. doi:10.2307/2780405

Graber, D. (1988). *Processing the news: How people tame the information tide*. London, United Kingdom: Longman Group.

Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models. In T. Calders, F. Esposito, E. Hüllermeier, & R. Meo (Eds.), *Machine learning and knowledge discovery in databases*. ECML PKDD 2014. Lecture Notes in Computer Science (Vol. 8724, pp. 498–513). Berlin, Heidelberg: Springer. doi:10.1007/978-3-662-44848-9_32

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. doi:10.1093/pan/mps028

Guggenheim, L., Jang, S. M., Bae, S. Y., & Neuman, W. R. (2015). The dynamics of issue frame competition in traditional and social media. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 207–224. doi:10.1177/0002716215570549

Guo, L., Vargo, C. J., Pan, Z., Ding, W., & Ishwar, P. (2016). Big social data analytics in journalism and mass communication: Comparing dictionary-based text analysis and unsupervised topic modelling. *Journalism & Mass Communication Quarterly*, *93*(2), 332–359. doi:10.1177/1077699016639231

Harcup, T., & O'Neill, D. (2001). What is news? Galtung and ruge revisited. *Journalism Studies*, *2*(2), 261–280. doi:10.1080/14616700118449

Iyengar, S., & Adam, S. (1993). News coverage of the gulf crisis and public opinion. *Communication Research*, *20*(3), 365–383. doi:10.1177/009365093020003002

Internet Archive (2013). Heritrix [Computer software]. Retrieved from http://builds.archive.org:8080/maven2/org/archive/heritrix/heritrix/3.1.1/

Jackson, D., & Moloney, K. (2016). Inside churnalism. *Journalism Studies*, *17*(6), 763–780. doi:10.1080/1461670X.2015.1017597

Jóhannsdóttir, V. (2018). Commercialization in the Icelandic Press: An analysis of hard and soft news in major print and online media in Iceland in times of change. *Journalism*. doi:10.1177/1464884918768494

John, P., Bertelli, A., Jennings, W., & Bevan, S. (2013). *Policy agendas in British politics*. London, United Kingdom: Palgrave Macmillan.

Kanagal, B., & Sindhwani, V. (2010). *Rank selection in low-rank matrix approximations: A study of cross-validation for NMFs*. Retrieved from https://www.cs.umd.edu/~bhargav/nips2010.pdf

Lancichinetti, A., & Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E*, *80*(56117), 1–11. doi:10.1103/PhysRevE.80.056117

Leupold, A., Klinger, U., & Jarren, O. (2018). Imagining the city. *Journalism Studies*, *19*(7), 960–982. doi:10.1080/1461670X.2016.1245111

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. doi:10.1080/19312458.2018.1430754

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, United Kingdom: Cambridge University Press.

Nicholls, T. (2018). *Churnalism, press releases and wire copy: Detecting textual reuse in large news corpora*. International Communications Association 2018 annual conference. Prague.

Nygren, G., Glowacki, M., Hök, J., Kiria, I., Orlova, D., & Taradai, D. (2018). Journalism in the Crossfire. *Journalism Studies*, *19*(7), 1059–1078. doi:10.1080/1461670X.2016.1251332

Patterson, T. E. (1998). Time and news: The media's limitations as an instrument of democracy. *International Political Science Review*, *19*(1), 55–67. doi:10.1177/019251298019001004

Peelo, M. (2006). Framing homicide narratives in newspapers: Mediated witness and the construction of virtual victimhood. *Crime, Media, Culture: An International Journal*, *2*(2), 159–175. doi:10.1177/1741659006065404

Pérez-Iglesias, J., Pérez-Agüera, J. R., Fresno, V., & Feinstein, Y. Z. (2009). Integrating the probabilistic models BM25/BM25f into lucene. *arXiv:0911.5046*.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228. doi:10.1111/j.1540-5907.2009.00427.x

Riffe, D., & Stovall, J. G. (1989). Diffusion of news of shuttle disaster: What role for emotional response? *Journalism Quarterly*, *66*(3), 551–556. doi:10.1177/107769908906600303

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open ended survey responses. *American Journal of Political Science*, *58*(4), 1064–1082. doi:10.1111/ajps.12103

Rosvall, M., Axelsson, D., & Bergstrom, C. T. (2009). The map equation. *European Physical Journal Special Topics*, *178*, 13–23. doi:10.1140/epjst/e2010-01179-1

Soroka, S. N. (2012). The gatekeeping function: Distributions of information in media and the real world. *The Journal of Politics*, *74*(2), 514–528. doi:10.1017/S002238161100171X

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. doi:10.1111/jcom.12259

Van Aelst, P., & Walgrave, S. (2011). Minimal or massive? The political agenda setting power of the mass media according to different methods. *The International Journal of Press/Politics*, *16*(3), 295–313. doi:10.1177/1940161211406727

Van Dalen, A., de Vreese, C., & Albæk, E. (2017). Economic news through the magnifying glass. *Journalism Studies*, *18*(7), 890–909. doi:10.1080/1461670X.2015.1089183

Vasterman, P. L. M. (2005). Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems. *European Journal of Communication*, *20*(4), 508–530. doi:10.1177/0267323105058254

Waldherr, A. (2014). Emergence of news waves: A social simulation approach. *Journal of Communication*, *64*(5), 852–873. doi:10.1111/jcom.12117

Wien, C., & Elmelund-Præstekær, C. (2009). An anatomy of media hypes: Developing a model for the dynamics and structure of intense media coverage of single issues. *European Journal of Communication*, *24*(2), 183–201. doi:10.1177/0267323108101831

Zhu, J.-H. (1992). Issue competition and attention distraction: A zero-sum theory of agenda-setting. *Journalism Quarterly*, *69*(4), 825–836. doi:10.1177/107769909206900403