A Computational Approach for Examining the Comparability of 'Most-Viewed Lists' on Online News Sites

Rodrigo Zamith

Abstract

This study introduces a computational approach for evaluating the lists of most-viewed items present on the homepages of many news organizations, focusing on two dimensions: the list's rate of change over the course of the day and the median time it takes a news item to appear on the list. That approach is then applied in an analysis of 21 news organizations over two months, revealing clusters across those dimensions which indicate the reporting of different data. Scholars are ultimately encouraged to perform their own analyses and cautioned against assuming the lists are comparable just because they appear alike.

Keywords

analytics, metrics, page views, methodology, computational content analysis

Scholars of digital journalism have, in recent years, taken growing interest in the relationship between page views and news content. This is partly due to the increasing prevalence of audience analytics, which allows data to be captured on a micro scale. For example, on an exploratory level, scholars have described the kinds of content that tend to yield high amounts of page views from readers and that which is deemed important by editors (e.g., Boczkowski & Mitchelstein, 2013; Schaudt & Carpenter, 2009; Kormelink & Meijer, 2017). On an explanatory level, scholars have treated page views both as an outcome variable (e.g., the effect of a story's prominence on the amount of page views it receives, see Lee, Lewis, & Powers, 2014) and as a predictor variable (e.g., the effect of the amount of page views an item receives on its likelihood of remaining on a page at a later point in time, see Bright & Nicholls, 2014).

The number of page views an item receives is often treated as a key indicator of the broader concept of popularity (e.g., Boczkowski & Peer, 2011; Lee et al. 2014; Tenenboim & Cohen, 2015) in light of the metric's central role within newsrooms (Anderson, 2011; MacGregor, 2007; Usher, 2012). Page views are easy to capture given the very nature of networked systems (Andrejevic, 2007; Kaushik, 2009), and news organizations invariably adopt systems like Google Analytics and Chartbeat to perform that task (Boczkowski & Mitchelstein, 2013; Cherubini & Nielsen, 2016; Graves & Kelly, 2010). These data may be recorded directly by the news organization whenever a page is requested by a client (i.e., reader) as well as by a third party through the inclusion of a small piece of code on a webpage.

Although page view data are generally readily available to editors and managers (and, in some cases, journalists) at those organizations (Hanusch, 2016), those data are often out of the reach of scholars. Such data may typically only be accessed through an agreement with the news

organization, and that access is often limited since news organizations may be reluctant to share such data given their potential commercial implications (see Napoli, 2011; Turow, 2005).

In lieu of that prized data, scholars often turn instead to the lists of popular items—typically titled 'most viewed,' 'most clicked,' or 'most popular'—that appear on the websites of many news organizations. The implications of using such lists as proxies for popularity have not received a great deal of scholarly attention, however. This is problematic because studies that use these lists may adopt an implicit assumption that they invariably represent similar kinds of data. For example, they may assume that these lists cover the same period of time (e.g., popularity over the past day) and that they are automatically updated with the same frequency (e.g., every hour). Complicating matters, news organizations rarely provide sufficient information on their websites to ascertain those key considerations. The limitations of scholars' understanding of these lists are especially problematic for comparative research, wherein divergences in the findings could be due to differences in what the data captures.

The present study adds to the understanding of what these lists represent by first reviewing the literature around the concept of 'liquidity' and the use of page views as a variable in digital journalism scholarship. A computational approach for extracting and analyzing content from lists of most-viewed items is then described. That approach focuses on two indicators that are useful for examining the comparability of lists of most-viewed items: the list's rate of change over the course of the day and the median time it takes a news item to appear on the list. The approach is then applied in an analysis of the lists of most-viewed items from 21 news organizations over two months. In doing so, contributions are made to the literature by examining assumptions about the similarities of lists across several news organizations and by providing a flexible approach for scholarly analysis that can adapt to evolving news websites.

Literature Review

As scholars of digital journalism have observed, online news can change continuously and unpredictably (Deuze, 2008; Karlsson & Strömbäck, 2010), resulting in "liquid news stories being published in different drafts … and essentially consisting of ever-changing and unruly processes" (Karlsson, 2012a, p. 388). In contrast to newspapers that offer a single snapshot of the previous day's news, online news sites may change throughout the day as stories develop and to fit to readers' preferences (Lim, 2012; Zamith, 2016b). For example, Boczkowski and de Santos' (2007) work suggests that online news sites may present more hard news in the evenings than in the mornings. More recently, Widholm (2016) found variations in the amount and type of content published online over the course of the day.

These observations point to the importance of considering the implications of a liquid Web, both in terms of how materials may change significantly depending on the parameters used to access the content and in terms of how the 'black box' aspect of journalism may be peered into through the study of the evolution of content (Deuze, 2008; Karlsson, 2011). For example, Karlsson (2012b) found that media framing of the swine flu epidemic would change continuously over the course of the day, especially during the initial stages of reporting. Saltzis (2012) similarly found that news content can change considerably over a story's lifespan, with updates being most frequent within the first two hours. Saltzis (2012, p. 704) thus argues of digital news content, "the continuously updated news story stops being a fixed entity, the 'final product' of the work of journalists, and it becomes always evolving and fluid."

The study of "liquid" content—rapidly changing digital artifacts—is an emerging area within media studies that has already pointed to a number of challenges associated with

traditional content analyses (Karlsson & Sjøvaag, 2016; Sjøvaag & Stavelin, 2012; Zamith, 2016a). One key challenge is to identify efficient ways to "freeze" liquid content into static objects that can be analyzed independently from time (Karlsson & Strömbäck, 2010; Widholm, 2016). Given that online content may theoretically refresh every time a page is loaded, researchers typically seek out computational solutions that are not only more efficient but also more reliable (Lewis, Zamith, & Hermida, 2013; Sjøvaag, Stavelin, & Moe, 2016; Zamith & Lewis, 2015).

The list of most-viewed items that is found on the homepages of many news organizations represents an object of interest to media scholars that is particularly liquid. Indeed, the prospect of frequently updated data reflecting audience behaviors may only be realized if that object is able to change constantly, which in turn creates a methodological headache for scholars. However, the ability to freeze such data offers a compelling reward: a better understanding of audience preferences at a time when audiences feel empowered and are becoming increasingly important in news production (Adornato, 2016; Holton, Lewis, & Coddington, 2016).

The number of page views a news item receives is often employed by scholars of digital journalism as the primary measure of that item's popularity (e.g., Boczkowski, Mitchelstein, & Walter, 2011; Boczkowski & Peer, 2011). On its face, this is a reasonable indicator: if a large number of news consumers clicked on an item, it is likely because that item had general appeal on some level (cf., Kormelink & Meijer, 2017). Furthermore, in studies involving newswork and newsworkers, the number of page views an item receives is particularly useful because much of the literature on audience metrics has found that page views serve as the dominant metric in newsrooms, and that the notion of popularity is often expressed in terms of page views in those environments (Anderson, 2011; Groves & Brown, 2011; MacGregor, 2007; Usher, 2012). Thus,

while the abstract concept of popularity may be operationalized through different measures (e.g., number of shares on social media or time spent on page)—either in isolation or as a multi-dimensional construct—studies exploring that notion from the perspective of newsworkers typically rely on the number of page views an item receives as it is viewed as the de facto measure popularity (e.g., Bright & Nicholls, 2014; Lee et al., 2014; Tenenboim & Cohen, 2015; Zamith, 2016b).

Scholars of digital journalism rarely have access to detailed, ratio-level data on page views (i.e., the specific number of page views an item receives), largely because of the potential commercial implications of that data for news organizations (see Napoli, 2011; Turow, 2005) and because of the technical challenges associated with making real-time data available in a format that can be readily used by researchers (Graves & Kelly, 2010). As such, researchers typically rely on ordinal-level data, in the form of rankings, that appear on the homepages of many news organizations through computer-generated lists of the site's most-viewed items. These data may convey that one news item received more or fewer page views than another, but they do not provide the absolute magnitude of the difference. Instead, they offer equidistant intervals (e.g., most popular and second most popular).

Despite their limitations, data from those lists have been used extensively and to good effect. Boczkowski and Peer (2011) used data from such lists to demonstrate a gap in the preferences of journalists and news consumers when it came to the subject matter and format of news stories. Boczkowski et al. (2011) used such data to show that a similar thematic gap persisted across six countries in Western Europe and Latin America. Looking beyond just the most-viewed items, Boczkowski and Mitchelstein (2012) used data from the lists of most clicked, most e-mailed, and most commented stories to assess the differences between those

forms of interactivity as they related to the subject matter of stories and whether the story occurred during periods of routine or heightened political activity. Drawing on the *New York Times*' list of most-viewed items, Lee et al. (2014) used data from the lists of most-viewed items of three different U.S.-based news organizations to show that the popularity of an item had an effect on its subsequent news placement, but that placement had no effect on the number of clicks an item subsequently received. Welbers et al. (2015) used those rankings in an analysis of five Dutch news organizations, finding that the popularity of a story impacted the likelihood it would receive follow-up reporting. Bright and Nicholls (2014) used rankings in their analysis of five U.K.-based outlets to show that, relative to their non-popular counterparts, popular news items had a lower risk of being removed from the homepage at a later point in time.

While these studies have collectively offered scholars a better understanding of the kinds of content that tend to be popular and how popularity influences, and is influenced by, other factors, they offer limited insight into the comparability of those lists. Specifically, all of the aforementioned studies engage in some form of comparative work, yet only one attempts to assess the comparability of the data source. That study, by Bright and Nicholls (2014), established that items appearing on the list of most-viewed items for five U.K. news organizations typically appeared there before they were removed from the page, leading them to conclude that "most read lists do provide a reasonably accurate picture of what is currently popular on the site, rather than what was popular over the last few days" (p. 176). More often, however, there appears to be an implicit assumption that those lists invariably represent the same kind of data, such as the time period covered by the list and the frequency with which the list is updated. Indeed, such an assumption would be necessary for those data to be comparable.

This assumption, however, may be problematic: different organizations use different software to gather traffic information and different content management systems to display content on their homepages, restricting what and how particular metrics may be displayed (Cherubini & Nielsen, 2016; Anderson, 2011; Graves & Kelly, 2010). Moreover, a lengthy body of literature on the social construction of technology has found that the meaning assigned to a technology and the perceptions of the environment external to that technology becomes central to how that technology is engaged with (Kraut, Rice, Cool, & Fish, 1998; Markus, 1994; Orlikowski, 2000). For example, some organizations may find their readers are better served by listing the stories that are trending (i.e., recently popular) while others favor listing stories that were popular over the past day or week (i.e., 'in case you missed it').

The prevalence of such lists as data sources in scholarly research and the potential that those lists represent different data demands an empirical evaluation of the potential implications of using those lists and the extent to which they may be comparable among oft-studied media. In order to perform that evaluation, it is necessary to first develop a process for systematically capturing and storing data from the liquid lists of most-viewed items. One must then identify empirical dimensions for comparing lists in order to assess whether they represent similar data.

Karlsson and Strömbäck (2010) have pointed to the promise of "mirroring" software that create copies of given webpages at predetermined intervals. Those software, in combination with custom computer scripts for organizing the resulting files, have been used by researchers to systematically create duplicate copies of webpages at predetermined intervals (see Hermida, Lewis, & Zamith, 2014; Sjøvaag, Moe, & Stavelin, 2012; Sjøvaag et al., 2016; Widholm, 2016). However, as Zamith (2016a) observes, oft-used "mirroring" software like HTTrack, WebCopy, and wget are becoming less useful for analyzing certain aspects of the modern web because

organizations are increasingly using JavaScript-based technologies to add interactive features to their websites, which those software fail to process.[i] Zamith proposes using an approach that emulates a full browsing session (e.g., automating Mozilla's Firefox browser) in order to create an exact replica—JavaScript features included—of liquid content viewed under predefined conditions. This software can be paired with computer scripts written in general-purpose languages like Python to create and organize snapshots (Sjøvaag et al., 2016).

Once liquid objects have been frozen into static ones, the researcher may then analyze the features of interest. Lists of most-viewed items are the output of machine-generated HTML code. Sjøvaag and colleagues (2012; 2016) and Zamith (2016) have pointed to the BeautifulSoup library for Python as being particularly useful for extracting features from HTML documents. BeautifulSoup can be used to turn the HTML code into a parseable object that can be navigated and searched, allowing information to be easily scraped from each snapshot. That information may then be stored in a structured text file (Sjøvaag & Stavelin, 2012) or relational database (Zamith, 2016a), and analyzed using statistical software like R and SPSS.

An Approach for Evaluating Most-Viewed Lists

The insights derived from the emerging literature on liquidity offer a sound foundation upon which one may develop an approach that enables the researcher to assess different dimensions of a list of most-viewed items, and comparability among multiple lists. The development and assessment of computational approaches for handling liquid content has been identified by Karlsson & Sjøvaag (2016) as being particularly important amid the rapid shift toward digital media production and consumption. Heeding this call and building on the existing

literature, a three-step process is proposed wherein the information is systematically downloaded, then computationally parsed, and finally statistically analyzed.

Many news organizations have multiple lists denoting the popularity of items. For example, *The New York Times* allows the reader to select between the list of items that were most e-mailed and the list of items that were most viewed. Similarly, *The Denver Post* allows the selection of a general list of most-viewed items, or lists specific to news, sports, business, arts and entertainment, and lifestyle. In order to select the desired list, the user must click on the appropriate heading, which initiates a JavaScript call to load the requested information in a specific area without refreshing the whole page. Additionally, some news organizations use JavaScript to load the initial list. Thus, in order to create a complete replica of the page, a mirroring solution that can process JavaScript and simulate user actions is sometimes required. Building on Zamith (2016), Selenium—a tool for automating user input and actions in browsers like Mozilla's Firefox and Google's Chrome—is recommended as it allows for the complete emulation of a typical client and couples nicely with the popular Python programming language.

Because some news organizations feature a single list of most-viewed items, or default to the primary list, a single, general purpose script that indicates which website to visit and where to store the snapshot is sufficient to complete the first step for many news organizations. When user input is required to ensure the right list is selected, additional code may be added to this base script to instruct the browser to click on certain elements based on unique identifiers or their XPaths (an element's position within the HTML document's structure). For example, one could instruct Firefox, via Selenium, to perform a click on the *Plain-Dealer*'s homepage by using the `find_element_by_xpath()` function, pointing it to the XPath of

`id('river_nav_inner')//li[@data-value='popular']`, and chaining the `click()` function. For complete code demonstrations, see the URL at the end of this section.

Once a snapshot is created, the information of interest may be extracted. Because the layout of each organization's website is different, individual scripts usually need to be developed, though much of the code can be reused. Like Sjøvaag et al. (2012; 2016) and Zamith (2016a), the BeautifulSoup library was found to be especially useful for turning the HTML code into a traversable object. The area containing the relevant list of most-viewed items must then be isolated, usually using the unique identifier of the "div" element containing the information, a unique combination of "class" attributes specific to that "div", or an XPath relative to an identifiable element. Then, hyperlink information from the appropriate child elements (typically, "li" elements within the isolated area that have an "a" child element of their own) may be extracted. Those URLs can be stored in a Python list object as they are typically already ordered by descending popularity. For example, in the case of the *New York Times*, one could use BeautifulSoup's `find()` function to identify a "div" element with the class combination of "tab-content most-viewed", and chain the `find_all()` function to identify all "li" children, which would be iterated through using the `find()` function to identify the "a" elements that have an "href" attribute and store those items in a Python list object. Hyperlinks, as opposed to the link text, are preferable because they are unique identifiers and remain static, even as headlines and other content change.

This strategy should prove robust against the dynamic nature of homepages, as the region containing the list of most-viewed items typically maintains a uniform code pattern, remaining identifiable even as other parts of the layout change in response to breaking news and special features. After a given snapshot is parsed, information from the Python list object containing the

URL for each item and its position on the list at the given time may be stored in a relational database using either the PySQL or SQLAlchemy libraries for easy filtering and retrieval.[ii] It may also prove useful to the researcher to use BeautifulSoup to identify all "a" elements with an "href" attribute that appear outside the region containing the most-viewed items to capture when an item first appeared on a page.

After parsing the collected snapshots, the researcher must analyze the data to assess what phenomena such lists likely represent.[iii] Because this is still a nascent area of study, there are no clear standards for how lists of most-viewed items should be compared. However, there are two key dimensions that are illuminating: the rate of change for a given list of most-viewed items and the length of time it takes a news item to appear on that list. These dimensions should be evaluated complementarily in order to determine the data represented by each list. Specifically, a high rate of change and a short median time would be indicative of a list that is continuously updated and reflects recent popularity (i.e. past hour). Conversely, a low rate of change and a long median time would be indicative of popularity over a longer time period (i.e. past day).

To evaluate the length of time it takes an item to appear on that list, the researcher may query the relational database and compare the time stamp of an item's first appearance on the list of most-viewed items against its first appearance elsewhere on the page. To evaluate the rate of change, a value may be calculated to reflect the proportion of items that appear on a given list at Time (t) that change by a subsequent interval, Time (t+1). Change can be effected both through the introduction of new items to the list as well as through changes in the rankings of existing items. Formally, this calculation may be expressed as $(\frac{M_1+M_2}{2} - I) / \frac{M_1+M_2}{2}$, where *I* refers to the intersection of the lists, or the number of items (including their positions within the list) that did *not* change, and M_1 and M_2 refer to the number of items on each list. For example, if a list

contained five items and two of those items changed from Time (t) to Time (t+1)—either two new items made it to the list at the expense of two other items, or two items swapped rankings between Time (t) and Time (t+1)—then the rate of change would be 0.4, or 40%.

To access a set of computer and analysis scripts that may be used to put this approach into action, see https://www.rodrigozamith.com/.

Comparing Lists Across News Organizations

In order to illustrate the aforementioned approach and make an empirical contribution to the literature on the noteworthiness of news, an analysis of the most-viewed lists of 21 of the 50 largest print news organization in the United States, based on their weekday print circulation (see Table 1), was performed. These organizations were selected because they are often studied by mass communication scholars, were part of a broader research project (see Zamith, 2016b), and had publicly accessible lists of their most popular news items. They represent a near-census of large news organizations with connections to a print product for which information from a public list of most-viewed can be obtained. The analysis focused on two research questions centered on the comparability of the lists of most-viewed items:

RQ1: Is the average rate of change for the lists of most-viewed items similar across news organizations?

RQ2: Is the amount of time it takes a news item to appear on the list of most-viewed items similar across news organizations?

Data collection, which began on October 18, 2014 and lasted until December 20, 2014, employed the aforementioned approach to systematically download and extract information from

the lists of most-viewed items for each of the 21 news organizations every 15 minutes.[iv] Because different news organizations had lists of most-viewed items of varying lengths, only the top five items from each list were considered in order to ensure consistency in the comparison and to be able to evaluate a sufficiently large number of organizations. To ensure accuracy, error-logging mechanisms were employed by the researcher as part of an iterative algorithm development process to call attention to instances where the algorithm failed to code an item, and an electronic interface that would automatically place a given snapshot alongside the respective algorithmic coding decisions was subsequently used to manually verify the final algorithms' coding decisions for 1,050 of the snapshots captured. These data were then entered into a MySQL database. All times noted in this report reflect an adjustment to the organization's native time zone and account for changes in Daylight Saving Time.

With regard to the first research question, there was a considerable amount of variation in the rates of change for the different news organizations when utilizing a one-hour interval. As shown in Figure 1, *The Denver Post* (68.7%), the *Plain-Dealer* (65.5%), and the *Oregonian* (63.3%) had the highest rates of change. For those organizations, nearly three-fifths of the news items were, on average, either added or removed from the list, or had their positions change within it, from one hour to the next. The *Kansas City Star* (11.4%), the *Miami Herald* (11.4%), and the *Seattle Times* (12.8%) had the lowest rates of change. For those organizations, there was less than a single-item change from hour to hour on average. Additionally, some organizations, like the *Miami Herald*, the *Kansas City Star*, and the *Register* had a sudden peak followed by low or declining rates of change, suggesting that the system reset at a preset period (e.g., 2 a.m. for the *Register*), and that page views accrued from that point in time. Most organizations,

however, have patterns of change that indicate that they cover a rolling period of time (e.g., change since the previous hour or previous 24 hours).[v]

With regard to the second research question, there were also notable differences in the median amount of time it took the average news item to appear on the list of most-viewed items for the different news organizations. As shown in Figure 2, for some organizations, like the *Oregonian*, the *Plain-Dealer*, and *The Star-Ledger*, it took, on average, less than an hour for an item to appear on the list of most-viewed items (for those items that appeared on the list of most-viewed items). In contrast, it took, on average, 19 hours for an item to appear on the *Miami Herald*'s list of most-viewed items, and 16.5 and 16 hours to appear on the lists of the *Seattle Times* and the *New York Times*, respectively.

Because a considerable amount of news organizations' traffic comes from social sharing (i.e., Facebook or e-mail) or through links from aggregators and blogs, it is unsurprising that it can take items longer than an hour to appear on a list covering traffic over the past hour. That is, although an item may appear on a website at 9 a.m., it may take that item multiple hours to gain sufficient traction on social networks and other media to displace existing popular items. For example, the *St. Paul Pioneer Press* is among the few organizations that explicitly states that its list covers the past hour, yet the median it takes a news item on its site to appear on its list of most-viewed items was just over 3 hours. Nevertheless, organizations that have high median times, like the *Miami Herald*, are highly unlikely to have lists covering activity over the previous hour. The *New York Times*, for example, explicitly notes on its website that its list covers the previous 24 hours, which is consistent with its high median time.

Discussion and Conclusion

The present work aimed to offer an approach for computationally evaluating the lists of most-viewed items on different websites, and to empirically assess the comparability of those data across a number of large news organizations. In short, it was found that an accessible set of flexible, open-source software running on consumer-grade hardware can be used for such an analysis, and that data obtained from such lists are not always comparable across organizations when it comes to two dimensions: the rate at which the lists of most-viewed items change and the median time it takes a news item to reach that list. Therefore, the central and overarching conclusion from this study is that scholars should not assume that such data are comparable, and should instead ensure such comparability through empirical analysis.

These findings should not automatically cast doubt on prior work that made use of lists of most-viewed items. For example, the finding from the work of Boczkowski and Peer (2011) that there is a gap in the preferences of journalists and news consumers when it comes to the subject matter and format of stories is unlikely to be substantially affected by the fact that the data for news consumers may have covered the previous day for one organization and the previous hour for another. Indeed, provided there are a sufficient number of data points to mitigate the effect of specific events (e.g., that data covers a terrorist attack in one case but only the follow-up reporting in the other), the finding should hold up. However, the findings of studies like that of Lee et al. (2014) that utilize strict parameters (e.g., assessing relationships over short periods of time) *could* be impacted if data for one organization represents page views over the past hour while that of another organization represents page views over the past day.[vi] More importantly, such designs are likely to become more common as immediacy becomes increasingly important to digital newswork (Karlsson & Holt, 2016; Usher, 2017), news products become more liquid (Karlsson & Sjøvaag, 2016; Saltzis, 2012), and scholars adopt computational methods that can

capture shorter time lags (Widholm, 2016; Zamith & Lewis, 2015). The findings from this study therefore primarily point to the challenges of utilizing lists of most-viewed items and the need to evaluate them in comparative work to ensure that they are comparable along at least some empirical dimensions. Such evaluations should be included in the methodological details of a research report—something currently found in little of the literature on digital journalism that makes use of data from those lists.

As guidance to future researchers, the 21 organizations analyzed in the previous section were grouped into four clusters based on where they aligned across the two dimensions proposed in this study. The organizations in these clusters, shown in Figure 3, should be comparable with other organizations within their cluster, based on the rates of change of their lists of most-viewed items and the median time it takes an article to appear in it. There are, of course, no natural cutoffs for those two measures. For the purposes of this illustration, an average rate of change between 6 a.m. and 10 p.m. (when one may reasonably expect most news consumers to access content) that exceeded 50%—that is, that at least half the items on the list changed in some manner from one hour to the next—was deemed to be high. If it took the average news item longer than 360 minutes (6 hours) to appear on the list of most-viewed items, then that list was considered to have a high median time. These thresholds were also developed while being mindful of the explicit information offered by the few news organizations that commented on the data, such as the *New York Times* and the *St. Paul Pioneer Press*.

Based on this classification procedure, and as shown in Figure 3, the lists of most-viewed items for the *Oregonian*, the *Plain-Dealer*, the *Salt Lake Tribune*, the *San Jose Mercury News*, the *St. Paul Pioneer Press*, the *Star Tribune*, *The Denver Post*, and *The Star-Ledger* comprise one cluster. This cluster represents lists of most-viewed items that are most likely to reflect what

is currently popular on the website based on their high rate of change and a low median time. The *Fort Worth Star-Telegram*, the *Milwaukee Journal-Sentinel*, the *Daily News*, the *Register*, and the *Washington Post* comprise a second cluster, and the *Wall Street Journal* a third.[vii] These two clusters have lists of most-viewed items that may or may not reflect what is currently popular on their websites. Finally, the *Honolulu Star-Advertiser*, *Houston Chronicle*, *Kansas City Star*, *Miami Herald*, *New York Times*, and the *Seattle Times* comprise a fourth cluster. This cluster is unlikely to reflect what is currently popular on their websites based on their low rate of change and a high median time. It must be noted that these systems are not static, and that the data reflected by them in the future may be different than the data reflected by them at the time of this study. Thus, scholars are encouraged to perform their own analyses at the time of their study, using the approach described earlier in this report.

As scholars have long observed, the adoption and use of a technology is not dependent solely on its technical features (Kraut et al., 1998; Markus, 1994; Orlikowski, 2000). For example, Tandoc (2014, p. 568) notes that "the homepage is the prime space for a news site" and some editors believe it is important to keep it looking "fresh." The data powering lists of most-viewed items can be a helpful resource for ensuring the homepage remains fresh with interesting content—that it remains liquid—but the implementation of the list itself on the homepage (and elsewhere) may be subjected to contrasting ideas. The present findings underline the notion that the most-viewed list isn't just a technical tool with uniform purpose, but rather a configurable one that communicates information about the organization's brand and mission. For example, a rapidly-changing list covering popularity over the past hour may highlight immediacy, while a relatively static list covering popularity over the past day may highlight curation. That such lists may be used so differently across organizations, practitioners, and audiences underscores the

importance of critically examining them and the data they represent, despite their aesthetic similarities across contexts.

While a two-month timeframe was selected for this study, scholars are likely to be able to make similar assessments over a shorter period of time. Longer periods are preferred to mitigate the impact of unusual events, such as a major terror attack that results in a sudden spike of concentrated coverage or a major, anticipated event like Election Day that significantly alters the allocation of organizational resources, users' expectations, and the manner in which information is presented. However, based on the author's experience, a 'typical' two-week period should be sufficient to permit comparisons as it would sufficiently account for the temporal rhythms of different organizations and mitigate the impact of small, isolated blips. Within any timeframe, the present analysis indicates that scholars should gather information on at least an hourly basis from the respective list of each organization they intend to study.

Scholars should also remain cognizant of the potential sampling biases that may arise when relying on lists of most-viewed items. Though not related as a finding in this report due to the omission of an intercoder reliability assessment at the time of the study, the author observed several systematic omissions. That is, an entire set of news organizations belonging to a single parent company may not report data pertaining to popularity via a publicly accessible mechanism like a list of most-viewed items. Studies that draw from these lists therefore should acknowledge their inability, where appropriate, to serve as representative samples.

Future work may build upon this study by considering an even broader set of news organizations. This may include other media (e.g., broadcast and digital-native news organizations) as well as smaller news organizations, including community newspapers. Additionally, scholars should consider other measures that may be used to empirically assess the

phenomena captured by lists of most-viewed items and their comparability. While the present work has offered both a starting point and guidance for researchers in the area of digital journalism, there are surely other worthwhile measures to consider. Lastly, while certain insights from this study may be presumed to apply to other lists (e.g., lists of most-discussed items may also update on distinct schedules), additional empirical work is necessary to confirm those expectations.

In conclusion, while ratio-level data on page views is generally preferable, ordinal-level data obtained from lists of most-viewed items can be a useful alternative. However, when using such data, researchers must recognize their limitations and be transparent about them, from the sampling biases they may introduce to the information that is lost when working with relative values. Moreover, researchers should avoid assuming that these lists are comparable across organizations just because they look similar and instead attempt to assess their comparability across some empirical dimensions, with the approach described in this report serving as a guide. Ultimately, this study serves as a reminder of the need to view data and data sources with a critical eye.

References

Adornato, A. C. (2016). Forces at the gate: Social media's influence on editorial and production decisions in local television newsrooms. *Electronic News*, *10*(2), 87–104. https://doi.org/10.1177/1931243116647768

Anderson, C. W. (2011). Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism*, *12*(5), 550–566. https://doi.org/10.1177/1464884911402451

Andrejevic, M. (2007). *iSpy: Surveillance and power in the interactive era*. Lawrence, Kan.: University Press of Kansas.

Boczkowski, P. J., & Mitchelstein, E. (2012). How users take advantage of different forms of interactivity on online news sites: Clicking, e-Mailing, and commenting. *Human Communication Research*, *38*(1), 1–22. https://doi.org/10.1111/j.1468-2958.2011.01418.x

Boczkowski, P. J., & Mitchelstein, E. (2013). *The news gap: When the information preferences of the media and the public diverge*. Cambridge, Massachusetts: MIT Press.

Boczkowski, P. J., Mitchelstein, E., & Walter, M. (2011). Convergence across divergence: Understanding the gap in the online news choices of journalists and consumers in Western Europe and Latin America. *Communication Research*, *38*(3), 376–396. https://doi.org/10.1177/0093650210384989

Boczkowski, P. J., & Peer, L. (2011). The choice gap: The divergent online news preferences of journalists and consumers. *Journal of Communication*, *61*(5), 857–876. https://doi.org/10.1111/j.1460-2466.2011.01582.x

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. https://doi.org/10.1177/0894439313506845

Cherubini, F., & Nielsen, R. K. (2016). *Editorial analytics: How news media are developing and using audience data and metrics*. Oxford: Reuters Institute for the Study of Journalism.

Graves, L., & Kelly, J. (2010). *Confusion online: Faulty metrics and the future of digital journalism*. New York: Tow Center for Digital Journalism.

Groves, J., & Brown, C. L. (2011). Stopping the presses: A longitudinal case study of the Christian Science Monitor's transition from print daily to web always. *#ISOJ*, *1*(2), 95–134.

Hanusch, F. (2016). Web analytics and the functional differentiation of journalism cultures: individual, organizational and platform-specific influences on newswork. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2016.1241294

Hermida, A., Lewis, S. C., & Zamith, R. (2014). Sourcing the Arab Spring: A case study of Andy Carvin's sources on Twitter during the Tunisian and Egyptian revolutions. *Journal of Computer-Mediated Communication*, *19*(3), 479–499. https://doi.org/10.1111/jcc4.12074

Holton, A. E., Lewis, S. C., & Coddington, M. (2016). Interacting with audiences. *Journalism Studies*, *17*(7), 849–859. https://doi.org/10.1080/1461670X.2016.1165139

Karlsson, M. (2012a). Charting the liquidity of online news: Moving towards a method for content analysis of online news. *International Communication Gazette*, *74*(4), 385–402. https://doi.org/10.1177/1748048512439823

Karlsson, M. (2012b). The online news cycle and the continuous alteration of crisis frames: A Swedish case study on how the immediacy of online news affected the framing of the swine flu epidemic. *Journal of Organisational Transformation & Social Change*, *9*(3), 247–259. https://doi.org/10.1386/jots.9.3.247_1

Karlsson, M., & Holt, K. (2016). Journalism on the web. In J. F. Nussbaum (Ed.), *Oxford Research Encyclopedia of Communication*. Oxford: Oxford University Press.

Karlsson, M., & Sjøvaag, H. (2016). Introduction. *Digital Journalism*, *4*(1), 1–7. https://doi.org/10.1080/21670811.2015.1096595

Karlsson, M., & Strömbäck, J. (2010). Freezing the flow of online news: Exploring approaches to the study of the liquidity of online news. *Journalism Studies*, *11*(1), 2–19. https://doi.org/10.1080/14616700903119784

Kaushik, A. (2009). *Web analytics 2.0: The art of online accountability and science of customer centricity*. Indianapolis, Ind.: Wiley.

Kormelink, T. G., & Meijer, I. C. (2017). What clicks actually mean: Exploring digital news user practices. *Journalism*. Advance online publication. https://doi.org/10.1177/1464884916688290

Kraut, R. E., Rice, R. E., Cool, C., & Fish, R. S. (1998). Varieties of social influence: The role of utility and norms in the success of a new communication medium. *Organization Science*, *9*(4), 437–453.

Lee, A. M., & Chyi, H. I. (2014). When newsworthy is not noteworthy. *Journalism Studies*, *15*(6), 807-820. https://doi.org/10.1080/1461670X.2013.841369

Lee, A. M., Lewis, S. C., & Powers, M. (2014). Audience clicks and news placement: A study of time-lagged influence in online journalism. *Communication Research*, *41*(4), 505–530. https://doi.org/10.1177/0093650212467031

Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34–52. https://doi.org/10.1080/08838151.2012.761702

MacGregor, P. (2007). Tracking the online audience. *Journalism Studies*, *8*(2), 280–298. https://doi.org/10.1080/14616700601148879

Markus, M. L. (1994). Electronic mail as the medium of managerial choice. *Organization Science*, *5*(4), 502–527.

Martin, F. (2015). Getting my two cents worth in: Access, interaction, participation and social inclusion in online news commenting. *#ISOJ*, *5*(1), 80–105.

Napoli, P. M. (2011). *Audience evolution: New technologies and the transformation of media audiences*. New York: Columbia University Press.

Orlikowski, W. J. (2000). Using technology and constituting structures: A practice lens for studying technology in organizations. *Organization Science*, *11*(4), 404–428. https://doi.org/10.1287/orsc.11.4.404.14600

Saltzis, K. (2012). Breaking news online. *Journalism Practice*, *6*(5–6), 702–710. https://doi.org/10.1080/17512786.2012.667274

Schaudt, S., & Carpenter, S. (2009). The news that's fit to click: An analysis of online news values and preferences present in the most-viewed stories on azcentral.com. *Southwestern Mass Communication Journal*, *24*(2), 17–26.

Sjøvaag, H., Moe, H., & Stavelin, E. (2012). Public service news on the web: A large-scale content analysis of the Norwegian Broadcasting Corporation's online news. *Journalism Studies*, *13*(1), 90–106. https://doi.org/10.1080/1461670X.2011.578940

Sjøvaag, H., & Stavelin, E. (2012). Web media and the quantitative content analysis: Methodological challenges in measuring online news content. *Convergence: The International Journal of Research into New Media Technologies*, *18*(2), 215–229. https://doi.org/10.1177/1354856511429641

Sjøvaag, H., Stavelin, E., & Moe, H. (2016). Continuity and change in public service news Online. *Journalism Studies*, *17*(8), 952–970. https://doi.org/10.1080/1461670X.2015.1022204

Tandoc, E. C. (2014). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, *16*(4), 559–575. https://doi.org/10.1177/1461444814530541

Tenenboim, O., & Cohen, A. A. (2015). What prompts users to click and comment: A longitudinal study of online news. *Journalism*, *16*(2), 198-217. https://doi.org/10.1177/1464884913513996

Turow, J. (2005). Audience construction and culture production: Marketing surveillance in the digital age. *The ANNALS of the American Academy of Political and Social Science*, *597*(1), 103–121. https://doi.org/10.1177/0002716204270469

Usher, N. (2012). Going web-first at the Christian Science Monitor: A three-part study of change. *International Journal of Communication*, *6*, 1898–1917.

Usher, N. (2013). Al Jazeera English Online. *Digital Journalism*, *1*(3), 335–351. https://doi.org/10.1080/21670811.2013.801690

Usher, N. (2017). Breaking news production processes in US metropolitan newspapers: Immediacy and journalistic authority. *Journalism*. Advance online publication. https://doi.org/10.1177/1464884916689151

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., Ruigrok, N., & Schaper, J. (2016). News selection criteria in the digital age: Professional norms versus online audience metrics. *Journalism*, *17*(8), 1037–1053. https://doi.org/10.1177/1464884915595474

Widholm, A. (2016). Tracing online news in motion. *Digital Journalism*, *4*(1), 24–40. https://doi.org/10.1080/21670811.2015.1096611

Zamith, R. (2016a). Capturing and analyzing liquid content. *Journalism Studies*. Advance online publication. https://doi.org/10.1080/1461670X.2016.1146083

Zamith, R. (2016b). On Metrics-Driven Homepages: Assessing the relationship between popularity and prominence. *Journalism Studies*. Advance online publication. http://dx.doi.org/10.1080/1461670X.2016.1262215

Zamith, R., & Lewis, S. C. (2015). Content analysis and the algorithmic coder: What computational social science means for traditional modes of media analysis. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 307–318. https://doi.org/10.1177/0002716215570576

Table 1

List of the 50 Largest U.S. Newspaper Organizations

Name	Location	Parent Company	Circulation
Arizona Republic	Phoenix, AZ	Gannett Company Inc.	290,653
Arkansas Democrat Gazette *	**Little Rock, AR**	**WEHCO Media Inc.**	**161,047**
Atlanta Journal-Constitution	Atlanta, GA	Cox Media Group	198,568
Boston Globe	Boston, MA	Boston Globe Media Partners	238,108
Buffalo News *	**Buffalo, NY**	**The Buffalo News**	**160,674**
Chicago Sun-Times	Chicago, IL	Wrapports, LLC	451,864
Chicago Tribune	Chicago, IL	Tribune Publishing Company	413,475
Cincinnati Enquirer	Cincinnati, OH	Gannett Company Inc.	130,968
Courier-Journal	Louisville, KY	Gannett Company Inc.	139,225
Daily News	**New York, NY**	**New York Daily News**	**501,130**
Dallas Morning News	Dallas, TX	A.H. Belo Corporation	409,696
Detroit News/Free Press	Detroit, MI	Gannett/MediaNews	331,005
El Vocero de Puerto Rico	San Juan, PR	El Vocero de Puerto Rico	216,723
Fort Worth Star-Telegram	**Fort Worth, TX**	**McClatchy Company**	**186,625**
Hartford Courant	Hartford, CT	Tribune Publishing Company	129,903
Honolulu Star-Advertiser	**Honolulu, HI**	**Oahu Publications, Inc.**	**200,682**
Houston Chronicle	**Houston, TX**	**Hearst Newspapers**	**332,954**
Indianapolis Star	Indianapolis, IN	Gannett Company Inc.	159,037
Kansas City Star	**Kansas City, MO**	**McClatchy Company**	**186,350**
Las Vegas Review Journal *	**Las Vegas, NV**	**Stephens Media Group**	**252,110**
Los Angeles Times	Los Angeles, CA	Tribune Publishing Company	647,723
Miami Herald	**Miami, FL**	**McClatchy Company**	**191,426**
Milwaukee Journal Sentinel	**Milwaukee, WI**	**Journal Communications, Inc.**	**202,573**
New York Post	New York, NY	News Corporation	547,508
New York Times	**New York, NY**	**New York Times Company**	**1,852,698**
Newsday	Long Isla., NY	Newsday Holdings LLC	427,721
Oregonian	**Portland, OR**	**Oregonian Publishing Co.**	**226,566**
Orlando Sentinel	Orlando, FL	Tribune Publishing Company	161,837
Philadelphia Inquirer	Philadelphia, PA	Philadelphia Media Network	301,639
Pittsburgh Post-Gazette *	**Pittsburgh, PA**	**Block Communications, Inc.**	**177,411**
Plain Dealer	**Cleveland, OH**	**Plain Dealer Publishing Co.**	**292,302**
Register	**Santa Ana, CA**	**Freedom Communications**	**320,628**
Sacramento Bee *	**Sacramento, CA**	**McClatchy Company**	**195,030**
Salt Lake Tribune	**Salt Lake City, UT**	**Newspaper Agency Corp.**	**237,493**
San Francisco Chronicle	San Franc., CA	Hearst Newspapers	223,225
San Jose Mercury News	**San Jose, CA**	**MediaNews Group, Inc.**	**232,272**
Seattle Times	**Seattle, WA**	**Seattle Times Company**	**259,138**
South Florida Sun-Sentinel	Fort Laud., FL	Tribune Publishing Company	161,933
St. Louis Post-Dispatch	**St. Louis, MO**	**Lee Enterprises, Incorporated**	**169,352**
St. Paul Pioneer Press	**St. Paul, MN**	**MediaNews Group, Inc.**	**236,279**
Star Tribune	**Minneapolis, MN**	**Star Tribune Media**	**303,929**
Sun	Baltimore, MD	Tribune Publishing Company	171,614
Tampa Bay Times	St. Petersburg, FL	Times Publishing Company	246,240
The Denver Post	**Denver, CO**	**MediaNews Group, Inc.**	**414,673**
The Star-Ledger	**Newark, NJ**	**Advance Publications, Inc.**	**305,903**
Tribune Review	Pittsburgh, PA	Trib Total Media	200,502
U-T San Diego *	**San Diego, CA**	**San Diego Union-Tribune**	**225,189**
USA Today	Washing., DC	Gannett Company Inc.	1,739,338
Wall Street Journal	**New York, NY**	**Dow Jones/News Corp.**	**2,320,915**
Washington Post	**Washing., DC**	**Nash Holdings, LLC**	**454,938**

Note: All names, locations, ownership, and circulation figures according to the *Alliance for Audited Media* on Sept. 26, 2014. Bolded organizations included a list of most-viewed items at the time of the study. Organizations with an asterisk had a list of most-viewed items but were not analyzed because data were unavailable to the author at the time of study.

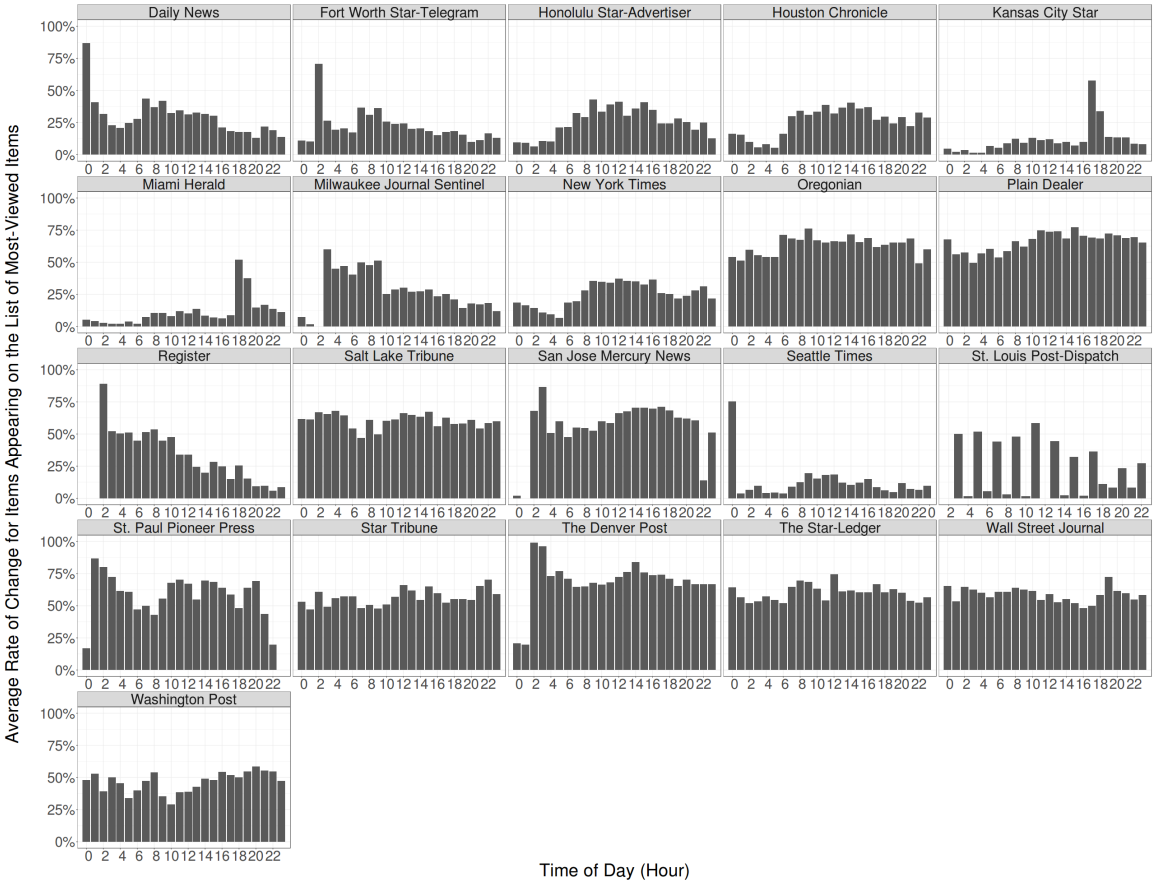

Figure 1. The average rate of change for the items appearing in the top five spots of the lists of most-viewed items of 21 news organizations over the course of 61 days.

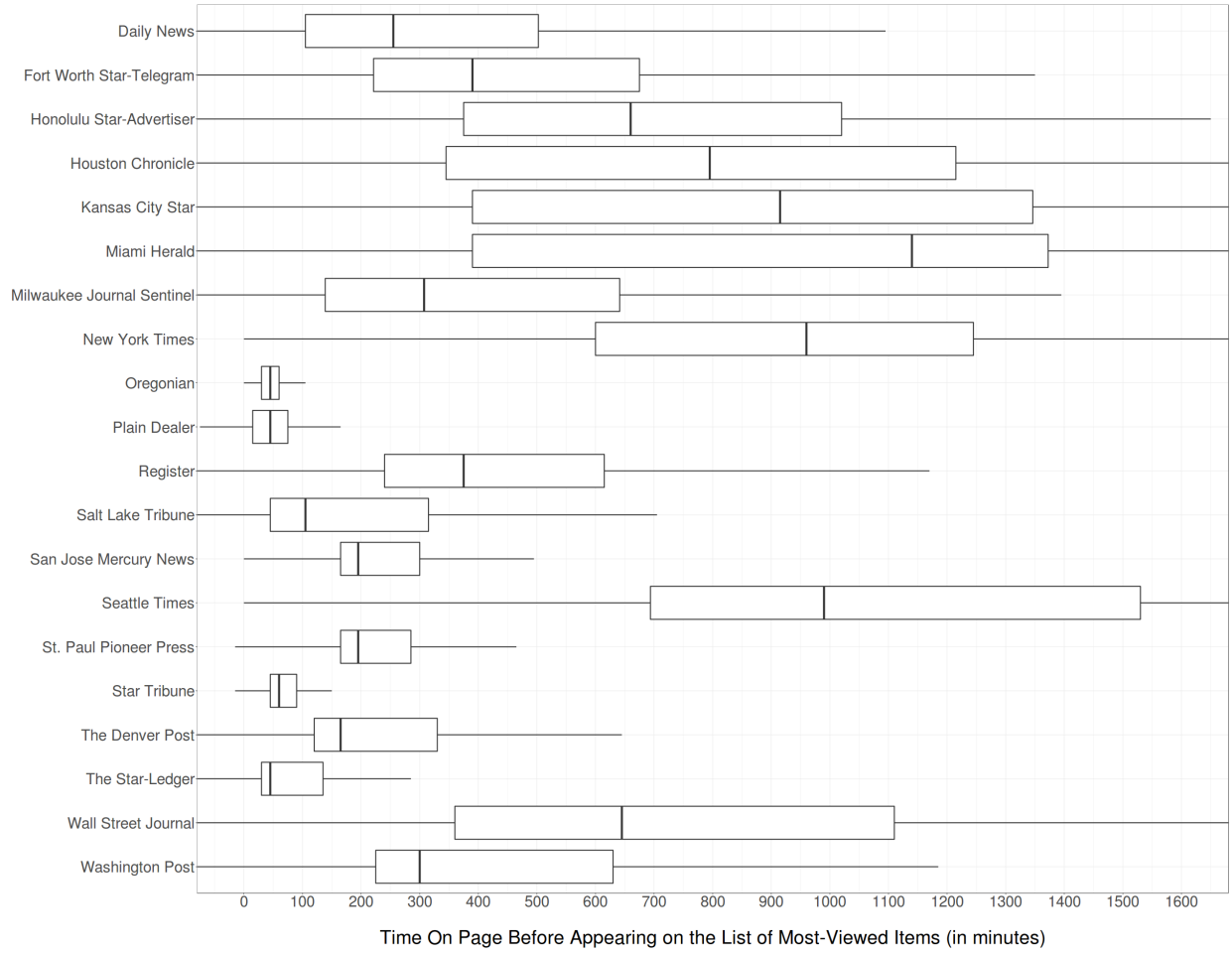

Figure 2. The amount of time it took news items to appear on the list of most-viewed items, from the time they appeared elsewhere on the homepage, for 20 news organizations over 61 days. The lines represent the range, with the left part of the box representing the lower quartile, the vertical line within the box the median, and the right part of the box the upper quartile.

	High Rate of Change	Low Rate of Change
High Median Time	• *Wall Street Journal* Quadrant II	• *Honolulu Star-Advertiser* • *Houston Chronicle* • *Kansas City Star* • *Miami Herald* • *New York Times* • *Seattle Times* Quadrant I
Low Median Time	• *Plain Dealer* • *Oregonian* • *Salt Lake Tribune* • *San Jose Mercury News* • *St. Paul Pioneer Press* • *Star Tribune* • *The Denver Post* • *The Star-Ledger* Quadrant III	• *Fort Worth Star-Telegram* • *Milwaukee Journal-Sentinel* • *Daily News* • *Register* • *Washington Post* Quadrant IV

Figure 3. The aggregation of the lists of most-viewed items from 20 news organizations into comparable clusters. Organizations in Quadrant III have lists that are good proxies of what is currently popular on the homepage. Organizations in Quadrant I have lists that are poor proxies of what is currently popular on the homepage. Items in the Quadrants II and IV have lists that are of an intermediate quality.

[i] For example, as Martin (2015, p. 92) observed in their analysis of the most popular news websites in four countries, "a key trend noted was the widespread use of integrated third-party platforms for managing commenting," such as Livefyre, Disqus, and the Facebook comments plugin. Similarly, news organizations rely on tools like Omniture, Google Analytics, Chartbeat, and Parse.ly to track website performance and user preferences (Cherubini & Nielsen, 2016; Zamith, 2016b). Such tools typically require the execution of JavaScript to permit user interaction (e.g., commenting or viewing the list of most-viewed items) with those third-party technologies.

[ii] It is recommended that the information be stored in a relational database, or a digital container with a relational model of data storage wherein data are stored in rows and columns, with a unique key typically used to identify each row. MySQL was used for this project since it offers stability, the ability to handle multiple transactions simultaneously, and advanced filtering mechanisms that make it easy to retrieve subsets of the data. MySQL, and its open-source sibling MariaDB, are free, work across multiple operating systems, and have been extensively tested.

[iii] While it might be sensible to solicit information about exactly what data are represented by lists of most-viewed items directly from a news organization, this often yields, in the author's experience, conflicting information depending on who is contacted within a given organization (see also Graves & Kelly, 2010). Thus, empirical evaluation is strongly recommended.

[iv] Because the U.S. mid-term elections—an exceptional and planned event that led to a focus on constantly-updated voting results and voter guides—occurred during this time period, data collected on Nov. 3rd, 4th, and 5th were discarded.

[v] All but one of the organizations updated their list of most-viewed items at least once an hour on average. The lone organization that did not do this was the *St. Louis Post-Dispatch*, which

generally updated its list every other hour. Additionally, there were a few points in time where there was no activity for some of the organizations (e.g., the *Register*), typically occurring during overnight hours. Given their consistent recurrence, and based on the researcher's observations while developing the computer scripts, this is likely because the requisite systems (e.g., server log information) were unavailable during those hours due to regular server maintenance or as aggregate reports were compiled.

[vi] To be clear, the author is not indicting that work as the report does not provide sufficient information to assess the potential for incomparability among lists. Instead, the reference points to a particular research design that may be used to study such phenomena.

[vii] The *Wall Street Journal*'s unique position is likely due to the fact that it combines page view data with social media data in its calculation of its most popular items. A high rate of change coupled with a high median time would indeed suggest a responsive algorithm affected by the time it takes popular items to gain traction on and diffuse through social media. Such complex algorithms are currently rare, however, as most outlets either use a single metric (e.g., most viewed) or split metrics into separate lists (e.g., most viewed and most shared).