Editorial Judgment in an Age of Data: How Audience Analytics and Metrics are Influencing the Placement of News Products

A DISSERTATION
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Rodrigo Zamith

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Seth C. Lewis, Co-Adviser
Marco Yzer, Co-Adviser

May 2015

Acknowledgements

The prominence of my name on the title page is misleading. This dissertation is not just my work. It is the product of a wonderful environment and an incredible cast of advisers, family, and friends. (These categories are far from mutually exclusive!) I should begin by acknowledging the contributions of my mother, father, brother, and sister. I would not have made it this far without their unceasing enthusiasm and optimism. A wonderful cast of friends similarly kept me sane and challenged my ideas. Anna, Brett and Kathleen, Konstantin and Emily, Liz and Brian, Meagan, and Sarah were far from the only colleagues who shared a pint with me one day and helped me interpret a finding the next, but they deserve special mention for their immense charity and good humor throughout my program. My advising team was nothing short of terrific. Dr. Fan forced me to think carefully about how to lay the foundation for my analyses well in advance of conducting them. Dr. Lewis went to tremendous lengths to help me grow as a scholar over the past four years, and I will consider myself a success if I ever come *close* to the high bar he sets for himself as a person and as a scholar. Dr. Watson kept me on my toes throughout my program, making sure my ego never got too inflated or deflated, all while pushing me to work harder, sharing far more of his time than a student should expect in the best of circumstances, and introducing me to a few new brews. Dr. Yzer alternated being a sounding board, a teammate, and the sort of deep thinker who gives you confidence in everything you do. In short, if any of these individuals were to be removed from this process, this dissertation would have turned out quite differently—if at all. Thank you.

Dedication

To my mother and father, who have offered me consistent and unconditional support and affection.

Abstract

In recent years, audience analytics (systems that collect and analyze digital trace data from users) and audience metrics (quantified measures of how content is consumed and interacted with) have gained currency within newsrooms, enabling them to influence editorial newsworkers' constructions of their audiences and, consequently, the shapes that news products take. The extent of that influence, however, remains largely unknown, with few studies examining the direct relationship between metrics and news content. The present work addresses this shortcoming by offering methodological guidance and an empirical assessment of the extent to which one particularly salient metric, page views, influences the prominence and de-selection of news items on the homepages of several news organizations. It demonstrates that algorithms can be leveraged to computationally analyze particular aesthetics of a large volume of homepages; that the 'most viewed' list can serve as a useful indicator of the popularity of news items, though such lists are not always comparable across organizations and introduce important limitations; and that the effect of an item's popularity on its subsequent placement on the homepage is fairly muted in the process of selection, though it is greater in the process of de-selection. In short, the present research indicates that the current content-related effects of audience metrics—at least as it pertains to a particular editorial function and metric—may be overstated in the literature, and offers pathways for further studying similar relationships.

Table of Contents

List of Tables

List of Figures

CHAPTER I:
INTRODUCTION

According to multiple theories of the role of the press in liberal democratic societies, the press plays a crucial role in enabling and facilitating democratic processes (Bennett, Lawrence, & Livingston, 2008; Carey, 1993; Habermas, 1989; Siebert, Peterson, & Schramm, 1956). As such, scholars have long taken an interest in the editorial process, attempting to better understand how journalists and editors select what to cover from a seemingly limitless pool of potential stories available to them and, subsequently, decide how to present that content. These decisions, in turn, exert considerable effects on how publics come to understand and prioritize issues (Breed, 1955; Shah, McLeod, Gotlieb, & Lee, 2009). The placement of news stories, for example, conveys to the consumer a strong signal of the editor's perception of the importance and news value of that news product, with more-prominent items being more likely to remain salient in that consumer's mind (McCombs & Shaw, 1972; Wanta, 1997).

Over the past two decades, newswork has changed dramatically: new technologies have been introduced (Boczkowski, 2005; Reich, 2013), audiences have been empowered (Benkler, 2006; Bruns, 2008), and traditional business models have been challenged (Fengler & Ruß-Mohl, 2008; Soloski, 2013), with the confluence of these developments yielding a very different media environment that demands changes to the process of doing journalistic work (Deuze, 2007). In particular, news organizations have developed their online presence, with many viewing digital distribution both as an opportunity and a vexing challenge (Boczkowski & Mitchelstein, 2013; Boczkowski, 2005). By their very nature, these digital platforms enable data about users and their

behaviors to be automatically collected and quickly processed (Andrejevic, 2007; Mullarkey, 2004). News organizations, following in the footsteps of other service and content providers, have come to recognize the potential of these data to gain a better understanding of the individuals who consume their products—and potentially use it to make important editorial decisions (Anderson, 2011a; Groves & Brown, 2011; Usher, 2012).

However, not only have such data been historically rejected by editorial newsworkers (Gans, 1979; Schlesinger, 1978) but they introduce a challenge to the occupational ideology and professional logic of journalism (Deuze, 2005; Lewis, 2012). The present research explores the emerging role of audience analytics through a focus on a particular practice in newswork: presenting news content. Specifically, it focuses on understanding this development and the tensions associated with it; offering methodological guidance to scholars interested in assessing the impact of metrics on content; and empirically assessing the extent to which one particularly salient metric, popularity, influences the prominence and de-selection of news items.

Constructing Audiences

A central consideration in the editorial process is the audience, or the individuals for whom a newsworker develops a news product (Shoemaker & Vos, 2009). The audience may be comprised of multiple groups of people, such as editors and sources (Gans, 1979), although one central group is the news consumer (DeWerth-Pallmeyer, 1997). While it may be said that there exists an actual audience—that is, a tangible set of

individuals with certain characteristics and preferences—that audience becomes a social construction in the mind of newsworkers during the process of crafting news products (Pool & Shulman, 1959). Put differently, in the process of envisioning the actual audience, newsworkers fashion a constructed audience based on multiple inputs (e.g., their interactions with a small portion of the audience).

Over the past century, there has been a movement toward ever-greater rationalization of audience understanding, or the use of scientific methods to construct the audience based on data (Napoli, 2011). This movement has been greatly influenced by technological changes that have altered the dynamics of media consumption and technological changes that have facilitated the gathering of new forms of information about the media audience, with major shifts occurring in the 1930s and 1970s (Napoli, 2011). In the contemporary media environment, a proliferation of networked communication technologies and low-cost publishing platforms has led to an explosion in the availability of news products and changes in news consumption patterns (Benkler, 2006; Deuze, 2001). Consequent to this has been the development of an unprecedented degree of audience fragmentation (Napoli, 2011). Additionally, these technologies, by their very nature, make it possible to capture large amounts of digital trace data and relay that data, often in real-time, back to the content producer (Andrejevic, 2007).

In light of these shifts, it may be argued that we are in the midst of a new wave toward the rationalization of audience understanding. This new wave is marked by the increasing adoption of a new audience information system, or data gathering and feedback mechanism used to measure exposure to and interaction with content (Napoli,

2011): audience analytics. Audience analytics, or the systems that enable the measurement, collection, analysis, and reporting of digital data pertaining to how content is consumed and interacted with (see Kaushik, 2009), allows content producers to harvest non-purposive feedback from news consumers. This information is then distilled into audience metrics, or quantified measures of audience preferences and behaviors (Anderson, 2011a), which offer the potential to dramatically alter the construction of the audience by offering a powerful new input. Indeed, while editorial newsworkers have traditionally relied on interactions with their immediate peers, friends, and family as well as letters to the editor as their primary inputs in constructing the audience (Gans, 1979; Schlesinger, 1978), audience metrics offer real-time, complete, and seemingly more 'objective' measures (MacGregor, 2007) that can lead to "a fundamental transformation … in journalists' understanding of their audiences" (Anderson, 2011b, p. 529; see also Couldry & Turow, 2014).

Environmental Challenges

However, the mere availability of a technology does not mandate its use. To that end, it is important to note that over the past two decades (and especially in the past decade), the news industry in the United States—and in particular newspaper companies—has had great difficulty adapting to these technological developments and sociocultural shifts in the consumption of news products (Soloski, 2013). Environmental changes like a continuous drop in print circulation, a prolonged decline in advertising revenue, and the dramatic increase in competition (for recent indicators, see Mitchell et

al., 2014) have all created a climate of uncertainty in the industry. One organizational response to such threats is to engage in the adjustment of internal processes, strategies, and goals by realigning organizational resources (Meyer, Tsui, & Hinings, 1993; Snow, Miles, & Miles, 2005). Newspaper companies have arguably done this by transitioning an ever-greater proportion of their dwindling resources to developing and supporting digital formats (Soloski, 2013; Tang, Sridhar, Thorson, & Mantrala, 2011). Furthermore, many organizations have reconfigured the very process of newswork by demanding greater integration of business units, or the tight-coupling of departments, so that knowledge—a key resource in modern economies and particularly in news enterprises—can be more readily shared (Gade, 2009a; Lowrey, 2011). Indeed, one of the key recommendations in the leaked and much-discussed *New York Times Innovation Report* was to increase inter-departmental integration, which the authors argued stifled innovation and the company's ability to compete (The New York Times Company, 2014).

The consequence of this process of integration is the erosion of the traditional 'wall' between business and editorial concerns and an even greater pressure to shift toward a market orientation wherein editorial decisions are made with greater deference to what the reader wants than what the reader needs in order to be a responsible member of a democratic society (Beam, 2001, 2003; Fengler & Ruß-Mohl, 2008).[1] In particular,

---

[1] It is important to recognize here that these are not mutually exclusive considerations. That is, what the reader wants could very well be what he or she needs in order to be informed citizens. Furthermore, these terms are used with the perspective of the newsworker in mind. Put differently, what the reader 'needs' refers to the value judgment of the news practitioner about what information is essential to his or her constructed audience.

audience metrics enable editors to close the gap between what journalists find to be newsworthy and audiences find to be noteworthy (Boczkowski, Mitchelstein, & Walter, 2011; Boczkowski & Mitchelstein, 2013; Boczkowski & Peer, 2011; Lee & Chyi, 2014; Tandoc, 2014a). In light of the shift toward a market orientation and the economic challenges faced by many news organizations, it would follow that institutional forces would promote the narrowing of this gap—that is, to move toward what Anderson (2011b, p. 529) has called the "agenda of the audience."

Newswork, Analytics, and Metrics

These developments, then, would seem to indicate that audience analytics and metrics are emerging as central objects in the newsroom that are capable of influencing editorial decision-making. However, the use of audience analytics may easily be understood to be at odds with the professional logic of journalism and its occupational ideology. Indeed, according to Lewis (2012), the notion of professional control over content is central to the professional logic of journalism. Furthermore, journalistic autonomy and the ability to ascertain newsworthiness are important components of the occupational ideology of journalism (Boczkowski, 2004; Deuze, 2005). As such, the emergence of audience analytics introduces a new source of tension to changing newsrooms, which must be resolved by individual newsworkers.

The emergent work in this area would seem to support the contention that editorial newsworkers are becoming increasingly sensitive to audience analytics. MacGregor (2007), for example, found "prolific use" (p. 286) of audience analytics, with

some journalists viewing them with "a hawk eye" (p. 289). Anderson (2011a) found that "major" (p. 559) editorial decisions were being made based on audience metrics. Simultaneous ethnographic studies by Groves and Brown (2011) and Usher (2012) painted pictures of an organization—the *Christian Science Monitor*—that became metrics-driven in its shift to the Web. Furthermore, recent surveys of editors have found extensive awareness—and in many cases use—of audience analytics and metrics (Lowrey & Woo, 2010; Mayer, 2011; Vu, 2014).

However, the rate of adoption of audience analytics has hardly been uniform, nor has its construction or use been homogenous. Anderson (2011a) found that the *Star-Ledger* made far less use of audience analytics than the Philadelphia newsrooms he visited. Similarly, Usher (2013) found that the use of audience analytics was very limited at *Al Jazeera English Online*, with several top-level managers downplaying its use. As MacGregor (2007) argues, there are multiple potential responses to audience analytics, from stimulus-response behavior characterized by immediate reactions to deliberated and mediated responses in which such data are weighted against other factors. Additionally, audience analytics may be consulted, but not yield a content-related response. For example, both Usher (2013) and Petre (2015) found that, at some news organizations like *Al Jazeera English Online* and the *New York Times*, audience analytics and metrics were used by journalists primarily to validate, rather than guide, their editorial decisions.

Gaps in the Literature and Contribution of the Dissertation

While it has generally been presumed that audience metrics have an extensive

impact on the shape content ultimately takes—from what gets covered to how it is presented—the scholarship in this area is sparse. Indeed, following her ethnographic examination of the popular analytics tool Chartbeat and the newsrooms at *Gawker* and the *New York Times*, Petre (2015, n.p.) calls particular attention to this gap and urges scholars to engage in "more studies using systematic content analysis to determine if and how metrics are influencing news content." This shortcoming is especially concerning given that much of the scholarship in this area has focused on digital-native organizations or newsrooms that are ahead of the curve in this sociocultural shift, such as the *Christian Science Monitor* and *Al Jazeera English Online* (e.g., Groves & Brown, 2011; Usher, 2012, 2013). Lee and colleagues (2014) provide a welcome contribution through their study of the impact of a story's popularity on its prominence on a homepage, but focus only on three news organizations in the state of New York over a short period of time. Bright and Nicholls (2013) also offer a useful contribution to the literature, looking at the impact of a story's popularity on the length of time it remains on a homepage, but focus on just five U.K.-based organizations. Thus, much remains unknown about the prevalence of such effects across the broader range of oft-studied large and mid-sized newspaper organizations, which are likely to be the sites of greatest tension over the use of audience analytics and metrics. Furthermore, these contributions point to several methodological challenges associated with this kind of work that have not been critically examined by scholars.

The present research seeks to address these gaps through a series of analyses that offer both methodological and theoretical contributions. In Chapter II, a comprehensive

review of the literature is conducted in order evaluate how and why audience analytics and metrics are being used in the modern media environment. In Chapter III, a framework for capturing 'liquid' Web content and computationally identifying and extracting important elements pertaining to its aesthetic, namely the items that appear in regions of prominence and on the list of most-viewed items is detailed, highlighting the utility of open-source software that could be leveraged in computational social scientific inquiry. In Chapter IV, the 'most viewed' lists of 21 news organizations are assessed across two dimensions in order to evaluate the phenomena captured by those lists and the extent to which they are comparable across organizations, demonstrating their adequacy and limitations as a general proxy for popularity. In Chapter V, a theoretical framework rooted in gatekeeping and the professional logic of journalism is drawn upon to inform a large-scale, computational analysis of the impact of an item's popularity on its subsequent prominence and de-selection across 14 news organizations. Finally, these insights are situated among broader questions in the field in Chapter VI as part of the overarching contribution of the present work to the body of literature on the emerging role of audience analytics and metrics in newswork.

CHAPTER II:
QUANTIFYING AUDIENCES

In a study of the use of audience metrics by organizations owned by Philadelphia Media Holdings, Anderson (2011a, p. 560) found that "website traffic numbers, no matter what the content of actual clicked articles, were invoked often at the *Philadelphia Inquirer* and almost obsessively at Philly.com." At the *Inquirer*, reporters had constant access to reports of 'click counts' sorted by author name, leading one journalist to lament that they were "probably headed toward a new model where reporters get paid by clicks" (p. 559). At a news meeting, one editor called attention to the fact that the website had faced two consecutive months of decline in terms of hits: "We're in a summer slump—and we aggressively need to find a way to end it. We will protect our growth in page views!" In describing the organizational identity, a web producer at Philly.com said, "we're trying to be a real strong local news site that appeals to our audience and gets traffic" (p. 562).

Anderson's (2011a) observations are hardly an outlier. There is growing evidence that audience analytics and audience metrics are becoming increasingly prevalent in newsrooms in the United States and elsewhere, leading to shifts in the way newswork is done and potentially in the shape that news products take (Bright & Nicholls, 2013; Lee et al., 2014; Tandoc, 2014b; Usher, 2012). This is a surprising development for many, given that data about the audience have traditionally been ignored, if not rejected, by many editorial newsworkers, and because they may be seen as an affront to the professional logic and occupational ideology of journalism (Deuze, 2005; Gans, 1979;

Lewis, 2012; Nguyen, 2013; Tandoc, 2014a). This raises two fundamental questions: How did audience analytics and metrics gain such currency? And exactly how are they being used in the current media environment?

In the following sections, a range of scholarly works is examined to address those questions. This examination begins by assessing how audiences become constructed in the minds of newsworkers and how key inputs in that construction have changed in recent decades. It then focuses on a particularly important development, the introduction of audience analytics—systems that automatically collect a range of data about the consumers of digital content—and how the distilled, quantified information they yield—audience metrics—can transform constructions of the audience. To aid the understanding of why these systems may be gaining traction among news organizations, theories of organizational change are drawn upon in light of the uncertain environment that has clouded the news industry in recent years. The current state of knowledge about how and the extent to which these systems and data are being used is then evaluated. Finally, the examination concludes with an analysis of what these developments may mean for newswork and newsworkers.

The Construction of the Audience

From a sociological and constructivist perspective, audiences and news products may be viewed primarily as social constructions that may or may not reflect some objective reality (Berger & Luckmann, 1966; Schudson, 2003). Such a perspective focuses, therefore, on the constructed audience—that is, the "images" (Gans, 1979),

"abstractions" (Schlesinger, 1978), and "fantasies" (Pool & Shulman, 1959) that newsworkers develop to represent their vision of the actual audience, or the tangible set of individuals who consume a given news product. It is important to distinguish between the constructed audience and the actual audience because individuals (e.g., editorial newsworkers) can only make decisions based on their perceptions of phenomena and those perceptions may vary considerably from reality. Indeed, editorial newsworkers long depended on letters to the editor and interactions with their immediate peers, friends, and family as the primary inputs for their construction of the audience, yielding constructions that were only marginally reflective of the individuals who consumed their work (DeWerth-Pallmeyer, 1997; Gans, 1979).

The constructed audience is of import because it has long been a source of influence on the development of news products. As DeWerth-Pallmeyer (1997, p. 34) notes, "although the central concern of journalists is creating the news, tacit understandings of the audience are imbedded in the news gathering process, in the news values they use, and in the technology they use." As an illustration of this phenomenon, consider Schudson's (1995) contention that there are five parties involved in a journalistic interview: the journalist, the source, the journalist's employer, the source's employer, and the absent audience. That is, when interviewing a source, journalists ask questions to which they think a constructed audience wants (or needs) answers.

The constructed audience is derived largely from newsworkers' tacit professional knowledge (DeWerth-Pallmeyer, 1997). This tacit professional knowledge may be conceptualized by what Schön (1983) calls as knowing-in-action, or the "actions,

recognitions, and judgments which we know how to carry out spontaneously; we do not have to think about them prior to or during their performance" (p. 28). The concept is derived from the work of Polanyi (1966), who argued that much of what individuals know is understood on a subconscious level and not easily articulated. These abstractions, therefore, are of great import because they inform decision-making at multiple levels, from calculations of newsworthiness values (DeWerth-Pallmeyer, 1997) to organizational strategy (Turow, 2005). As Napoli (2011) writes:

> The perceptions that producers of culture have about their audience (be they informed or uninformed, narrow or well-rounded) naturally feed into judgments as to what kinds of content will succeed and what kinds of content will fail, as well as into assessments of which audience interests are being well served and which are not. (p. 18)

The constructed audience, therefore, can influence decision-making at both a conscious and an unconscious level. As such, it is important to understand how the audience is constructed and how that process has 'evolved' (Napoli, 2011). Indeed, scholars have argued that the basis for the constructed audience has changed considerably over the past century as a result of technological developments, changing patterns of media consumption, and economic imperatives (Andrejevic, 2013; Lowrey & Gade, 2011; Napoli, 2003).

The Rationalization of Audience Understanding

According to Napoli (2011), the twentieth century was marked by an increasing

rationalization of audience understanding. By this, Napoli means that "over time, media industries' perceptions of their audience have become increasingly scientific and increasingly data-driven" (p. 11). Napoli contends that media organizations largely utilized an 'intuitive model' of audience understanding during the early twentieth century. Under this model, decisions were made based on "subjective, often instinctive judgments of content producers, distributors, and exhibitors regarding audience tastes, preferences, and reactions" (p. 32). This was largely possible because economic conditions—namely the centralization of media production and distribution, relatively limited competition, and considerable demand for mass media vehicles by advertisers—were favorable, thereby making more sophisticated and rigorous analyses unnecessary.

Napoli notes that in the 1930s, the economic hardships introduced during the Great Depression drove marketers and advertisers to demand 'tangible' evidence that their campaigns were effective and their ad dollars well spent. Consequently, the 1930s marked the first wave of progression toward the rationalization of audience understanding. For example, it was during this time that magazines began to compile and publish detailed readership reports that included demographic and behavioral characteristics of its readers. A second wave of progression toward the rationalization of audience understanding occurred in the 1970s as computers began to facilitate the collection and analysis of large quantities of statistical data and news outlets began to turn toward news consultants that could help them attract larger audiences (see also Allen, 2007; Barnes & Thomson, 1988; Buzzard, 2003). Driving this second wave was the desire to further quantify audiences so that more 'scientific' managerial decisions

could be made.[1]

The Emergence of Audience Analytics

According to Napoli (2011), there are two interrelated factors that drive the rationalization of audience understanding: (1) technological changes that alter the dynamics of media consumption; and (2) technological changes that facilitate the gathering of new forms of information about the media audience. With regard to the first process, the contemporary media environment is characterized by fragmentation and audience autonomy (Napoli, 2011). That is, there is presently a large and growing array of content delivery platforms (inter-media fragmentation), resulting in the disaggregation of content (intra-media fragmentation) and the diffusion of audience attention (audience fragmentation) (Anderson, 2006; Napoli, 2003). Furthermore, audiences now have considerably more control over when, where, and how they consume media (Benkler, 2006; Deuze, 2001) and have greater capacity to produce their own content at marginal costs (audience autonomy) (Bruns, 2008; Croteau, 2006), thereby providing them with an abundance of choices.[2]

These shifts in fragmentation and autonomy have created a significant challenge for traditional audience information systems—that is, the "data gathering and feedback

---

[1] For a thorough review of these developments, see Napoli (2011).
[2] It should be noted that while modern technologies have drastically reduced the costs of production, the most disruptive development in the modern media environment is actually the unprecedented ability of individuals to distribute their creations (or the creations of others) to mass audiences with trivial costs (Napoli, 2001).

mechanisms used by media industries and advertisers not only to measure audience exposure to media content, but also to predict content preferences and consumption patterns, target content to specialized audience segments, and gather information on audiences' reactions and behavioral responses to content" (Napoli, 2011, p. 10)—since they are unable to accurately capture such dispersed and empowered audiences (Webster, 2008). For example, traditional audience measurement, which has relied on small yet purportedly representative samples of media consumers, performs especially poorly in capturing much of the activity that occurs in the 'long tail' of the Web (Anderson, 2006). Similarly, the ability of the audience to take a multitude of paths to a given media product and often to interact with it has led to the pursuit of a more sophisticated, complete, and accurate understanding of behavior (e.g., how a news consumer moves about a website and the keywords they use to seek out media content) in addition to exposure (e.g., the pages they view) (Napoli, 2011).

Parallel to this development has been the introduction of electronic devices and services that can systematically collect information about users—their 'digital footprints' (Madden, Fox, Smith, & Vitak, 2007)—in the process of providing users with their desired content. According to Andrejevic (2007), modern technologies have created a 'digital enclosure,' or "an interactive realm wherein every action and transaction generates information about itself" (p. 2). These digital trace data, or micro-data about every action a user takes in the digital realm, may be subsequently transmitted back to content producers, providing them with new streams of information about the consumer's habits and preferences. With regard to digital media content, these technologies do away

with the traditional need to sample (for data may be inexpensively gathered about every user) and are capable of capturing all digital behavior (removing the need to rely on diaries or other self-reports).[3] Ultimately, the technological abilities and interactive nature of these devices and services makes it "possible to record data about individual consumers at an unprecedented level of detail" (Mullarkey, 2004, p. 42).

In light of these technological and sociocultural shifts in media consumption and data gathering, it may be argued that the media industry is in the midst of a third wave toward the rationalization of audience understanding. In particular, this wave is characterized by the intense development of systems to capture, link, and organize these digital trace data. In particular, there has been great focus on audience analytics, or the systems that enable the measurement, collection, analysis, and reporting of digital data pertaining to how content is consumed and interacted with (see Kaushik, 2009). The requisite data are in many cases automatically collected by servers in order to process requests from the client, although website and application designers may also introduce additional pieces of code (e.g., scripts and cookies in the context of a website) to capture additional data about user habits.[4] Audience analytics suites, such as Chartbeat and Omniture, then analyze and report, in a form that is easier for humans to interpret, counts,

---

[3] However, focus groups and surveys remain invaluable for assessing certain attitudinal constructs that behavioral data provide only limited insight into.

[4] It is possible for the client to forge a great deal of information in their request—for example, the client may claim to be using a different browser or operating system—and cookies and scripts can be selectively rejected, frustrating some data collection efforts. However, one metric that cannot be spoofed is the page view, or hit, because it is assessed on the server side. That is, for a hit to occur, a page must be requested by a client and served by the server. It should be noted, however, that a request for a page does not guarantee that it will be consumed (e.g., read) or that it was requested by a human rather than a computer program, like a search engine crawler, leading to overstatement (Graves & Kelly, 2010).

proportions, and patterns derived from these data.

The introduction of new audience information systems, and in particular audience analytics, allows for the construction of a more sophisticated understanding of the audience across certain dimensions. Indeed, all user behavior on a website or digital device can be potentially tracked and stored, from the items users click on to each pixel movement of the mouse cursor (Andrejevic, 2007; Kaushik, 2009). Thus, whereas traditional audience measurement has focused on exposure and relied on data obtained through the use of small, though supposedly representative, samples of the audience (Webster, Phalen, & Lichty, 2000), audience analytics enables low-cost, automatic gathering of relatively sophisticated data on exposure and certain behavioral responses for all consumers of a digital media product. Put differently, audience analytics can capture a wide range of non-purposive feedback—that is, information about an individual's preferences and behaviors that the individual did not intentionally provide.[5]

Audience Metrics and the Constructed Audience

The distilled and enumerated form of data yielded through audience analytics is referred to in the present research as audience metrics. That is, audience metrics refer to the quantified and aggregated measures of audience preferences and behaviors generated by the data collection and processing systems termed audience analytics (see Anderson,

---

[5] An example of purposive feedback would be the reader writing to a reporter to express his or her satisfaction with the reporter's story. In contrast, non-purposive feedback would include the fact that the reader viewed the story or shared it with colleagues on social media. The distinction is that the reader intends to convey information directly to the reporter in one instance but not the other.

2011a). For example, whereas audience analytics would refer to the algorithms that capture and display sophisticated information about the pathways taken by a user to a given endpoint (e.g., a news article), the audience metric would be the number of individuals who took a particular route (e.g., through the homepage). This distinction enables the separation of the artifactual nature of the technology and the textual nature its content (see Siles & Boczkowski, 2012). That is, it enables the researcher to be cognizant of the fact that although different analytics suites focus on many of the same metrics, they often collect, synthesize, and present that information in very different ways (Graves & Kelly, 2010). In combination, audience analytics and audience metrics offer the potential to dramatically alter editorial newsworkers' construction of the audience by introducing powerful new inputs.

Among editorial newsworkers, the constructed audience has often relied on what Napoli (2011) has described as the 'intuitive model.' For example, as Schlesinger (1978, p. 107) noted, the "audience remains an abstraction, made real on occasion by letters or telephone calls, encounters of a random kind in public places, or perhaps more structured ones such as conversations with liftmen, barmen and taxi-drivers." Indeed, for most journalists, letters to the editor and interaction with their immediate peers, friends, and family have offered the primary basis for the construction of the audience (Gans, 1979). Consequently, newsworkers have traditionally relied on 'gut feelings'—or rather, the unconscious application (knowing-in-action) of news values and editorial judgment associated with their occupational ideology and professional logic, and cultivated through education and experience—to drive their editorial decision-making. Indeed, as one editor

interviewed by DeWerth-Pallmeyer (1997) reflected while discussing the selection process for the stories that appear on the front page: "It's a gut decision. It's something that we make on an individual case every day. And it's almost thoughtless after a period of time" (p. 27). These gut feelings, in turn, are the product of a composition of news factors like timeliness, proximity, conflict, and prominence—and, as DeWerth-Pallmeyer (1997) argues, notions of the audience are a tacit part of every one of these factors.

In contrast to the gut feelings based on intuitive rationalizations of the audience, audience metrics offer a data-driven and ostensibly more comprehensive picture of the audience by virtue of its ability to capture an array of information about the actual choices and behavioral patterns of all users (Kaushik, 2009). As such, audience metrics offer the potential to make the constructed audience to become more reflective of the actual audience. Furthermore, in contrast to waiting for a letter to the editor or until a future encounter with a reader, the nature of audience metrics enables it to be continuously updated without the need for intervention, and thus provide editorial newsworkers with real-time information (Anderson, 2011a).

News Media in Uncertain Times

As scholars have long observed, the availability of a technology does not mandate its use (Pinch & Bijker, 1984). That is, just because editorial newsworkers have access to audience analytics and audience metrics does not mean that they will make use of them. Formal audience research—and audience metrics certainly qualify as that—has long been available to editorial newsworkers, from magazine reader surveys in the 1930s to the

emergence of television ratings in the 1950s to formal research by news consultants in the 1970s (Napoli, 2003). However, as researchers have long noted, editorial newsworkers have tended to be skeptical and mistrustful of formal audience research (Gans, 1979; Green, 2002; Stuart, 2010). In particular, Gans (1979) noted four reasons why journalists tend to be skeptical of formal audience research: (1) journalists usually have liberal arts educations and dislike statistics; (2) journalists are rarely shown how that information could be useful; (3) such research may cast doubt on their news judgment and affect their professional autonomy; and (4) audience research is typically conducted by non-journalists. In a series of profiles of the 'American journalist', Weaver and colleagues (2007) have found that, since the 1980s, a decreasing proportion of journalists perceive audience research to be very influential in their work, with just 29 percent of journalists at daily newspapers thinking it was influential in their most recent survey.[6]

According to a broad set of theories in the field of organizational change, organizations must adapt in the face of environmental changes—those that do not, or those that select inferior responses, often fail (Barnett, Greve, & Park, 1994; Raisch & Birkinshaw, 2008; Snow et al., 2005; Tushman & Romanelli, 1985). Understanding the contemporary media environment, therefore, is important for comprehending why news organizations may be driven to adopt and encourage the use of audience analytics and metrics.

---

[6] More recent figures from Vu (2014) and Tandoc (2014b), discussed later in this chapter, indicate that this trend has shifted.

Current Challenges Faced by the Industry

As scholars have observed, the news industry in the United States—and in particular newspapers—has been subjected to disruptive environmental changes, or shifts in the conditions surrounding an organization, over the past decade (for recent indicators, see Mitchell et al., 2014). Many of these shifts, like growing audience fragmentation and increasing audience autonomy have upended the traditional economic models driving the industry (Napoli, 2011). As Lowrey and Gade (2011) note, the news industry, and in particular newspapers, are currently faced with a climate of uncertainty. This widespread anxiety has been fostered by a number of troubling developments, including a considerable and continuous drop in print circulation, a prolonged decline in advertising revenue, and a dramatic increase in competition.[7]

Between 2003 and 2009, the weekday print circulation of newspapers declined by 17 percent (from 55.2 million to 45.7 million). In particular, there has been a trend for paid subscribers to dump the print edition and access the same content online for free (Doctor, 2010). In response to this shift, the Alliance for Audited Media, formerly the Audit Bureau of Circulations, has in recent years changed its reporting procedures to include 'digital circulation.' Recent figures indicate that digital editions account for more than 19 percent of U.S. daily newspapers' total average circulation (Alliance for Audited Media, 2013). Using these more-inclusive figures, indicators suggest that large daily

---

[7] Unless otherwise noted, the figures presented in the following paragraphs were obtained from the 2014 State of the Media report by the Pew Research Journalism Project. That report uses data from a variety of sources, such as the Newspaper Association of America and Scarborough Research, in addition to their own primary data collection. For more information, see Mitchell et al. (2014).

newspapers—the small set of 5 newspapers with a daily circulation in excess of 500,000—have seen a boon in circulation (a 22 percent increase between 2012 and 2013). However, all newspapers categories below that have continued to see contraction in circulation.

Parallel to these shifts in circulation—and likely more alarming for newspaper companies—has been an exceptional decrease in annual revenue. This is especially true with regard to advertising revenue, which has long been the lifeblood of most newspapers (Holcomb et al., 2014). Between 2003 and 2013, advertising revenue for newspaper companies declined more than 55 percent ($46.1 billion to $20.7 billion), with a $2.2 billion increase in online revenue doing little to offset the $27.6 billion drop in print revenue. Notably, these declines have occurred across categories (e.g., retail, national, and classified advertising). Furthermore, while many newspaper executives were once hopeful that digital advertising rates would eventually become comparable to their print counterparts, this has failed to materialize, leading to considerable uncertainty about the long-term prospects of relying primarily on digital advertising revenue (Doctor, 2010).

Finally, whereas competition with traditional competitors has in many cases decreased, a litany of new competitors has emerged. With regard to traditional competitors, in 2009 there were 65 fewer daily newspapers in the U.S. than in 2005—and since 2009, a number of high-profile newspapers like the *Rocky Mountain News* have ceased to exist while others, such as the *Times-Picayune* and the *Oregonian*, have reduced their publishing frequency. Furthermore, traditional media use for the purposes of news consumption has been in decline. A series of biennial surveys by the Pew

Research Center found that the percentage of Americans who obtain their news from newspapers has been in a consistent decline over the past decade, contracting from 42 percent in 2004 to 29 percent in 2012. In contrast, there has been a marked increase in the use of the Internet to obtain news. The aforementioned Pew surveys also indicated that the percentage of Americans who obtain their news from networked platforms has increased from 23 percent in 2006 to 39 percent in 2012. Thus, even as demand drops for their products, print news organizations find themselves no longer competing strictly against a small number of traditional media organizations in their local market, but rather against large regional, national, and international competitors and alternative information sources like blogs and social media.

The Emergence of a Market-Oriented Culture

One response to these environmental changes, according to rational-choice economic perspectives, is for an organization to reconfigure its resources and processes to overcome uncertainty and ensure that it is operating in the most efficient manner possible (Meyer et al., 1993). Indeed, according to McGahan and Porter (1997), whereas industry conditions account for 19 percent of a firm's performance, the competitive strategy the organization adopts accounts for 32 percent of its performance. In the news industry, evidence of this may be seen in the increased attention devoted to digital products, as news organizations, undoubtedly aware of changes in consumption habits, have invested ever-greater amounts of their dwindling resources on strengthening their position online (Soloski, 2013; Tang et al., 2011). While many newspapers continue to

view their print editions as their bread and butter, there is widespread belief that the future lies in adequately monetizing their digital editions.

In particular, one rational response to these developments would be for a news organization to engage in greater organizational integration, or enhance coordination among its autonomous units (Ettlie & Reza, 1992; Lowrey & Woo, 2010; Nooteboom, 2000). Organizations with strong integration are generally viewed as being more flexible in the face of uncertainty (Ven, Delbecq, & Koenig, 1976) and better able to maximize resource utility by sharing resources across organizational boundaries (Grant, 1996; Koster, Stokman, Hodson, & Sanders, 2007). In particular, knowledge—a key asset in the modern economy and in particular information industries—may be more freely shared, and thus better exploited, in well-integrated organizations (Esper, Ellinger, Stank, Flint, & Moon, 2010; Rhodes, Hung, Lok, Lien, & Wu, 2008).

Historically, news organizations, and newspaper organizations in particular, have been organized in a segmented fashion—that is, following a strategy of weak coupling, there has long been a 'wall' separating editorial staff from business staff (Gade, 2004). This has largely made sense since news organizations typically operate in dual markets: they sell their content to an audience and they sell their audience to advertisers (Gade, 2009b). This has been a traditional source of tension for many news organizations, with journalists viewing their mission as enhancing democracy and providing content that serves the public good while the business staff focuses on ensuring that the organization is able to generate sufficient revenue to ensure that economic objectives are met (Achtenhagen & Raviola, 2009; Raviola & Hartmann, 2009; Tunstall, 1971).

The shift toward greater organizational integration in newspaper organizations has been described by scholars as a movement toward 'reader-driven' or 'market-driven' journalism (e.g., Attaway-Fink, 2005; Beam, 1996; Cohen, 2002; McManus, 1994). More broadly, this may be seen as a shift toward a market-oriented culture, or "the dominant, dynamic segment of an organization whose orientation, attitudes and actions are geared towards the market" (Harris, 1998, p. 360), which is not a transformation that is unique to journalism but one that marks a crucial shift in the work of journalists. Scholars have noted that this shift began in the 1980s and gained traction in the 1990s amidst a wave of hostile takeovers and public offerings for newspaper organizations (McManus, 1994; Underwood, 1993). Indeed, as a 2000 cover story in the *American Journalism Review* indicated, newspapers were becoming more "reader friendly," viewing them as partners in decision-making about news, focusing on stories that were unlikely to offend readers, and promoting greater information-sharing with marketing departments (Stepp, 2000).

However, it is worth keeping in mind that newspaper companies, even during the turbulent early 1990s, were still generating exceptional profit (Cranberg, Bezanson, & Soloski, 2001). Indeed, as Martin (1998) found in a study of corporate profits between 1984 and 1994, newspaper companies averaged profits that were more than 90 percent higher than those of publishing companies, and nearly 54 percent higher than corporate bond yields. Thus, the incentive for organizations to adopt a market-oriented culture was primarily driven by a desire to increase profits and maintain high share prices.

In contrast, newspaper companies today face exceptional challenges that threaten their very existence—indeed, websites like *Newspaper Death Watch* (Gillin, 2007) have

emerged to chronicle the woes of the industry—and their employees work under the shroud of voluntary buyouts and involuntary downsizing (Soloski, 2013). Thus, while the traditional 'wall' between the editorial and business staffs has been eroding for some time, this process has accelerated in recent years as organizations face major annual losses and declining prospects, and as some journalists become more concerned with their jobs than their journalistic values (Coddington, 2015). That is, whereas there was previously some economic pressure on traditional media organizations to yield more profit, there is now an outright imperative to increase revenue and trim costs just to remain operational. In 2012 alone, the newspaper industry shed 2,600 jobs—and across the industry, there was a 29.1 percent drop in the number of newsroom jobs between 2005 and 2012 (Mitchell et al., 2014). There is, therefore, more pressure than ever for journalists and editors—and not just management and business staff—to treat the audience as a market and downplay the so-called separation of 'church' and 'state' (Fengler & Ruß-Mohl, 2008).[8]

Toward the Agenda of the Audience

Because the primary economic model for online news is premised on advertising—that is, providing content at nominal cost in order to capture large, attractive audiences—there is increasing pressure on news organizations to ensure that news

---

[8] The "separation of 'church' and 'state'" is a phrase often used among news professionals to refer to the importance of keeping business concerns separate from editorial considerations. It is, of course, largely a product of journalism's own myth-making, rather than a reflection of journalistic practice (Hampton, 2010).

products align with audience demands.[9] More specifically, news organizations in an online environment are being increasingly asked to cater to audience interests in order to generate more 'clicks' (Tang et al., 2011), even if it comes at the expense of journalistic values (Cohen, 2002).

As DeWerth-Pallmeyer (1997) notes, there is often tension between what audiences find to be interesting and what editorial newsworkers find to be important. These concepts are by no means mutually exclusive: a news product may be both interesting *and* important.[10] However, in many occasions, this is not the case. For example, a story about the city council's vote on appropriations for snowplowing equipment may be of great import to a community (e.g., a lack of such equipment may result in workers not being able to go to work during a bad winter, causing economic damage to the community) but be of little interest to anyone besides local government wonks.

Boczkowski, in particular, has conducted extensive research on the convergence and divergence of the preferences of editorial newsworkers and their audiences, terming the discrepancy a choice gap. In one study of four large U.S. online news websites, Boczkowski and Peer (2011) found that there was a 'sizable' gap between journalists'

---

[9] While many news organizations have introduced various forms of paywalls to monetize the access to content, advertising remains a substantial, and often dominant, source of revenue.

[10] As Dennis and Merrill (1984, p. 146) note: "A good editor is one who recognizes that it is a journalistic responsibility to provide the reader with some significant and useful news which may or not be of great immediate interest or appeal; at the same time, the editor knows that, in order to get the readers exposed to such news, he must also provide types of news of a more shallow—perhaps even more sensational—nature. The good editor is a pragmatist and a realist, not some one-dimensional person seeking either to entertain or to educate."

choices and consumers' choices, with journalists favoring news items that pertained to public affairs more often than audiences. In a second study, Boczkowski and colleagues (2011) found a similar phenomenon with 11 online news websites in six different countries in Western Europe and Latin America. These findings were expanded in a broader synthesis by Boczkowski and Mitchelstein (2013), and are supported by the work of Lee and Chyi (2014), who found that content that was deemed to be 'newsworthy' by media professionals is not always considered 'noteworthy' by audiences. In a survey of 767 U.S. adults, Lee and Chyi found that respondents believed that little more than one-third of news content produced by mainstream media was noteworthy—that is, nearly two-thirds of the content shown to them was deemed to be irrelevant or uninteresting. In another study, Boczkowski (2010) observed that journalists in two Argentinian newspapers were becoming increasingly aware of what their online audience was interested in. Boczkowski noted that in the face of a "tension" (p. 153) between readers' preferences and their own news judgment, journalists tended to stick to occupational values.

The existence of such a gap is notable because a rational-choice economic perspective would suggest that organizations should respond to uncertain market conditions (i.e. the current news media market) by reconfiguring their resources and processes to ensure tighter organizational integration (i.e. sharing knowledge about the audience) and the emergence of a market-oriented culture would demand that news products become more closely aligned with audience demands (Gade, 2009a; Lowrey, 2011; Meyer et al., 1993; Snow et al., 2005). That is, in the face of troubling economic

prospects, a news organization should seek to engage in greater information sharing among its editorial and business (and information technology) departments to learn more about what the audience wants and provide them with that content. The findings of Boczkowski and colleagues (2010, 2011) and Lee and Chyi (2014) offer evidence that news organizations have not yet maximized the economic potential of appealing to their audience—and could seek to in the face of difficult economic conditions.

The consequence of this shift should not be understated: it would realize a shift toward what Anderson (2011b, p. 529) has termed the "agenda of the audience." Indeed, it is a stark contrast to the traditional view of professional journalism wherein audiences are seen as being passive consumers who are easy to ignore and journalists, as experts in separating information individuals need from information they do not, set the agenda (Shoemaker, 1991). The shift toward an "agenda of the audience" would represent a movement toward a more conversational form of journalism at best and mere populism at worst. That is, it could present the opportunity for journalists to understand the kinds of content that their audience members want and to provide them with rich, potentially civic-minded information pertaining to that demand (Hindman, forthcoming). However, under a market-oriented logic, it more simply presents the opportunity to minimize the public-service mission that is central to the occupational ideology of journalism in order to maximize the return on investment in a news product (Fengler & Ruß-Mohl, 2008; Nguyen, 2013).

Audience Analytics and Metrics in the Newsroom

Much of the recent research on the use of audience analytics in newsrooms indicates that audience analytics have been adopted in some fashion by the majority of news organizations, with the most commonly used systems being Chartbeat, Omniture, Parse.ly, comScore, Nielsen NetRatings, and Google Analytics (see Anderson, 2011a; Graves & Kelly, 2010; Groves & Brown, 2011; MacGregor, 2007; Schaudt & Carpenter, 2009).[11] These systems can be easily incorporated into any website, often by including a small piece of code that seamlessly collects and directs information to the third party. In many cases, these third parties benefit by being able to collect information about user behaviors, which they can then anonymize, aggregate, and resell to other clients. News organizations, meanwhile, are given access to data that has been distilled and is presented in a comprehensible manner. Alternatively, news organizations may utilize systems that are developed in-house. It is not uncommon for a news organization to use multiple software to gather and analyze audience analytics, with different categories of employees (e.g., managers) having access to more sophisticated analytics than others (e.g., journalists), as well as more training on how to utilize such data (Graves & Kelly, 2010; Usher, 2013).

According to Nguyen (2013), newsworkers may turn to two distinct sets of audience metrics: internal metrics and external metrics. Internal metrics include data about how a site is utilized by a user during their visit, which Nguyen decomposes into two subgroups: data about traffic to and from the site (including hits, visits, unique

---

[11] Omniture was purchased by Adobe Systems in 2009 and rebranded as Adobe Marketing Cloud in 2012.

visitors, geographic origin, referrer sites, time of visit, whether they are new or returning visitors, and where they go afterward) and data about user behaviors (including how many comments a story receives, how often an item is shared via e-mail or on social media, the most common search terms used on the site, and the average time spent on a story). External metrics consist of information about what is trending on the web, which includes information about what is buzzing on Facebook and Twitter, and the specific keywords that are being utilized on search engines.

The Use of Analytics and Metrics by Editorial Newsworkers

The growing role of the audience has received substantial attention in the scholarly literature in recent years, with a handful of scholars making it central to their inquiries. Their scholarship broadly indicates that newsworkers are becoming increasingly sensitive to audience metrics, although its use—and how it is thought and talked about—appears to vary across newsrooms. In particular, the limited evidence appears to indicate that audience analytics and metrics have, at minimum, the capacity to affect different aspects of newswork, such as the production and presentation of news.[12]

MacGregor (2007) conducted in-depth interviews with 19 journalists in senior or

---

[12] While the present research focuses on changes in editorial decision-making by news professionals, it is worth noting here that, in a small number of news organizations, humans play a marginal role in determining what news to cover and how much prominence to accord those items. Indeed, Anderson (2011b) uses the term "algorithmic intelligence" (p. 536) to describe the methodology employed by companies like Demand Media wherein computer algorithms are leveraged to learn about audience preferences, determine the story topics that will yield the greatest return, and make decisions about what to cover based strictly on computer-generated metrics. While such organizations certainly offer great potential for academic study, they largely operate on the margins of the online news ecosystem, and introduce a set of questions that extend beyond the present research.

mid-level positions at a range of media outlets and found "prolific use" (p. 286) of audience metrics that were largely collected by third-party companies like Omniture. Of note was MacGregor's finding that while a minority of journalists "obsessively" (p. 289) viewed the metrics with "a hawk eye," (p. 289) the majority did not look at audience metrics about a story until the following day—and in both cases, journalists across print, broadcast, and net-native organizations said it rarely led to an instant, real-time modification of a story.

More recently, Anderson (2011a) echoed some of MacGregor's findings, noting that newsworkers in some of the newsrooms he visited had access to, and were "obsessed" (p. 558) with, traffic statistics. However, Anderson found that organizations made "major" (p. 559) editorial decisions based on those figures. This led Anderson to conclude that, for at least one of the newsrooms, "it is not an exaggeration to say that website traffic often appeared to be the *primary ingredient* in *Philly.com* news judgment" (p. 561, emphasis his). Indeed, these figures were often invoked in presentations and in monthly reports, but also consulted over the course of the day to determine how much play to give an item. As one web producer told him:

> We're trying to be a real strong local news site that appeals to our audience and gets traffic. … As far as the spotlight versus the biggie goes, it's intuitive, but we just put about anything we think will get clicked up there at this point. You just have a gut feeling about it. Like for an article about Michelle Obama: your gut instinct is that it's not going to get picked up, but if it's getting clicked, we'll bump it up. (p. 562)

Vu (2014) similarly found that the majority of the 318 U.S. editors he surveyed monitored web traffic on a regular basis, with nearly half of those doing so on a daily basis. According to Vu, they were most likely to use it to make popular articles more prominent, followed by looking for follow-up articles to popular content, looking for additional multimedia elements to incorporate into that content, trying to update popular content more often, run articles similar to popular content, make unpopular articles less prominent, and try to find possible editorials for popular content. Finally, Vu found that when considering whether or not to run an article, editors were most influenced by two audience-centric considerations: whether readers would read it and whether readers needed to know about it. This led Vu to conclude that "editors are willing to adjust their editorial decision-making based on metrics" (p. 11).[13]

Studying the *Christian Science Monitor*'s transition from a print newspaper to an online-only venture, both Groves and Brown (2011) and Usher (2012) painted pictures of an organization that became metrics-driven and had to deal with considerable tension in reconciling its journalistic mission with the demands of a click-driven culture.[14] Their overlapping fieldwork indicated that, before the change to online-only, journalists at the *Christian Science Monitor* feared that the shift would undermine their journalistic values

---

[13] However, one should exercise some caution in interpreting these findings as key questions were answered using a four-point scale that included "not likely," "somewhat likely," "likely," and "very likely" as the intervals. Such a scale may have biased the findings since respondents often avoid the polar ends of scales, thus privileging the options of "somewhat likely" and "likely," and thereby ostensibly overstating the extent to which such data are used and how influential they are.

[14] The *Christian Science Monitor* did retain a weekly print magazine, although Groves and Brown (2011) and Usher (2012) both indicate that, at least strategically, it received considerably less attention and was not central to the organization's view of its long-term future.

(e.g., report on inconsequential stories that generated traffic), reduce the control they had over their work (e.g., having less latitude to pitch stories), and alter their workflow and routines (e.g., contact a single, dependable source for a quick quote in order to get stories up as quickly as possible). Shortly after the shift, those who stayed—it is important to note that dozens of newsroom jobs were eliminated through attrition and buy-outs, ostensibly affecting those who were most skeptical and reluctant to change—were beginning to overcome their fears and were able to adapt their practices. Of note was their belief that they, and the organization, were able to stay true to their core values; thus, "journalists did not feel a challenge to their role as authoritative storytellers" (Usher, 2012, p. 1907). Additionally, at this point, awareness of audience metrics was becoming more prevalent in the newsroom, but these measures had not yet begun to influence editorial decision-making.

However, within eight months of the shift to online-only, a metrics-oriented culture had begun to take root—driven largely by editors and managers—with journalists becoming increasingly aware of traffic figures and story performance quickly becoming the "primary measure of success" (Groves & Brown, 2011). Indeed, daily news meetings began with a presentation of the traffic figures for specific stories, with editors trying to figure out what generated traffic and what did not. According to Usher (2012), journalists were becoming increasingly demoralized by the emergent culture—they wanted to resist the influence of metrics, but were cognizant of the reality that their success was largely measured by it. Groves and Brown (2011) do not explicitly touch on the budding demoralization of the newsroom, but provide extensive examples of the frustration felt by

several journalists who perceived that they were short-changing their readers.

Although Usher's fieldwork ended in February 2010, Groves and Brown indicate that, by January of 2011, this tension had started to resolve itself, with most of the staff having adapted their routines to fit traffic-oriented aims. Furthermore, Groves and Brown note that the skepticism over strategy and technology had largely, though not entirely, abated. Groves and Brown attribute this shift to the perception of success vis-à-vis its venture online; thus, "as page views rose, success validated and embedded new routines, turning implementation into confirmation."[15] It should be noted, however, that the *Christian Science Monitor* represents an exceptional case; unlike newspapers that have devoted more resources to their online presence (e.g., *The New York Times* and *The Washington Post*) and those who have just cut back on their print offerings (e.g., the *Detroit Free Press*), the *Christian Science Monitor* largely divested itself of its print product and strategically oriented itself as a digital-native news organization.

MacGregor (2007) has noted that there are multiple potential responses to metrics data, from stimulus-response behavior characterized by immediate reactions to deliberated and mediated responses in which such data are weighted against other factors. Specifically, MacGregor observed that the journalists he spoke with saw three uses for such data: to see which particular stories are popular; to assess trends over time across the site; and to offer "objective" (p. 290) evidence to supplement journalists' perceptions. MacGregor concluded that audience metrics are refining professional practices and

---

[15] According to Groves and Brown (2011), the *Christian Science Monitor*'s website began to near its ambitious traffic goals and its stories routinely placed among the top results on Google and Google News.

emerging tensions, but not independently producing new professional procedures and beliefs.

In the case of the *Christian Science Monitor*, audience metrics were utilized extensively to determine the kinds of content that should be produced and how content should be presented to yield maximum exposure (Groves & Brown, 2011; Usher, 2012). Content that failed to garner page views, such as podcasts with journalists and other multimedia offerings, were quickly done away with. Additionally, the term "riding the Google wave" quickly became part of the newsroom's lexicon, with editors making use of data from Google Trends to figure out what readers were looking for and employing search engine optimization (SEO) strategies to write headlines that would increase an article's chances of coming up as a top Google search result.

However, even as metrics have "become inextricably linked to the heart of newsroom operations and editorial decision making" (Usher, 2012, p. 1911) for some organizations like the *Christian Science Monitor*, the scholarship also indicates that some newsrooms do not make much use—or particularly sophisticated use—of audience analytics and metrics. Anderson (2011a) observed that while the use of audience metrics was extensive and strategic in the Philadelphia newsrooms he visited, it was far less so at the Newark *Star-Ledger*. Anderson attributed this difference to organizational management strategy, with the Philadelphia newsrooms and *Philly.com* in particular, placing far greater emphasis on the diffusion of audience metrics. Similarly, Usher (2013) found that top online editors at *Al Jazeera English Online* largely ignored metrics and attributed this to the organization's culture. Indeed, in analyzing the response from one of

the company's senior Web editors—who articulated opposition to chasing traffic—Usher noted:

> His voice is an especially important one to acknowledge as he actually set up and organized the page during the day, deciding where to place each story and finalized each headline. Thus, for him to ignore traffic patterns and his sense of prioritizing his news judgment over what was important for the public to know versus "giving the masses what they want" was quite important. He was on the front lines of determining whether to take traffic into account, and chose not to encourage his editorial team to consider these numbers. (p. 344)

This is of import, Usher aptly notes, because it set the tone for the organization—that is, it established a culture of not deferring news judgment to metrics data—and sent a signal to reporters that they need not pay attention to such data. However, although the journalists and editors at *Al Jazeera English Online* argued that they did not turn to audience metrics to guide their judgment, they did make use of them for the purposes of self-validation. As one journalist noted, "it is used for moral uplift more than anything else" (quoted in Usher, 2013, p. 346). This observation is notable because it may suggest the possibility of a more subtle effect from metrics data: rather than consciously driving decisions—both immediate decisions about how to alter the presentation of a story over the course of a day and long-term decisions relating to editorial strategy—these data may be lead to an acculturation process whereby journalists and editors unconsciously adopt feedback from the audience and use it to please their ego (Usher, 2013). Put differently, it is possible that newsworkers may begin to slowly alter their judgment to ensure greater

appeal to the audience not as a cognizant response to audience demands, but rather to increase the likelihood that their decisions will be validated. Indeed, as Napoli (2011) notes, such data can be utilized both for the formulation of decisions as well as their justification.

Critically, Usher (2013) did note that *Al Jazeera English Online* is a fairly unique case in that it is funded almost exclusively by the Qatari government, thereby sheltering it from economic concerns. However, these observations indicate that the use of metrics is not ubiquitous, and that there is a need to consider the context under which a given newsroom operates. Indeed, the importance of economic orientation is also observed by MacGregor (2007, p. 282), who notes that audience metrics may "be understood in terms of market goals to 'serve' customers and expand markets."

Additionally, even in newsrooms that have sophisticated systems, access to audience metrics is not universal. That is, certain metrics may only be available to some newsworkers and not others. Indeed, in her analysis of *Al Jazeera English Online*, Usher (2013) observed that although top managers had access to more sophisticated metrics, reporters did not. Additionally, Usher observed that most journalists had difficulty interpreting the data they did have access to, largely because of the lack of a metrics-driven culture in that newsroom. Thus, even when a newsworker has access to a set of metrics, there is no guarantee that he or she will know how to use it. Indeed, as Graves and Kelly (2010) have noted, confusion over how to interpret and make use of audience metrics may be more prevalent than scholars often estimate. Petre (2015) argues that the context within which a newsroom operates plays a significant role in the extent and

manner in which audience analytics and metrics are used to inform editorial decision-making, with her investigation pointing to very different uses by the Gawker network of websites and the *New York Times*.

Tandoc (2014b), in a survey of 276 online news editors, found that online editors used metrics primarily for monitoring traffic, but found some evidence that they were also factoring them into editorial decision-making processes. Specifically, Tandoc found two pathways influencing the use of analytics. In the first pathway, considering the audience in terms of economic capital (i.e. value to advertisers) led the editors to also consider them in terms of symbolic capital (i.e. the organization's credibility), which in turn had a small effect on certain perceived content-related decisions, like selecting which stories to cover, which stories to do follow-ups on, and how to cover those stories. In the second pathway, the extent of competition faced the organization led to the adoption of audience metrics, which in turn led to the use of metrics for monitoring traffic and keeping the website functional and then to perceived content-related uses. This led Tandoc to conclude that while online editors maintain some agency in mediating the impact of metrics on their work, they are also constrained by organizational and socio-institutional structures.

The Impact of Metrics on Content

Little of the work in the burgeoning stream of research on audience analytics and metrics has directly assessed the relationship between metrics and journalistic content—that is, how news products are changing in response to the use of audience metrics. Such

analyses are important to complement self-reported data and the interactions observed by researchers because potential effects may be easily overstated or understated due to social desirability and confirmation biases (Kreuter, Presser, & Tourangeau, 2008). To that end, the work of Lee and colleagues (2014) and Bright and Nicholls (2013) serve as welcome, though isolated, contributions to the literature.

In an analysis of three New York-based news websites using time-lagged structural equation models, Lee and colleagues (2014) found that audience clicks affected the placement of news stories, and that this effect varied over the course of the day. Of particular interest is that Lee et al. found the effect of audience clicks to be negative—that is, as stories became more popular, they became less prominent. They reason that this is because editors recognize that those stories have already gained sufficient prominence by virtue of being on the 'most viewed' list; thus, perhaps, they can replace them with other content in the hopes of attracting traffic to that content. Notably, however, when Lee et al.'s findings are disaggregated, one finds that there is a positive effect for one website (nytimes.com), no effect for a second website (nypost.com), and a negative effect for a third website (nydailynews.com). These findings would thus lend support to the contention that some websites are more sensitive to audience metrics than others—or that they may use that information for different purposes.

Bright and Nicholls (2013) took a different approach to measuring the effect of story popularity on the manner it is presented: assessing if it had an impact on the short-run likelihood of the article being removed from the front page. In an analysis of a month's worth of data from five leading U.K. newspapers, Bright and Nicholls found that

the average article spent 15 hours on the front page of the website and that roughly one-tenth of the articles on the front page were featured on the most-read list at one point. More importantly, they found that articles appearing on a most-read list remained on the front page for three hours longer than articles that did not appear on the list, and through the use of Cox Proportional Hazards models—enabling them to measure the effect that being on a most-read list would have on an article's 'survival' time 15 minutes later—they found that the risk of being subsequently removed from the front page was 26 percent lower for articles that appeared on a most-read list than those that did not. Bright and Nicholls also found little difference in the effects between 'soft' entertainment news than 'hard' political news. Ultimately, they concluded that there was a measurable impact of the popularity of a story on its lifespan on the front page of a news website.

Synthesizing the Literature

As this examination has shown, it is useful to view audiences through a constructivist lens when studying the influence of audiences on newswork and news products. That is, while a tangible audience may exist, it is only of relevance insofar as it aligns with the audience constructed in the mind of the newsworker. Furthermore, there is reason to believe that technological advances and shifts in media consumption, and in particular growing audience fragmentation and increased audience autonomy, have initiated a third-wave toward the rationalization of audience understanding. In response to this development, a new set of audience information systems, audience analytics, has emerged—and the outputs of those systems, audience metrics, are capable of

significantly altering how audiences are constructed.

To help explain why audience analytics and metrics have become appealing to news organizations, the state of the industry was evaluated with the preponderance of evidence suggesting that, in recent years, it has been, and continues to be, in a state of uncertainty. Drawing on theories of organizational change, a rational response to environmental uncertainty would be to reconfigure organizational resources and processes to facilitate integration and the sharing of knowledge. Such integration has been complemented, if not driven, by the development of a market-oriented culture in the newsroom that has accelerated in recent years. Nevertheless, a series of studies have shown that there exists an exploitable gap between what editorial newsworkers consider to be newsworthy and what audience members noteworthy, and that audience metrics could reduce that so-called choice gap.

This examination has also indicated that the majority of the scholarship suggests that audience analytics have been adopted by a number of news organizations and that metrics have become important objects within many newsrooms. In particular, the literature indicates that there are several different relevant metrics that can be used and rationalized in different ways—though the 'page view' remains central. For example, metrics may be used to formulate long-term strategy like what to cover and how to cover it as well as short-term decisions about what content to promote and demote on a homepage. Furthermore, they can be used in a justificatory manner, such as to validate decisions regarding a story as well as to reward newsworkers that perform well according to specific metrics.

Though these studies indicate that there is widespread access, they also point to distinct differences in the extent to which they are used. While some individuals and organizations prioritize chasing clicks and making data-driven decisions, others do not. These studies, however, are usually only able to speculate about the effects metrics are having on news content, with the literature assessing their direct relationship lagging behind. Notably, the two small-scale studies that have empirically assessed that relationship indicate that audience metrics have some effect on content, though they vary somewhat across organizations.

It seems, then, that audience analytics and metrics are worthwhile objects of study because of their potential to alter journalistic processes like gatekeeping and the ideals associated with journalism, like the watchdog and communitarian roles of the press (Christians, 1997; Croteau & Hoynes, 2001; Lee Plaisance, 2005; Siebert et al., 1956). As Hindman (forthcoming) notes, it can be beneficial to include the audience in considering what and how to cover different topics. However, it is equally important that journalism does not employ the rhetoric of "empowerment" to justify unreflexive use of audience data to guide editorial decision-making, effectively treating the audience strictly as a market (see also Couldry & Turow, 2014; Turow, 2005). It is this latter point that drives anxiety for many scholars, who fear the prospect of a data-driven shift to catering to the demand of news consumers (Nguyen, 2013; Tandoc & Thomas, 2015).

In assessing the extent to which journalism is effectively changing, both in terms of practices and substance, as a result of the proliferation of audience analytics and metrics, Tandoc (2014b) is correct in stating that much remains unknown. For example,

scholars continue to have a limited understanding of the factors that drive the use of audience analytics and audience metrics by certain organizations and classes of newsworkers, and not others. Additionally, the literature has only recently begun to analyze how specific aspects of journalistic work (i.e., specific editorial decisions and functions) are being influenced by audience analytics and metrics, rather than viewing them in aggregate. Furthermore, the vast majority of research has focused on the impact that audience analytics and metrics have on journalistic practice, with comparatively little attention paid to the effects they have on journalistic content. Lastly, the vast majority of the attention has been devoted to digital-native news organizations and those far ahead of the digital curve, with few studies incorporating mid-size and community news organizations. Thus, although the scholarship on audience analytics and metrics has yielded a number of insights in recent years, there are still several critical questions that remain unanswered and avenues that scholars may pursue in adding to that body of literature.

CHAPTER III:
COMPUTATIONALLY ANALYZING LIQUID CONTENT

According to Shah and colleagues (2015), we are in the midst a turn toward computational social science, a paradigmatic shift characterized by the use of large and complex datasets drawn from digital media sources that must be analyzed through the use of computational and algorithmic solutions. This is not to imply that small-scale or manual analyses are going away; they are not and in many cases offer the most appropriate approach (Crawford, Gray, & Miltner, 2014; Karpf, 2012). Rather, it points to a growing recognition of the benefits (though with their own unique limitations) of using computational tools to drive analyses (Zamith & Lewis, 2015). Within the field of mass communication alone, a diverse array of social scientific research adopting computational methods has been published in recent years, involving the analysis of anywhere from tens of thousands to billions of units—tweets, forum postings, news articles, and the like (e.g., Bruns, 2013; Burgess & Bruns, 2012; Grimmer & Stewart, 2013; Hermida, Lewis, & Zamith, 2014; Kirilenko & Stepchenkova, 2012).

In conjunction with these empirical analyses, mass communication scholars have made a series of methodological contributions in recent years to guide automated and semi-automated analyses of content (e.g., Baek, Cappella, & Bindman, 2011; Bruns & Burgess, 2012; Grimmer & Stewart, 2013). One area in this stream of research that has received substantial attention in recent years has been the analysis of 'liquid' content, or content that is not only mutable but constantly changing (Deuze, 2008; Karlsson & Strömbäck, 2010; Karlsson, 2012; see also Lowrey, 2006; Matheson, 2004; Sjøvaag, Stavelin, & Moe, 2015; Tremayne, Weiss, & Alves, 2007). Of particular concern within

that stream is how to effectively capture and analyze rapidly changing web content, which introduces an array of methodological challenges (Herring, 2010; Sjøvaag & Stavelin, 2012).

This chapter focuses on delineating a process for computationally capturing and analyzing aesthetic facets of rapidly changing web content, using as a case study the analysis of the homepages of 21 U.S.-based news organizations. It begins with a review of the literature on liquid content and the methodological challenges associated with studying it. Then, existing approaches to 'freezing' such content and computationally analyzing Web documents are considered. Finally, these insights are synthesized and built upon as part of an analysis of more than 125,000 documents, with the advantages and disadvantages of the process discussed.

Background

As several scholars have observed, online news is distinct from its analog counterpart (e.g., the newspaper), with the characteristics of immediacy and interactivity often touted as primary distinguishing factors (Boczkowski, 2004; Deuze, 2003; Karlsson, 2011; Massey & Levy, 1999; Pavlik, 2000). Interactivity has been described as "the extent to which users can participate in modifying the form and content of a mediated environment in real time" (Steuer, 1992, p. 84), and broadly refers to the additional control granted to the consumers of the content. This may include commenting on a news article, filtering the content that appears on a site, or choosing whether or not to play a video or maximize an image that is appended to a page. Interactivity effectively

sets online journalism apart by enabling greater integration of user-generated content, and by allowing users to engage with content in different ways, such as by selecting what features to display on a crime map visualization (Smit, de Haan, & Buijs, 2014).

Immediacy, in turn, refers to the absence of a delay between when producers create or update content and when consumers can view that content (Lim, 2012). In a different context, immediacy may also refer to the related expectation among online news consumers that content will be updated frequently to ensure that fresh material appears the next time a page is accessed (García Avilés, León, Sanders, & Harrison, 2004; Pavlik, 2001). This, in turn, has altered journalism as it has become "less a product than a process, witnessed in real time and in public" (Tumber, 2001, p. 98), therefore moving from a 'black box' model to something less opaque (Karlsson, 2011; Singer, 2005). Collectively, interactivity and immediacy have promoted the development of what researchers have varyingly termed 'liquid,' 'fluid,' and 'dynamic' journalism (Deuze, 2008; Engebretsen, 2006; Karlsson & Strömbäck, 2010; Singer, 2006).

Although there are various ways to study a website's liquidity (see Karlsson, 2012), one approach is to focus on the extent to which content in key areas of a page changes. The homepage is one particular section of a website that news consumers would expect to be fluid through the addition of breaking news and updates to existing articles (Sjøvaag et al., 2015). Though homepages no longer serve as the primary point of entry for many news consumers (see The New York Times Company, 2014), they are nevertheless generally viewed by news organizations as an important page through which to build an audience and engage in core functions of newswork, like communicating

news priorities (Benton, 2015).

However, as Lim (2012) notes, the notion that the homepages of news organizations are constantly changing is has long been taken for granted by scholars, and that empirical work assessing the extent to which news homepages change over the course of the day remains fairly limited. One exploratory study, drawing on a small sample of data from 30 U.S. daily newspapers in 2003, found that large news organizations 'frequently' updated their homepages, though many smaller news organizations treated their websites as 'shovelware' (Alves & Weiss, 2004). In another study, Lim (2010) found that the websites of four large news organizations—the *Los Angeles Times*, the *New York Times*, *USA Today*, and the *Washington Post*—typically updated between 69.0% and 83.4% of their content in the most prominent regions every seven hours. In contrast, however, Lim (2012) found relatively little change in the most prominent areas of 13 popular Korean news websites when using a shorter window of time, with anywhere between two to nine percent of the items changing every 30 minutes.

Though the evidence pertaining to the extent of fluidity on these homepages is mixed, the prospect of constantly changing websites offers mass communication researchers several opportunities for studying the process of how particular news content evolves over short periods of time (Karlsson, 2012; Lim, 2012; Sjøvaag et al., 2015). However, that prospect also introduces a number of methodological challenges for the researcher (Herring, 2010; Sjøvaag & Stavelin, 2012). In particular, as Deuze (2008, p. 861) notes, "the study of content has always rested on the premise that content actually exists, that it genuinely can be considered as a finished, static object of study. In the

current media ecology of endless remixes, mashups, and continuous edits, that is a problematic assumption." This begs the questions: how should one capture dynamic objects that may constantly be in a state of flux?

Freezing Liquid Content

In order to analyze dynamic objects like homepages, it is often necessary to first "freeze" them (Karlsson & Strömbäck, 2010, p. 16; see also Leetaru, 2012; Sjøvaag & Stavelin, 2012). In a manual content analysis, for example, turning dynamic objects into static ones is essential for ensuring that multiple coders are able to view the same content when establishing intercoder reliability. Similarly, in a computational content analysis, freezing objects is essential for developing a corpus on which an algorithm may be repeatedly trained and/or tested. In both instances, it is important to freeze content in order to reproduce the research.

Unfortunately for researchers, the act of freezing objects must often be performed by the analyst. For example, although there are a few large online archives of websites, even the most reobust of these, the Internet Archive, will, at most, take only a handful of electronic snapshots each day of the homepages of large news organizations like the *New York Times*. Furthermore, it will often fail to archive homepages of smaller organizations for several days—and even weeks—at a time. In the absence of a third-party archive, researchers must create their own archive for the content they wish to study.

Karlsson and Strömbäck (2010) point to three techniques for accomplishing this. The first technique is to simply take screenshots of the content. As they note, the

advantage of using a screenshot is that it accurately captures the appearance of content as it is displayed under certain conditions (e.g., using a specific browser on a particular operating system). However, they argue that screenshots only capture what appears in a specific frame—that is, it fails to capture content that one would need to scroll up and down to see. Furthermore, several pieces of information are lost when interactive objects are transformed into flat images, such as the URL that a headline links to.

The second technique is to print a copy of the object, either on paper or to a PDF document. Karlsson and Strömbäck (2010) note that this allows the entire page to be captured as a single static object, but that it often triggers a 'printable version' of the page that excludes important information (e.g., certain images or interactive features like reader comments). Moreover, when a printable version is not available or bypassed, aesthetic elements are often forcibly rearranged in the printed copy to ensure that the content fits the printed page. As with the screenshots, a great deal of information is lost when a webpage is transformed into a static object like an image or print-out.

The third technique is to "mirror" the page—that is, to download the source code necessary to render the page, including all of the associated media content such as the images and style sheets. One may do this manually by using the page saving feature of his or her preferred browser. Alternatively, one may do this computationally through the use of free and commercial software, such as HTTrack or WebCopy. While this technique is advantageous in its ability to capture the greatest amount of information about the dynamic objects (e.g., information about the structure of the document, such as the styling of a particular element and link information), these approaches often fail to

capture all of the information necessary to exactly replicate the page (e.g., downloading external JavaScript files and adjusting hard-linked elements to the archived content). That is, even when all the constituent elements are downloaded, they often fail to later replicate the exact appearance of a page at the time it was mirrored.

The study of the liquidity of content often requires researchers to look at short periods of time (Lim, 2012). For example, during a breaking news event, an article may be updated several times over the course of a day and an organization's homepage may change the art associated with that item every few minutes during the initial stages of reporting. This is an important consideration because research designs that involve short intervals tend to generate large datasets. Consider, for example, an analysis of the news items appearing on the homepages of five news organizations that have, on average, 75 links to news items on each page, using an interval of an hour over the period of a single week. Such an analysis would involve capturing 840 distinct snapshots and coding information for 63,000 URLs.[1] Under a manual framework of data collection and analysis, even this fairly small analysis would likely be time and cost prohibitive (and likely error-prone given the magnitude and type of data). Indeed, as Karlsson (2012) notes, the lack of large-scale empirical analyses of the liquidity of content is largely due to the difficulty of the work. How, then, can such an analysis be comprehensively performed through computational means?

---

[1] Though the number of distinct news items would be far lower than that (several items would certainly persist over multiple intervals), it would nevertheless require some review every one of those URLs.

Computationally Analyzing Homepages

In order to tackle larger-scale analyses, it is preferable to identify computational solutions (Shah et al., 2015). Computers are sequential, deterministic machines capable of processing vast amounts of numerical and textual data with exceptional speed and perfect reliability. These two features—speed and reliability—have long made computers an enticing aide for scholars interested in content analysis, with early applications of computer-assisted content analysis dating back to the 1950s and 1960s (see Krippendorff, 2013; Riffe, Lacy, & Fico, 2014).

According to Zamith and Lewis (2015), computational approaches to content analysis offer a range of benefits, such as increased efficiency, transparency, and post-hoc malleability, even as they are limited by the kind of content they are able to process. Indeed, computational aides for content analysts have been critiqued as often being difficult to use and yielding results of questionable validity (Conway, 2006; Mahrt & Scharkow, 2013). This has led some scholars to argue that, in contexts where latent meanings are of interest, a hybrid approach that blends computational and manual traditions is favorable (Lewis, Zamith, & Hermida, 2013; Sjøvaag, Moe, & Stavelin, 2012; Zamith & Lewis, 2015). However, when the variables of interest are unambiguous, such as the presence or absence of a specific piece of computer code in the source code of a Web document, computational analyses are preferable (Leetaru, 2012).

The study of liquidity often centers on change—the addition of some elements and the removal of others—on digital documents like webpages (Karlsson & Strömbäck, 2010; Karlsson, 2012; Lim, 2012; Sjøvaag et al., 2015). Web pages are, by design, highly

structured digital documents. Among the websites of news organizations, these pages typically consist of a mixture of HyperText Markup Language (HTML)—the standard markup language of the Web—and JavaScript, a dynamic programming language that adds more sophisticated functions to a page through client-side scripts.[2] Both the HTML and JavaScript code are interpreted by the browser when a user accesses a page and the browser renders that code into the audiovisual objects the user sees and hears when the page is loaded.

The de-facto nature of the Web thus provides the researcher with access to the source of the documents. This is important because it presents semi-structured data to the researcher.[3] HTML, in particular, is written in the form of elements that consist of tags enclosed in angle brackets. For example, '<h1>' denotes the beginning of an important heading and '</h1>' denotes its end, with the text in-between representing what the user would see, such as a headline. Furthermore, these elements can be stylized through the use of a 'class' or 'style' attribute and identified through the use of an 'id' attribute.

Because each of these elements must be specified, it is possible to identify

---

[2] The use of client-side here and elsewhere in this work refers to the actions that must be performed by the user's computer, rather than the server serving the content. For example, a client-side script would have the user's computer download an array of items (e.g., information to populate a table or spreadsheet) once and then use the computer's processing power to sort that information from low to high every time a different column was clicked. In contrast, a server-side script would have the server generate a new table (or array of information) every time a column was clicked, forcing the page to be downloaded all over again. Oftentimes, client-side and server-side scripts are combined to load page elements on demand. For example, a page may only load a picture slideshow once the user has scrolled down to the part of the page where the picture is supposed to appear.

[3] The HTML code for a webpage is often generated through the use of server-side scripting languages like PHP and ASP. Though the code to generate pages is not visible to the user, the result—the HTML source code the browser interprets to render a webpage—is.

elements based on their syntax. Indeed, it is this very nature that allows large-scale analyses of links to be performed: the source code is systematically reviewed for the presence of anchor ('a') tags that have an 'href' attribute (De Maeyer, 2013; Etling, Kelly, Faris, & Palfrey, 2010). More importantly, however, it allows the use of selectors—pieces of code that can select other pieces of code based on syntactical features, like the 'id' attribute of an element. The combination of different selectors allows researchers to accurately identify and extract various features of a Web document (Sjøvaag & Stavelin, 2012).

The literature thus indicates that there are different strategies that a researcher can employ for computationally freezing and analyzing features of liquid content. However, how can these strategies be effectively leveraged to enable a comprehensive analysis of the liquidity of news websites, or to assess the relationships between features of dynamic documents like homepages? And, more specifically, how can they be synthesized into a process that can account for large datasets covering multiple different websites using short intervals over a long period of time?

Case Study

In order to demonstrate a process for automatically capturing and coding for structural features of liquid documents, a computational analysis of the homepages of 21 news organizations was performed over two months and using 15-minute intervals. The organizations analyzed, shown in Table 1, are among the largest U.S.-based news organizations and are often studied by researchers interested in U.S. news media (e.g.,

Denham, 2014; Luther & Radovic, 2014; Xu, 2013). The unit of analysis was the individual news item, appearing on the homepage of a given news organization at a given time. A news item was operationalized as a package that includes a headline and a hyperlink (Boczkowski et al., 2011; Boczkowski & Peer, 2011; Lim, 2012). Of particular interest to this study were two variables: the popularity of a given news item and how prominent it was, if at all (for an example of a similar study utilizing a manual approach, see Lee et al., 2014).

The process of computationally capturing and coding the liquid content consisted of four steps that are described in detail: (1) writing computer scripts to freeze the homepages in static snapshots that could be stored locally in an organized fashion; (2) identifying popular and prominent content and coding for their relative popularity and prominence; (3) verifying that the algorithm accurately coded the content; and (4) cleaning the coded entries and generating new datasets from them.

Freezing the Homepages

The first step in the process was to acquire the data. Because the homepages of the news organizations could change at any time as news items were either manually or automatically pushed onto the page and as automated functions were executed (e.g., computer scripts updating the 'most viewed' list), a data collection strategy with short intervals was favored. Consistent with prior work looking at the homepages of news organizations (e.g., Bright & Nicholls, 2013), a 15-minute interval was selected. Shorter intervals are generally preferable because they allow the researcher to later discard

unnecessary data (e.g., if a longer interval is deemed more appropriate at a later point in time) and because they reduce the likely impact of missing data if a page cannot be loaded during a point in data collection (e.g., due to a server or network malfunction). However, short intervals also increase the amount of data that must be stored (and, potentially, later analyzed), creating unique challenges.

Of particular importance to the ability to computationally identify and code for the popularity and prominence of individual news items is the acquisition of the computer code that enables the browser to render a homepage—that is, to display what the user ultimately views when he or she accesses a given homepage. While Karlsson and Strömbäck (2010) point to a manual approach or to the use of different third-party tools, neither option was found to be suitable for this analysis. The first option—manually saving each homepage—required human resources that were unavailable to the researcher (e.g., an individual to load 21 different homepages every 15 minutes for two months and organize each saved file).

The second approach—using third-party tools like HTTrack—was found to be unsatisfactory because some of the homepages made extensive of advanced Web features, namely client-side scripting languages like JavaScript to enable greater interactivity. For example, some homepages included a script to dynamically load the most viewed stories. This code is designed to be loaded and processed by default, but it requires a modern browser to interpret and execute it. This limits the range of existing website mirroring tools, which typically use lightweight solutions that cannot perform those automated, client-side actions. Moreover, some homepages had multiple lists of top

stories (e.g., 'most e-mailed') and in order to access the appropriate list (the 'most viewed' or 'most read' list), a mouse action had to be simulated to select that list and thus execute the script that would retrieve the appropriate information from the server and add it to the page. In light of this, a customized solution was deemed to be necessary.

Custom Python scripts—one for each news organization—were thus developed to freeze the homepages into organized, static files comprising of the page's source code and a screenshot of the entire page.[4] Specifically, each script simulated a browsing session through the use of the Selenium framework, which enables the programmatic control of popular web browsers.[5] A new browser (Mozilla Firefox) window was opened for each news organization and the respective URL for the homepage was automatically entered. Then, there was a forced two-minute delay intended to allow all of the website's elements to finish loading and for any automated and scripted actions to take place (e.g., for video and interstitial ads to disappear). The computer script then performed any additional actions necessary to load all of the necessary components of the page (e.g., clicking on the appropriate 'most viewed' list or loading that information from an external page). The HTML source code *processed by the browser* was then saved. The resulting file was then automatically named, tagged with a UTC date and time, and

---

[4] Python is an a general-purpose, high-level programming language that is commonly used to scrape and analyze online content. For more information, see https://www.python.org/.
[5] Selenium is a software testing framework for web applications that allows for the simulation of web browsers like Mozilla's Firefox and Google's Chrome. For more information, see http://seleniumhq.org/.

organized into an appropriate subdirectory.[6]

It is important to note that the code saved in this process differs from a copy downloaded by a mirroring program like HTTrack because it includes all of the processed client-side actions. For example, instead of having the 'most viewed' list area comprise of a call to a script that would need to be executed (and thus not be ready to be parsed by an algorithm), it was an HTML object with a list of the most popular stories. Furthermore, because these scripts often involve requesting data from the server, executing them at a later point in time would likely fail to yield the desired result (i.e., obtaining data from that point in time). Because only the homepage was reviewed in this analysis, the paywalls that typically restrict access to an organization's news content were not an impediment.

Additionally, in order to aid in the development and assessment of the algorithms that would code each captured snapshot, a screenshot of the entire webpage was also taken. Although Karlsson and Strömbäck (2010) are correct that screenshots typically only capture the visible portion of a page on a screen, the Selenium framework enables the researcher to capture the entire page as a single, full-size image. Although these images were not used in the analysis by the algorithms, they were essential for creating the algorithms and, later, for ensuring that they accurately coded the content.

These scripts were run automatically every 15 minutes over a two-month period

---

[6] UTC refers to the Coordinated Universal Time. It is the primary standard by which time is regulated worldwide. In particular, it facilitates calculating dates and times across time zones and is not sensitive to changes in daylight saving time. Storing date objects in UTC thus reduces the likelihood of error when working with multiple time zones, as was the case in the present work.

(from October 18 to December 20, 2014) using a dedicated Linux server with a quad-core, 3.1 gigahertz AMD processor, 16 gigabytes of RAM, and 1.5 terabytes of hard drive space, and running only free or open-source software. A total of 126,473 snapshots were captured, with the interpreted HTML source code taking up 40.5 gigabytes and the screenshots taking up 610.5 gigabytes of hard drive space.

Coding for Popularity and Prominence

Because most large news organizations—and all news organizations in this study—use content management systems, the structure of each news organization's homepage will generally only contain minor variations that are part of an otherwise consistent design.[7] In the present analysis, each organization had at most a handful of distinct layouts. Thus, although the content on a given website would be in flux over the course of the day, the layout (and, most importantly, the HTML elements used to instruct the browser how to render that layout) would only have minor variations. It was therefore possible to leverage this uniformity to automate the content coding process by accounting for those variations and, for each variation, seeking out common patterns in the code.

With this consideration in mind, a second set of Python scripts were therefore created to automatically code the frozen source code files. As with the data collection process, individual scripts were developed for every news organization. For each snapshot, the BeautifulSoup library for Python was used to transform the source code into

[7] This is not some special feature of news websites. Well-designed websites will maintain consistent interfaces so that users can quickly locate the desired content based on their developed familiarity.

a navigable object that allowed specific elements to be located based on their attributes (for a similar use, see Sjøvaag et al., 2015; Sjøvaag & Stavelin, 2012).[8] For example, with a navigable object, a script may easily locate a 'div' element with a specific 'id' attribute, and then locate all of the heading elements (e.g., 'h1' and 'h2') that appeared within that 'div' element. For an example of the code necessary to do this, see Figure 1.

Each script was designed to first identify the layout being used in the given snapshot based on predefined sets of structural attributes. For example, if a 'div' element with a 'class' attribute of 'big-story clearfix two' was found in the source code of a snapshot for the *Plain-Dealer*, the procedure for the first layout variation was followed. Alternatively, if a 'div' element with a 'class' attribute of 'big-story clearfix topic-two' was found, the procedure for the second layout variation was followed.

Additionally, in order to ensure that only news items were being coded, rather than links to an author's biography or to another website, the researcher identified distinct, publication-specific patterns in the URLs that would separate news content from non-news content. For example, all of the URLs pertaining to news content on the *Plain-Dealer*'s website either contained the string '/(YYYY)/(MM)/' and concluded with '.html' or contained the string '/news/article/'.

Following the detection of the layout, the script then gathered the information necessary to assess an item's popularity. While there are several ways to operationalize

---

[8] BeautifulSoup is a Python library used to parse HTML documents, creating navigable objects that can be easily extracted and manipulated. For more information, see http://www.crummy.com/software/BeautifulSoup/.

the popularity of a news item (e.g., number of times it has been emailed or tweeted about), it was operationalized here as the number of times it had been viewed. This operationalization was preferred because it is not only consistent with relevant literature but also because the number of times an article has been viewed has been repeatedly found to be the most salient metric in newsrooms (Anderson, 2011a; Groves & Brown, 2011; Usher, 2012, 2013). As with prior work (e.g., Boczkowski et al., 2011; Boczkowski & Peer, 2011; Bright & Nicholls, 2013; Lee et al., 2014), the determination for the relative popularity of an item was derived from the 'most viewed' lists that appeared on the homepages of the organizations analyzed (see Chapter IV).

The script thus identified the region of the page containing the list of 'most viewed' items and extracted all of the relevant links appearing in it in the order they appeared. For example, a selector was written for the *New York Times* snapshots to locate the 'div' element with a 'class' attribute of 'tab-content most-viewed' and then extracted all list elements ('li') appearing within it. Each of those elements was then scanned for the presence of an 'a' child element that had an 'href' attribute that matched the predefined URL pattern for news items.[9] All matching URLs were then temporarily stored as a Python list object for popular items.[10]

The next step was gather the information necessary to code for the prominence of

---

[9] Child elements refer to a sub-unit of another element. For example, in the code, ' Hello World! ', the 'b' (bold) element is a child of the 'p' (paragraph) element. (In contrast, 'p' is the parent element of 'b'.)

[10] Python lists contain a series of objects (e.g., [url1, url2, url3]) that may be accessed by selecting their position within that list. Notably, Python lists differ from Python sets in that the former may contain duplicate items (e.g., two instances of url1) while the latter only contains unique items (e.g., a single instance of url1).

the items appearing on the page. The prominence of an item refers to its relative position on the homepage, with items appearing in more noticeable spots deemed to be more prominent and those appearing in less noticeable spots deemed less prominent (Lim, 2012). Building off the traditions of its analog counterpart (the newspaper's front page), the homepage of an organization is typically designed such that the most important content appears in the most prominent areas of the page (Boczkowski & Peer, 2011). To assign a prominence ranking, the conventions used by Lee et al. (2014) and Boczkowski and colleagues (2011, 2013) were adopted. Specifically, the prominence of a given region was determined by following an F-shape pattern that privileged distinct spots where items could be placed from left to right and then top to bottom (see also Lim, 2010). The present analysis departed from those scholars, however, by privileging areas of the homepage that were clearly intended to draw readers, such as those that included large pictures and larger font sizes. The rationale for this decision is that those stories are clearly distinguished from competing items, thus indicating the intention on the part of the editorial staff to draw greater attention to those stories.[11]

Following this modified F-shape pattern, the five most prominent regions were identified for each variation of the layout for the 21 news organizations (for examples, see Figure 2, Figure 3, Figure 4). Based on the detected layout, the script would thus look for each one of those five regions using predefined selectors and identify the dominant item appearing within it. Through the use of different techniques, only the headlines for

---

[11] This is sometimes made evident in a website's code, with a specific section (e.g., the central spot with a large accompanying picture) assigned an 'id' or 'class' attribute like 'top_story' or 'lead_story'.

the main items appearing within the designated areas of prominence were extracted. For example, with the first layout variation of the *Plain-Dealer*, a selector as written to located the 'div' element with an 'id' attribute of 'main' and all child header elements matching 'h1', 'h2', or 'h3' were then extracted. Each of those elements was then scanned for the presence of an 'a' child element that had a 'href' attribute that matched the predefined URL pattern for news items. This procedure facilitated the exclusion of any 'related items,' since they are typically reserved for stories of lesser import or content from previous days. All matching URLs were then temporarily stored as a Python list object for prominent items.

After creating list objects for the popular items and the prominent items, all of the hyperlinks that matched the URL pattern for news items were extracted from the page and added to a third temporary Python list object. All three list objects were then combined into a single set of unique hyperlinks. Each of those unique hyperlinks was compared against the popularity and prominence list objects and subsequently coded based on if and where it appeared within those list objects. For example, if a link appeared as the second element in the prominence list object, it was assigned a value of 2 for prominence; if it did not appear on that list, it received a value of 0.

Each unique hyperlink was then stored as a separate row in a MySQL database. Although Sjøvaag and Stavelin (2012) advocate for the use of comma-separated values (CSV) files, an ACID-compliant, relational database was preferred because of its ability to handle multiple transactions at once (i.e., have multiple pieces of content be analyzed at the same time), its superior reliability, and its ability to easily filter and access specific

groups of entries.[12] Additionally, although alternative database paradigms like NoSQL and Hadoop were available, MySQL was deemed to offer adequate performance with proper indexing, and provided a well-tested and well-documented system.[13]

Verifying Coded Information

Computers are able to execute a given set of instructions with perfect reliability as a result of their deterministic nature (Grimmer, 2010; Leetaru, 2012). Consequently, there is no need to assess intercoder reliability when using an algorithm to perform the coding, a function that distinguishes computational approaches to content analysis from manual ones (Zamith & Lewis, 2015). However, in order to ensure that the algorithm accurately coded the content, coded data are often compared against a "gold standard," which is typically a human-coded dataset that is presumed to represent the "correct" coding decisions (Grimmer & Stewart, 2013).

Because of the straight-forward and mechanical nature of the adopted approach, the researcher opted to follow an iterative process that involved multiple revisions to the algorithm to ensure that all data were coded properly. Heeding Sjøvaag and Stavelin's

---

[12] ACID (Atomicity, Consistency, Isolation, Durability) refers to a set of properties intended to ensure that database transactions are processed reliably (i.e., no data is lost). Specifically, they ensure that if any part of the transaction fails, the entire transaction fails; that the database is always in a valid state; that each transactions is separate from other concurrent transactions; and that a transaction is permanently stored after it has completed. These properties help define many classical relational database management systems that focus on data integrity (e.g., MySQL, MariaDB), and help to distinguish them from systems that focus on performance (e.g., NoSQL and MongoDB).

[13] Indexing refers to an optimization strategy that stores key information about the entries in a database table, facilitating the look-up of information. In particular, an index enables the database system to quickly locate an item stored in any part of the table without having to iterate through all rows of the data.

(2012) recommendation, all scripts included functions to log failures at every step in the process—that is, to store in a separate database an entry every time the script encountered something it was not programmed to handle. This enabled the researcher to quickly identify problematic snapshots and tune the algorithm to properly deal with them. Additionally, the researcher created an electronic interface that would display the screenshot associated with a given snapshot alongside the respective coding decisions made by the algorithm (see Figure 5). This enabled the researcher to quickly review and confirm that the correct coding decisions were made by the algorithm for randomized subsets of the data.

This iterative process was preferred because it allowed the researcher to quickly identify when and where the algorithm was failing and because it reduced the amount of content that had to be manually reviewed for the researcher to have confidence in the accuracy of the algorithm. That is, if one were to review even just 10% of the snapshots analyzed—that would be 12,647 snapshots—he or she would need to set aside several days if not weeks every time the algorithm was tuned. Instead, the error-logging functionality allowed the researcher to identify and diagnose problems, and the electronic interface allowed the researcher to manually verify that the correct coding decisions were being made, as expected.

After multiple revisions of the algorithm, the researcher ensured that there were no systematic errors being logged and then reviewed 50 random snapshots for each organization—1,050 snapshots in all. Though this was just a fraction of the total number of snapshots, the fact that the variables analyzed were unambiguous and that the

algorithm identified and coded all items as the researcher would have gave the researcher sufficient confidence to proceed with the machine-coded data.

Cleaning Data and Generating New Datasets

A total of 13,077,079 units—individual news items appearing on a given news organization's homepage at a particular point in time—were identified and entered into the database.[14] Each of these records included information like the URL of the item, the publication associated with it, the timestamp of the snapshot both in UTC and adjusting for the organization's native time zone, whether it appeared in a prominent area or was popular, and, if so, the prominence and popularity ranking for that item.

These data were then cleaned through the use of a last set of Python scripts because of specific issues with the way the pages were presented. For example, in what was likely an unintentional error by their programmers, the *Denver Post*'s website routinely included links to a small number of stories that were more than a year old. These stories were hidden from view through a styling attribute but were included in the source code. Additionally, a few links—typically, static objects that happened to fit the URL pattern for news items—were present in all of the snapshots. These entries were removed from the database.

Some news organizations like the *St. Paul Pioneer Press* also used symbolic URL paths, such that the same item would have distinct URLs, though a unique identifier was

[14] A single news item was recorded multiple times if it appeared, as many did, on multiple snapshots of a given news organization's homepage.

shared. For example, a story might have the string ‘ci_26749631’ in its URL, but it would be prefaced by ‘/business/’ in one instance and ‘/popular/’ in another. This could be addressed computationally by reducing the URL to a string consisting only of the unique identifier.

Once the data had been cleaned, a range of new datasets could be quickly created from it. Because all data were stored in a relational database (MySQL), rather than a CSV file, the data could be easily and efficiently filtered and combined. Moreover, the database could be easily queried by Python scripts through the use of libraries like PyMySQL and SQLAlchemy to create more sophisticated datasets for assessing different aspects pertaining to the liquidity of an item and of a page, such as the amount of time each news item spent in an area of prominence, for how long it was popular, and the highest and lowest rankings for an item’s prominence and popularity across snapshots.[15] These data could also be easily transformed and reshaped to fit the expectations of popular structural equation modeling software like MPlus and AMOS in order to assess complex relationships among variables. Finally, the data could be quickly extracted into smaller files for analysis by software like R, which stores data objects in the system’s memory and therefore poses challenges when the size of the dataset exceeds the amount of RAM a machine has.

---

[15] PyMySQL and SQLAlchemy are Python libraries that facilitate the interfacing of Python scripts with MySQL and MariaDB databases. For more information, see https://github.com/PyMySQL/PyMySQL and http://www.sqlalchemy.org, respectively.

Discussion

The liquid quality of online news offers researchers the ability to peer into the 'black box' of journalism and assess how content changes and evolves (Deuze, 2008; Karlsson, 2011; Singer, 2005). However, the prospect of mutable, and in some cases constantly changing, content introduces a range of methodological challenges for researchers (Deuze, 2008; Herring, 2010; Lewis et al., 2013). Indeed, the immediacy afforded by digital and networked technologies allows content producers to constantly produce new work on irregular schedules and update existing work. Furthermore, to meet consumer expectations of constantly-updated content, such content is likely to become ever more liquid, especially as user-generated content becomes more popular.

Different scholars have proposed a number of strategies and techniques for addressing different challenges associated with liquid content, from how to freeze it to how to analyze it (Karlsson & Strömbäck, 2010; Sjøvaag & Stavelin, 2012). The present research has shown how those contributions could be effectively synthesized and built upon to develop a process for computationally capturing and analyzing large volumes of data using inexpensive, consumer-grade hardware.

While Karlsson and Strömbäck (2010) point to different ways of capturing data, from manual approaches like saving the pages with the browser to the use of third-party tools like HTTrack (see also Sjøvaag et al., 2012, 2015), the present work has shown that a superior approach is to use the Selenium framework. That framework allows modern browsers to process JavaScript code and other advanced web features, which are becoming increasingly prevalent as the so-called 'Web 2.0' proliferates and more of the

interactive affordances of the technology are put to use. In particular, JavaScript is now routinely used to make asynchronous calls to the server to load and modify specific content within a page.[16] These may include features that are of great interest to researchers, like the lists of most-viewed items (Boczkowski et al., 2011; Boczkowski & Peer, 2011; Bright & Nicholls, 2013), and, increasingly, core functionality like Facebook's automatic loading of new content when the user approaches the bottom of his or her news feed.[17] In order to accurately freeze pages that make use of this advanced functionality and perform the actions necessary to have the document reach the state of interest for the researcher, it is necessary to adopt more robust solutions—like mimicking a Web surfing session.

Additionally, adopting a technology like Selenium allows researchers to capture full-page screenshots of the content exactly as it appears on the browser screen. This effectively bypasses the limitations aptly noted by Karlsson and Strömbäck (2010) of traditional screenshots with the 'print screen' and 'screen grab' functions of popular operating systems (and the software that automate those functions). Such screenshots can prove to be invaluable to the development and verification of algorithms for coding aesthetic features of documents, and can even serve as the content that human coders could then analyze. Indeed, one could easily envision their use as complementary

---

[16] An asynchronous framework allows a webpage to appear and be interacted with as content is downloaded. This is a contrast to a synchronous framework where subsequent page elements will only load after earlier ones have completed loading or failed to load.

[17] That is, these websites will detect the user's position on a page and append new content to that page when the user reaches a certain point (e.g., the third-to-last story in Facebook's News Feed). This is a contrast to prior standard functionality, wherein the user would click a link (e.g., "Next Page") and an entirely new page would be loaded.

documents in a hybrid form of content analysis (Lewis et al., 2013; Zamith & Lewis, 2015).

However, although the Selenium framework offers several advantages, it must be noted that it brings with it a considerable computational cost. Specifically, it requires a complete browser like Mozilla's Firefox or Google's Chrome to be loaded, which consumes far more RAM and CPU cycles than a simple document mirroring tool like HTTrack. Although the hardware used for this analysis is readily available to any consumer for a moderate cost, simultaneously loading the homepages of the 21 news organizations consumed nearly all of the available CPU and RAM during the brief periods of access—even after shifting the browsing sessions to a virtualized environment. This computational cost thus impairs the scalability of this approach. Researchers seeking to analyze a larger number of websites simultaneously are thus advised to seek out a light-weight solution, with technologies like Node.JS and PhantomJS serving as alternatives worth investigating.

Additionally, the present research lends support to Sjøvaag and Stavelin's (2012) advocacy for the use of selectors to locate and extract content from HTML pages. Python and the BeautifulSoup library were effectively used to perform syntactical analyses in order to identify regions of significance and select the items of interest. Moreover, the uniform nature of the content management systems used by the news organizations meant that only a handful of variations of a website's layout had to be accounted for. This means that researchers only need to write a relatively small amount of computer code to accurately analyze large amounts of data. Moreover, the coupling of Python and

BeautifulSoup was fairly efficient: over 13 million units of data could be extracted and content analyzed using consumer-grade hardware in less than a day. This offers promise to researchers interested in computationally assessing the liquidity of different content, and allows for the analysis of larger amounts of data with greater precision (cf. Lee et al., 2014).

That the coding decisions made by the algorithm were not compared against an independent "gold standard" is a limitation of this research. Indeed, in an ideal world, an amount comparable to the conventions of intercoder reliability under a manual framework for content analysis would yield greater confidence in the accuracy of the algorithm. However, given that the analysis relied on the analysis of recurring syntax—HTML code generated by a content management system that had to be precise in order to render correctly—the efficiency gained through the procedure utilized greatly outweighed the potential loss of accuracy. Specifically, the error-logging mechanisms employed allowed the researcher to quickly identify the different variations of the layouts and, over time, minimize the number of instances in which the algorithms encountered a snapshot they were not programmed to handle. In order to ensure that the instructions were not too broad—resulting in miscoding items—the electronic interface made the confirmation of the algorithmic coding decisions expedient. The positive results from this process offer a pathway for researchers with limited resources.

As this analysis indicated, researchers must be careful to evaluate their data in different ways after the content has been collected and/or coded in order to identify anomalies that may point to broader issues with the source of the data. For example, the

use of symbolic URL paths by the *St. Paul Pioneer Press* meant that a single news item might have been inappropriately treated as two separate news items in an analysis. Rather than recoding all of the items, a simple computer script could be used to automatically clean up that potential issue for all relevant items in the MySQL database. This, among other examples, indicates why in many instances it is preferable to store data directly in a formal database rather than CSV files (cf. Sjøvaag & Stavelin, 2012).

In conclusion, computational techniques can be effectively used to freeze and analyze liquid content like the homepages of news organizations. While the preferable solutions may require some custom programming, there are several free and open-source programs, libraries, and frameworks that can facilitate the creation of powerful scripts. Moreover, such a process may be performed on consumer-grade hardware to analyze large amounts of content. This, in turn, enables researchers to engage in computational social scientific inquiry, and in particular assess the evolution of journalistic content and relationships between key variables both in short intervals and over long periods of time.

CHAPTER IV:
DECIPHERING 'MOST VIEWED' LISTS

Scholars of digital journalism have, in recent years, taken increasing interest in the concept of popularity as it relates to individual news items. This is partly due to the prevalence of audience analytics, which allows data to be captured on a micro scale (e.g., tracking the precise amount of time a particular user spends on a specific page). For example, on an exploratory level, scholars have described the kinds of content that tend to be popular on particular websites, both among readers and among editors (e.g., Boczkowski & Mitchelstein, 2013; Schaudt & Carpenter, 2009). On an explanatory level, scholars have treated popularity as an outcome variable (e.g., the effect of a story's prominence on its popularity, see Lee et al., 2014) and as a predictor variable (e.g., the effect of popularity on an item's likelihood of remaining on a page at a later point in time, see Bright & Nicholls, 2013). Indeed, much of the emerging literature on audience metrics focuses on the concept of popularity as it is understood by practitioners (Graves & Kelly, 2010; MacGregor, 2007), as it pertains to journalism ethics (Hindman, forthcoming; Tandoc & Thomas, 2015), and in terms of how it influences the work of journalism and the valorization of journalistic products (Anderson, 2011a; Groves & Brown, 2011; Usher, 2013).

A common measure of an item's popularity is the number of hits it receives—that is, the number of times that item is accessed (Boczkowski & Peer, 2011; Lee et al., 2014; Tenenboim & Cohen, 2013). Page views are easy to capture given the very nature of networked systems (Andrejevic, 2007; Kaushik, 2009), and news organizations invariably adopt systems like Omniture and Chartbeat to perform that task (Boczkowski

& Mitchelstein, 2013; Graves & Kelly, 2010). Specifically, these data may be recorded directly by the news organization whenever a page is requested from the server by a client (i.e., the reader), and by a third party through the inclusion of a small piece of code on a webpage.[1]

Although page-view data are generally readily available to editors and managers at those organizations, those data are often out of the reach of scholars. That is, such data may only be accessed through an agreement with the news organization, and that access is often limited to periodic reports. Moreover, news organizations may be reluctant to share that data given their potential commercial implications (Couldry & Turow, 2014; 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 homepages of many news organizations. Such lists are often intended to serve as a shortcut for readers to content that has wide appeal, keeping readers on the website for even longer periods of time (Thorson, 2008).

The implications of using such lists as a proxy for popularity have not received a great deal of scholarly attention, however. This is problematic because many studies that use these lists adopt an implicit assumption that they invariably represent similar kinds of data. For example, they 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

---

[1] While it is possible that a "most viewed" list would have been compiled by a human being without regard to any metrics, it would serve as an exceptional case. That is, the available evidence indicates that such lists are exclusively the byproduct of algorithms (Anderson, 2011a; Bright & Nicholls, 2013; Graves & Kelly, 2010), which is consistent with the reports from online editors from the author's ongoing research.

frequency (e.g., every hour). Complicating matters, the websites of news organizations rarely provide sufficient information to ascertain those key considerations. The limitations of scholars' understanding of these lists are especially problematic for comparative research, wherein much of the divergence in the findings could be due to differences in what the data reflects.[2]

The present study adds to the understanding of what these lists represent through a systematic assessment of the 'most viewed' lists of 21 different news organizations. First, a content analysis was performed to assess the information provided about those lists on the websites of the 21 news organizations. Then, data were collected from those lists over two months in order to assess the extent to which the different lists changed and the speed at which news items made it onto the list. In doing so, a contribution is made to the literature by assessing the comparability of the lists of different organizations and by categorizing them into good, intermediate, and poor proxies for what is currently popular on those homepages based on those two dimensions.

Background

The concept of popularity is rarely explicitly defined in the scholarship on digital journalism that uses it, though there does appear to be implicit consensus on what it means and how to measure it. For the purposes of this study, an item's popularity may be

---

[2] Although the term "comparative" is often used to refer to cross-national analyses, it is used here to refer to any research that seeks to contrast multiple news organizations. This would include, for example, the work of Lee et al. (2014), who compared three New York-based news organizations, as well as that of Bright and Nicholls (2013), who compared 5 U.K.-based news organizations.

conceptually defined as the extent to which a large body of news consumers find it to be appealing. Given this broad definition, an item's popularity can therefore be measured in multiple ways. For example, the number of times an item is shared on social media or commented on can serve as a measure of appeal (e.g., Boczkowski & Mitchelstein, 2012). Alternatively, researchers can survey individuals and ask them to identify the items they thought were most appealing (e.g., Lee & Chyi, 2014).

The most common measure of the popularity of a news item, however, is the number of times it was accessed, or the number of page views it receives (e.g., Boczkowski et al., 2011; Boczkowski & Peer, 2011; Bright & Nicholls, 2013; Lee et al., 2014; Tenenboim & Cohen, 2013). On its face, this is a perfectly reasonable measure: if a large number of news consumers clicked on an item, it is likely because it had general appeal on some level.[3] Furthermore, in studies involving newswork and newsworkers, the number of page views an item receives may serve as a superior measure to alternatives like the number of times it is shared because page views have traditionally served as the dominant metric in newsrooms and popularity is generally described in terms of page views by newsworkers (Anderson, 2011a; Groves & Brown, 2011; MacGregor, 2007; Usher, 2012). Thus, at least from the perspective of most newsworkers, the number of page views an item receives is the de facto measure of popularity.

In adopting the number of page views as a measure of an item's popularity, there

---

[3] That appeal may be fleeting or ultimately be unfulfilled. This is often the case with 'clickbait' headlines that appear enticing but lead to unsatisfying content. Regardless of the feeling readers are left with, there was, at one point, sufficient appeal to convince a large number of news consumers to click on an item.

are two common types of data that may be accessed: continuous and ordinal. Continuous data are typically the preferred type because it offers the greatest amount of information. For example, a researcher that has access to continuous data is able to calculate the exact difference in the number of hits between two items (e.g., 535 views vs. 231 views). Such data, however, are difficult to come by. First, news organizations may be reluctant to offer that information because of how important they are for essential commercial functions, such as setting advertising rates (Couldry & Turow, 2014; Napoli, 2011; Turow, 2005). Second, news organizations may find it too challenging to offer access to real-time data in a format that can be readily used by researchers (Graves & Kelly, 2010). Continuous data on page views may be published on the page containing the article or be accessible via an application programming interface (API), though these instances are rare. Given these difficulties, studies that incorporate page views as a measure of popularity rarely use continuous data (for an exception, see Tenenboim & Cohen, 2013, who used exact page view counts published alongside each article by the Israeli website *Walla!*).

Ordinal data, on the other hand, is comparably limited but easier to access. These data typically manifest themselves in the form of a ranking and are therefore effectively compressed. That is, they are able to convey that one news item received more or fewer page views than another, but they do not provide the absolute magnitude of the difference, offering instead equidistant intervals (e.g., most popular and second most popular). These data, however, are readily available on the homepages of many news organizations, through computer-generated lists of the ‘most viewed’ items.

Given the limited access to continuous data, it is perhaps unsurprising that the majority of the research utilizes ordinal data obtained from publicly accessible lists of frequently accessed items. Despite their limitations, those data have been used extensively and to great effect. Boczkowski and Peer (2011) used data from such lists to demonstrate a gap in the preferences of journalists and news consumers when it comes to the subject matter and format of news stories. Boczkowski et al. (2011) also 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. Lee et al. (2014) used data from the 'most viewed' lists of three different U.S.-based news organizations to show that the number of clicks an item received had an effect on its subsequent news placement, but that placement had no effect on the number of clicks an item subsequently received. Bright and Nicholls (2013) similarly used those rankings in an 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 (2013), 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. 7). More often, however, there appears to be an implicit assumption that those lists invariably represent similar kinds 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 (Anderson, 2011a; Graves & Kelly, 2010). Additionally, some organizations may find that their readers are best served by listing the stories that are trending (i.e., recently popular) while others favor listing stories that readers might have missed (i.e., popular over the past day or week). The potential that these data represent different things demands, at minimum, an empirical evaluation of the potential implications of using those lists and the extent to which they may be comparable among oft-studied media like large U.S. newspaper organizations (e.g., Denham, 2014; Luther & Radovic, 2014; Xu, 2013).

Research Questions

The first two research questions focus on the prevalence of 'most viewed' lists

and the number of items that they typically consist of. This is an important consideration because of the implications it has for researchers seeking to utilize that data, both in terms of sampling as well as the range (and thus potential variance) of the data. With this in mind, the following research questions are posed:

RQ1: How prevalent are 'most viewed' lists among large U.S. newspaper organizations?

RQ2: How many items are typically listed on the 'most viewed' lists for the organizations that have those lists?

The third and fourth research questions pertain to the comparability of the data from different lists as well as the likelihood that they serve as a useful proxy for what is currently popular on those websites. This is an important consideration because much of the research in digital journalism that has leveraged data from those lists assumes that they are comparable and immediate. While a few organizations explicitly note the frequency of updates and the period covered by their 'most viewed' list—for example, in a separate page, the *St. Paul Pioneer Press* states that their list is updated hourly and covers popularity over the past hour while the *New York Times* simply states that their list covers the previous 24 hours—several organizations do not. As such, the following research questions are posed:

RQ3: Is the average rate of change for the 'most viewed' lists similar across the large U.S. newspaper organizations that have them?

RQ4: Is the amount of time it takes a news item to appear on the 'most viewed' list similar across the large U.S. newspaper organizations that have them?

Method

This study was conducted in two stages. The first two research questions were addressed in the first stage, wherein the author conducted a content analysis of the homepages of the top 50 print news organizations in the United States, based on the weekday print circulation figures reported by the Alliance for Audited Media (AAM) on September 26, 2014 (see Table 1). These organizations were selected because they are among the organizations most often studied by scholars in the field of journalism studies (e.g., Denham, 2014; Luther & Radovic, 2014; Xu, 2013) and because they were part of a larger research project (see Chapter V). This analysis was conducted on September 29, 2014 and focused on two variables. The first variable was whether the homepage contained a list of most viewed items, coded in a binary fashion (present and not present). These lists could manifest themselves under varying titles, such as 'most viewed,' 'most popular,' 'most read,', and 'most clicked,' with the lone requirements being that they listed items that seemingly received the greatest number of page views and that they appeared somewhere on the homepage. The second variable was the number of items that appeared on that list. In some instances, the 'most viewed' list was abbreviated, and linked to the complete list of items. These links were followed in order to identify the maximal number of items that a researcher might have access to. Because of the simple nature of the variables and the mechanical nature of the analysis, a formal assessment of intercoder reliability was not conducted. Instead, the analysis was replicated by the researcher two days later.

Then, in the second stage, the author conducted a computational content analysis of the organizations that had a 'most viewed' list consisting of five or more items in order to evaluate the third and fourth research questions, which focused on the comparability of 'most viewed' lists and their suitability as proxies for what is currently popular on those websites. The five-item threshold was put in place because researchers seeking to use such data in a comparative sense are likely to need at least five data points from those lists, and because of the requirements of the larger research project associated with this study.

Twenty-one news organizations were analyzed in this second stage. Data collection, which began on October 18, 2014, and lasted until December 20, 2014, involved the use of computer scripts, developed by the author, to simulate a browsing session and systematically download the browser-interpreted source code of each news organization's homepage every 15 minutes. Specifically, these scripts made use of the Python programming language and the Selenium framework to simulate multiple Mozilla Firefox sessions and store the page's source code—after all elements had been loaded—in an organized fashion. 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, all data collected on November 3, 4, and 5 were discarded.

Additional Python scripts were then developed for each news organization in order to analyze those source code files and computationally identify and extract the elements in the 'most viewed' list and in other regions of the page using the

BeautifulSoup library. When multiple lists were accessible, the one with the shortest interval was selected (i.e., 'past hour' was selected over 'past day') in order to better categorize them as proxies for what is currently popular on those sites. In order to ensure that only news items were being collected, and not links to static pages like section fronts, a URL pattern for news items was identified for each organization and only those items fitting the pattern were recorded. 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 was subsequently used to manually verify the final algorithms' coding decisions for 1,050 lists. These data were then cleaned and entered into a relational database. A final computer script then used those data to calculate a series of variables for each unique news item, such as the amount of time that item spent on the page before appearing on a 'most viewed' list. For more information on this process, see Chapter III.

Results

The Prevalence and Length of 'Most Viewed' Lists

The first research question focused on the prevalence of 'most viewed' lists among the organizations studied. As shown in Table 1, 27 of the 50 largest U.S. print news organizations (54.0%) listed their most viewed items on the homepage. There was a positive correlation between the weekday circulation of the organization's print product and the presence of a 'most viewed' list, though it was weak ($r_s = 0.25$). This indicates that, among large news organizations, there is little relationship between an

organization's size and whether or not it has such lists on its homepage. Indeed, there were notable exclusions among the 10 largest organizations, such as the *Los Angeles Times*, the *New York Post*, and *Newsday*. Notably, none of the news organizations owned by Tribune Publishing Company listed their most popular items on their homepage. Similarly, only one Gannett Company-owned organization (*USA Today*) did so, though it listed fewer than five items.

The second research question inquired about the length of the 'most viewed' lists for the organizations that had them. The mean amount of spots was 13.2, with a lower median of 10 items. The smallest amount was four items (by *USA Today* and the *Chicago Sun-Times*) and the largest amount was 50 items (by *The Denver Post*, the *St. Paul Pioneer Press*, and the *San Jose Mercury News*). There was a negative correlation between the weekday circulation of the organization's print product and the length of its 'most viewed' list, though it was small (r = -0.13) and is skewed by the three 50-item organizations, only one of which appears in the top 25. Nevertheless, there appears to be no clear link between an organization's size and the length of its 'most viewed' list among large news organizations.

The Comparability of 'Most Viewed' Lists

The third and fourth research questions focused on the extent to which data obtained from different organizations are comparable and, more broadly, the likelihood that 'most viewed' lists represent what is currently popular on a homepage (as opposed to what has been popular over longer periods of time, such as the past day or week). This

was evaluated across two dimensions that indicate the frequency of updates and the period of time the list likely covers: the rate at which the list changed and the median time it took a news item to appear on the list, relative to its first appearance on the homepage. In order to ensure consistency in the comparison and to be able to evaluate a large number of organizations, the 'most viewed' list was reduced to the top five items. Organizations that had fewer than five items on their 'most viewed' list were therefore excluded from the analysis. Additionally, six news organizations (the *Arkansas Democrat Gazette*, the *Buffalo News*, the *Las Vegas Review Journal*, the *Pittsburgh Post-Gazette*, the *Sacramento Bee*, and *U-T San Diego*) were excluded because of resource constraints and because they did not fit into a larger project using those data (see Chapter V). A total of 115,533 'most viewed' lists containing 17,541 distinct news items from 21 news organizations were analyzed over 61 days.

To assess the third research question, which inquired about the rate at which the 'most viewed' lists changed, a value was calculated to reflect the proportion of items that appeared on a given list at Time (t) that had changed by Time (t+1). Change could 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 is 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. Since the length of each list was always five items, that equation can be simplified to $\frac{5-I}{5}$. Thus, if two items on the 'most viewed' list 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 if two items swapped positions between Time (t) and Time (t+1)—then the rate of change would be 0.4, or 40%. A one-hour interval was utilized as it would allow sufficient time for the servers to update their data.

As shown in Figure 6, all but one of the organizations updated their 'most viewed' list at least once an hour on average. The lone organization that did not do this was the *St. Louis Post-Dispatch*, which appeared to update its list every other hour. Notably, there were a few points in time where there was no activity for some of the organizations (e.g., the *Register*), typically occurring during the 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 nightly reports were compiled.

However, these data also show a considerable amount of variation in the rates of change for the different news organizations. The highest rates of change were for *The Denver Post* (68.7%), the *Plain-Dealer* (65.5%), and the *Oregonian* (63.3%). For these 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 lowest rates of change were for the *Kansas City Star* (11.4%), the *Miami Herald* (11.4%), and the *Seattle Times* (12.8%). For these 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., the past hour or the past 24 hours).

To assess the fourth research question, the time stamp of an item's first appearance on the homepage was compared against the time stamp of that item's first appearance on the 'most viewed' list. While it is possible for a popular item to appear in another part of the website first, this was rarely ever the case in these data as fewer than 0.1% of the news items appeared on the 'most viewed' list before appearing elsewhere on the homepage.

As shown in Figure 7, there were notable differences in the median amount of time it took the average news item to appear on the 'most viewed' list for the different news organizations. 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 'most viewed' list (for those items that appeared on the 'most viewed' list). In contrast, it took, on average, 19 hours for an item to appear on the *Miami Herald*'s 'most viewed' list, 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 (e.g., Facebook) or through links from aggregators and blogs (e.g., Google News), 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 'most viewed' list was just over three hours. Nevertheless, organizations that have high median times, like the *Miami Herald*, are unlikely to have lists covering activity over the previous hour. The *New York Times*, for example, explicitly noted on its website that its list covers the previous 24 hours, which is consistent with its high median time.

Discussion

This study aimed to offer a better understanding of the data derived from the lists of most viewed items on the homepages of large news organizations, from their availability to their comparability. In short, it was found that those data are only available for roughly half of the 50 largest U.S. newspaper organizations; that data are typically offered for at least the five most popular items, and in many cases the top 10; that the data can be gathered in, at minimum, hourly intervals; and that the data are not readily comparable across the entire range of organizations when it comes to two dimensions: the rate at which the 'most viewed' lists change and the median time it takes an article to reach that list. Therefore, the central and overarching finding of this study is that data from 'most viewed' lists have clear limitations that must be explicitly noted and that it should not be taken for granted that the data are comparable.

These findings, it must be noted, should not automatically cast doubt on prior work that has made use of 'most viewed' lists. 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 subject matter and the 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 findings should hold up if one accepts the assumption that organizations follow regularized (i.e., routine-driven) patterns that would make their coverage fairly consistent over a long period of time. However, the findings of studies like that of Lee et al. (2014) that require strict parameters (e.g., assessing relationships over short periods of time) could be colored to a substantial degree by differences in what the data represents.

Instead, the findings from this study point to the importance of being clear about the limitations of these data and the need to evaluate them to ensure that they are comparable along at least some empirical dimensions. As guidance to future researchers, the 21 organizations analyzed in the second stage were grouped into four clusters based on where they aligned across the two dimensions analyzed in this study. The organizations in these clusters, shown in Figure 8, should be comparable with other organizations within their cluster based on the rates of change of their 'most viewed' lists 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 analysis, 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 (six hours) to appear on the 'most viewed' list, then that list was considered to have a high median time. These thresholds were also developed while being mindful of the explicit information by organizations like the *New York Times* and the *St. Paul Pioneer Press*. While it might be sensible to solicit information about exactly what data are represented by those lists directly from a news organization, this often yields, in the author's experience, unreliable information.[4]

Based on this classification procedure, and as shown in Figure 8, the 'most viewed' lists for the *Oregonian*, the *Plain-Dealer*, the *Salt Lake Tribune*, the *San Jose Mercury News*, and the *St. Paul Pioneer Press*, the *Star Tribune*, *The Denver Post*, *The Star-Ledger* comprise one cluster. This cluster represents 'most viewed' lists that are most likely to reflect what is currently popular on the website based on their high rate of change and 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. These two clusters have 'most viewed' lists that may or may not reflect what is currently popular on their websites based on their combination of either a low rate of change and low median time or high rate of change and high median time. Finally, the *Honolulu Star-Advertiser*, *Houston Chronicle*, *Kansas*

---

[4] As part of an ongoing study, the author has found a considerable amount of misunderstanding among online editors regarding the 'most viewed' list published on their organization's homepage. Specifically, when two parties in the same newsroom were asked to describe the frequency of updates and the period of time covered by that list, conflicting answers were repeatedly given. For at least one organization, this uncertainty appeared to extend to the editors' immediate superiors as well.

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 high median time. It must be noted, however, 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.

The findings from this study also point to the sampling bias that arises when using data from lists of most-viewed items. Just over half of the 50 news organizations had such lists, and there were several systematic omissions, including all of the Tribune Publishing Company properties and nearly all of the Gannett Company papers. 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 news organizations) as well as smaller news organizations, like community newspapers, which remain largely understudied. Additionally, scholars should consider other measures that may be used to empirically assess the phenomena captured by lists of most-viewed items as well as their comparability. While the present work has offered both a starting point and guidance for researchers in the area of digital journalism, there are likely to be other worthwhile measures to consider.

In conclusion, while continuous data on page views—the precise number of 'hits' a story receives—is generally preferable, ordinal data obtained from lists of most-viewed items can be a suitable alternative. However, when using such data, researchers must

recognize their limitations and be transparent about them, from the sampling biases they introduce to the information that is lost when working with relative values. Moreover, researchers should avoid assuming that these lists are automatically comparable across organizations just because they look similar. Ultimately, this study serves as a reminder of the need to view data and data sources with a critical eye.

CHAPTER V:
THE EFFECTS OF POPULARITY ON PROMINENCE

Central to gatekeeping theory, in the context of journalism, is the notion that editors serve as central nodes in the process of creating news products (White, 1950). In manning their gates, these individuals decide not only what information gets through but also what it looks like once it has passed that gate. For decades, conceptualizations of the gatekeeping process largely minimized the role of audiences, often leaving them out of models. In recent years, however, scholars have asserted that audiences have a greater role in that process, leading to the inclusion of an audience channel in Shoemaker and Vos' (2009) revision of the gatekeeping model. This should not be viewed as a natural development, as it forces newsworkers to reconcile values and beliefs like the professional authority of journalists and the need to be insulated from non-editorial considerations (Deuze, 2005; Lewis, 2012) with an increasingly active audience (Napoli, 2011) and pressures to make use of the readily available information about their preferences (Groves & Brown, 2011; Usher, 2012).

As noted in Chapter II, the study of audience analytics and metrics in the context of journalism is not itself novel, and a growing number of scholars have taken an interest in this area in recent years (e.g., Nguyen, 2013; Tandoc & Thomas, 2015; Tandoc, 2014a, 2014b; Usher, 2013; Vu, 2014). This research has pointed to the growing influence of audience analytics in newsrooms, with editors reporting that they pay extensive attention to metrics and journalists sometimes believing that the value of their work is tied to the number of clicks it receives (Anderson, 2011a; Groves & Brown, 2011; Usher, 2012). The balance of evidence indicates that news practices are changing in response to this

phenomenon, with many scholars assuming that news products themselves are consequently changing as well.

There has been, however, little empirical work looking at how the content of news products is changing, which would be necessary to support that contention (Petre, 2015). Put differently, content-level effects—that is, the impact that audience metrics are having on the news products themselves—remain under-studied. Instead, much of what is known is derived from self-reported information gleaned from surveys and interviews, which represent only the extent to which the newsworker believes her or she is using the data, and could be subject to over-reporting and under-reporting as a consequence of social desirability biases (Kreuter et al., 2008), especially considering the culturally charged nature of audience analytics and audience metrics. Moreover, in order to further the understanding of this phenomenon, there is a need for scholars to further delineate the effects of these technologies on particular editorial behaviors, such as the placement of news content in particular areas of the homepage.

The present research adds to the growing body of work on audience analytics and metrics by assessing the amount of overlap between the editorial and audience agendas as well as the effect a news item's popularity has on its subsequent prominence on the homepage and on the likelihood that it will remain in a prominent area of the homepage at a later point in time. In particular, this study builds on the scholarship of Boczkowski and Peer (2011), Bright and Nicholls (2013), and Lee et al. (2014) by looking at a larger and more heterogeneous set of organizations over a longer period of time, and by using a novel computational approach. In doing so, it offers insight into the extent to which

gatekeeping may be changing and evaluates the prospect of what has been called a turn toward an "agenda of the audience" (Anderson, 2011b, p. 529).

Background

Journalism as Profession, Ideology, and Logic

As scholars have observed, journalism, as practiced in the United States, is not a classical profession (Kaplan, 2006; Schudson, 1978). According to Freidson (2001, p. 12), "professionalism may be said to exist when an organized occupation gains the power to determine who is qualified to perform a defined set of tasks, to prevent all others from performing that work, and to control the criteria by which to evaluate performance."[1] By this definition, journalism, as practiced in the United States, is not a profession (Lewis, 2012). For example, there is no formal mechanism for the inclusion or exclusion of would-be 'journalists.' That is, not only is there is no licensure of journalists in the United States but any attempt to monopolize the legitimation of practitioners (e.g., through certification) would likely be construed as an attack on individuals' freedom of expression (Witschge & Nygren, 2009). Furthermore, while there has long been a code of ethics in place—the Society of Professional Journalists adopted its first code of ethics in

---

[1] Greenwood (1957) defines a profession more broadly, arguing that it must possess five attributes: (1) a systematic body of knowledge; (2) professional authority and credibility; (3) regulation and control of members; (4) a professional code of ethics; and (5) a culture of values, norms, and symbols. While some of these traits are applicable to journalism in the United States (e.g., the existence of a professional code of ethics and a strong culture), others are less so (e.g., the existence of a systematic body of knowledge is disputed and there has been a marked decline in the public's trust in journalists). However, as argued here and elsewhere, the regulation and control of members—the central feature noted by Freidson (2001)—is a trait that does not apply to journalism in the United States.

1926—to enable self-regulation, adherence to those guidelines is voluntary and there is minimal power to enforce penalties against those who violate them (Kaplan, 2006).

While scholars may consider journalism to be a "semi-profession" (Witschge & Nygren, 2009, p. 39) at best, there is little question that among its practitioners in the United States, journalism is widely regarded as being more than an ordinary occupation (Weaver et al., 2007). This is because journalism in the United States is marked by a strong and shared set of values and role conceptions that is derived from a robust occupational ideology (Golding & Elliott, 1979; Schlesinger, 1978). Occupational ideology, in the context of journalism, may be defined as the "dominant way in which news people validate and give meaning to their work" (Deuze, 2005, p. 446).

Deuze (2005) argues that the occupational ideology of journalism consists of five central values: (1) journalists should provide a public service; (2) journalists should be impartial, fair, and objective; (3) journalists must be autonomous and independent in their work; (4) journalists must have a sense of immediacy and the ability to be expedient in their reporting; and (5) journalists must have a strong sense of ethics that is consistent with professional codes. Lewis (2012) adds that beyond occupational ideology, there is a professional logic among U.S. journalists that involves assumptions about the role of the journalist in society: "They take for granted the idea that society needs them as journalists—and journalists alone—to fulfill the functions of watchdog publishing, truth-telling, independence, timeliness, and ethical adherence in the context of news and public affairs" (p. 845).

The nature of journalism's ideology, professional logic, and even whether it is a

'true' profession or not, is not merely an academic affair. Indeed, these elements not only define a shared sense of purpose but further enable journalists to make jurisdictional claims in contested spaces—that is, to claim an exclusive right to perform a certain task for society (Abbott, 1988). These claims are common and often embedded in journalists' discursive practices. For example, journalists often draw upon their values and professionalism to claim and self-legitimize a prominent position in society as a central actor in ensuring a properly functioning democracy (Deuze, 2005; Schudson, 1978) and as the primary sense-maker of current events (Coddington, 2013; Singer, 1997, 1998). Indeed, the value of objectivity was largely adopted by American journalists in order to gain social authority, namely by enabling them to claim that their work was value-free, credible, and a representation of 'truth' (Schudson, 2001). The struggle for jurisdiction is especially relevant in the present media environment, with the decentralization of the production and distribution of media—facilitated by both technological and sociocultural shifts (Benkler, 2006; Jenkins, 2006; Shirky, 2008)—and the rise of new classes of media actors (e.g., bloggers) blurring the boundaries of who is a journalist and what journalism is (Shirky, 2008; Witschge & Nygren, 2009). This struggle raises important questions, for example, about the sorts of legal benefits and protections that journalists should receive and what the threshold should be to entitle one to those benefits (Gant, 2007).

However, occupational ideologies and professional logics are dynamic: they change over time as some ideas and values become marginalized and others codified (Deuze, 2007). For example, the aforementioned occupational value of impartiality and objectivity as we know it today only became widely adopted in 1920s (Schudson, 2001).

Given the notable shifts in the industry, both technological and cultural, it is possible that such values and beliefs, and how they should be enacted, may also be shifting.

Gatekeeping Theory

For much of the 20th century, journalism was based on a model of scarcity and exclusivity, with media organizations gaining social and economic power through their domination of the means of production and distribution of news (Bruns, 2005; Shirky, 2008). According to Lewis (2012), the notion of professional control over content is central to the professional logic of journalism: "professional journalists derive much of their sense of purpose and prestige through their control of information in their normative roles" (p. 845). As such, the work of gatekeeping—or the act of deciding what elements should be included or excluded—has long been key to the work of journalism. As Boczkowski (2004, pp. 206–207) notes: "All occupations and professions have certain traits that make them stand apart as a distinctive domain of activity. For modern journalism, one such trait is the notion of gatekeeping."

According to Shoemaker and Vos (2009, p. 22), "the basic premise of gatekeeping scholarship is that messages are created from information about events that has passed through a series of gates and has been changed in the process." Gatekeeping, therefore, involves both the selection of what passes and does not pass a gate as well as the shaping of the item as it passes each gate (Shoemaker, 1991). For example, a homepage editor may not only decide which news items make it onto the homepage, but alter the headline, blurb, or art associated with that item in the process of putting it on the

page. The ability to make decisions like these gives gatekeepers considerable power (Breed, 1955).

The origins of gatekeeping may be traced back to the work of Lewin (1947), though its first application in journalism may be found in White's (1950) study of 'Mr. Gates.' During his week-long study of the selection choices of a wire editor at a mid-sized Midwestern morning newspaper, White found that more than 90 percent of the wire stories received were not used and that the editor's decisions were "highly subjective" (p. 386). Indeed, much of the time, Mr. Gates rejected the stories based on his personal assessment of the story's merit—that is, the perceived newsworthiness of the item—as well as his personal values and taste (e.g., favoring interpretive stories over those filled with statistics).

Gatekeeping theory has been applied in a variety of contexts, and the general model has received several updates in recent decades to account for the different roles of various actors and multiple sources of influence (key works include Gieber, 1956; McNelly, 1959; Shoemaker & Vos, 2009; Shoemaker, 1991; Snider, 1967; Westley & MacLean, 1957). In the majority of these models, however, the constructed audience—the individuals a journalist thinks about when he or she thinks of his or her audience (DeWerth-Pallmeyer, 1997)—received limited attention as a potential source of influence in the gatekeeping process. Although the models of gatekeeping presented by Westley and MacLean (1957) and Shoemaker (1991) account for the potential for audience feedback, the audience is given a minor role in the process. This is perhaps unsurprising given the traditional widespread skepticism toward (if not outright rejection of) formal

audience research on the part of journalists and editors (DeWerth-Pallmeyer, 1997; Gans, 1979; Tuchman, 1978) and the limited avenues for engagement available to members of the actual audience during that time (Deuze, 2001).

More recently, Shoemaker and Vos (2009) proposed a new model that accords the audience a more prominent role in the gatekeeping process. In contrast to earlier models, the audience is considered to be a channel in this model, pointing to their increasingly active role. Specifically, this revised model contains three channels: a source channel, a media channel, and an audience channel. The source channel is comprised of non-media sources (e.g., an airplane crash survivor) who have information about an event (e.g., an airplane crash). These individuals, for example, may choose to withhold information from journalists or simply forget pieces of what they observed. The media channel is comprised of newsworkers (e.g., journalists) who may witness events first-hand or learn about them from sources (e.g., a press release). These individuals may, among other actions, choose to ignore an event or focus on a specific angle. Finally, the audience channel is comprised of the individuals who consume and redistribute content (e.g., tweet about an article or share it via e-mail), which extends the reach of the news product and offers a cue about its appeal.

This latest major revision of the model both implicitly and explicitly calls attention to the notion of "audience gatekeeping" (Shoemaker, Johnson, Seo, & Wang, 2010, p. 61) and ideas of news driven by the "agenda of the audience" (Anderson, 2011b, p. 529). In particular, it points to the growing perception of the influence of non-purposive forms of audience feedback gathered by sophisticated audience information

systems (audience analytics) and distilled into real-time, quantified measures of user behaviors (audience metrics). That is, the proliferation of these systems and measures within newsrooms has led some scholars argue that it effectively grants the audience a greater role in the gatekeeping process than in decades past (Lee et al., 2014; Shoemaker & Vos, 2009; Vu, 2014). However, the scope and extent of the role played by the audience in the gatekeeping process remains contested.

Content-Level Effects

Despite the perceived growing import of metrics in newsrooms, few studies have examined the relationship between metrics and content. Instead, most of what scholars know about that relationship is based on interviews, surveys, and ethnographies that focus on how newswork is performed. This work has indicated that newsworkers are becoming increasingly sensitive to audience metrics, with major editorial decisions sometimes being influenced by the number of times items are viewed (Groves & Brown, 2011; Usher, 2012; Vu, 2014). However, the extent of the use of metrics and the kinds of editorial activities informed by them appears to vary across newsrooms (Anderson, 2011a, 2013). These studies can only speculate about the effects metrics have on the shape news content takes, however. Furthermore, those speculations often diverge from the insights of Boczkowski and colleagues, who have repeatedly found sizable thematic gaps between the content that editors and readers find to be most important (Boczkowski et al., 2011; Boczkowski & Mitchelstein, 2013; Boczkowski & Peer, 2011). This would indicate a more muted or sophisticated use of audience metrics, and ease fears among

scholars and practitioners alike of an overreliance on audience data (Nguyen, 2013, p. 529; Tandoc & Thomas, 2015).

In the modest stream of work looking at content-level effects of audience metrics, two studies stand out as key contributions. The first study, by Lee and colleagues (2014), looked at the time-lagged effect of the popularity of a news story on its prominence on a news organization's homepage, and vice versa. Specifically, they analyzed the websites of three New York-based news organizations—the *Daily News*, the *New York Post*, and the *New York Times*—over a two-week period, collecting data at four different times of the day: 9 a.m., 12 p.m., 3 p.m., and 6 p.m. They found that a story's popularity (measured by its ranking on a list of most-viewed items) affected subsequent news placement; that the strength of that effect intensified over the course of the day; and that there was a stronger effect of story popularity on placement than of placement on popularity. Notably, however, when their findings were disaggregated, there was a positive effect for one website (nytimes.com), no effect for a second website (nypost.com), and a negative effect for a third website (nydailynews.com). Put differently, the *New York Times* rearranged its homepage in response to a story's popularity by promoting popular stories whereas the *Daily News* did the opposite, ostensibly to make room for content that had not yet been widely consumed. However, the effects found by the authors were quite small. For example, the overall effect of a one-rank increase in popularity was a -0.15-rank decrease in prominence.

An alternative approach to the study of the effect of audience metrics on content was adopted by Bright and Nicholls (2013). They analyzed a month's worth of data from

five news organizations in the United Kingdom—the *BBC*, the *Daily Mail*, the *Daily Telegraph*, the *Guardian*, and the *Mirror*—to assess the impact that being on a most-read list had on an article's likelihood of appearing somewhere on the homepage 15 minutes later. They found that articles appearing on a most-read list had a lower (26 percent less) risk of being removed from the homepage than articles that did not; that this effect occurred, with little difference, for both 'soft' and 'hard' forms of news; and that the effect was more extensive for the 'quality' publications (e.g., the *Guardian*) than the 'tabloid' ones (e.g., the *Daily Mail*).

The present research adds to that body of work by focusing on a larger set of more heterogeneous, U.S.-based news organizations over a longer period of time and by using a novel computational approach. In particular, it addresses a shortcoming in the existing literature by examining both large and mid-size newspaper organizations, and by assessing different indicators of the evolution of the audience and editorial agendas, and the extent to which the editorial agenda may be affected by the audience agenda, within a single study, thereby increasing confidence in the comparisons. Like the works of Lee et al. (2014) and Bright and Nicholls (2013), this study focuses on short-term, stimulus-response-type effects through the use of a time-lagged research design. Specifically, it focuses on a news item's popularity and its prominence, and evaluates the influence the former may exert on the latter at a subsequent point in time across two dimensions: changes in its position among the most prominent areas of the homepage and the likelihood it will remain in one of those prominent areas at a later point in time. Additionally, it evaluates the so-called editorial and audience agendas, and the extent to

which they converge (see also Boczkowski et al., 2011; Boczkowski & Mitchelstein, 2013; Boczkowski & Peer, 2011).

Research Questions and Hypotheses

Given the growing role of the audience in the gatekeeping process (Shoemaker & Vos, 2009) as well as reports of the growing salience of metrics in the newsroom (Anderson, 2011a; Groves & Brown, 2011; MacGregor, 2007; Tandoc, 2014a; Usher, 2012, 2013) and how it is perceived to influence decisions relating to the homepage (Tandoc, 2014b; Vu, 2014), one would expect that popular items would also tend to be prominent—that is, that there would be extensive overlap in the items that became popular and those that were placed in prominent areas of the homepage. However, an alternative perspective rooted in the traditional view of the gatekeeping process (Shoemaker, 1991)—which affords the audience a limited role in the process—and building on the professional logic (Lewis, 2012) and occupational ideology (Deuze, 2005) of journalism—which emphasize the need for journalists to play a central role in the production of content and separated from external influences—would contend that there should be limited overlap, especially in light of the considerable differences in the preferences of editors and readers (Lee & Chyi, 2014).

The empirical work of assessing news content appears to primarily lend support to the latter theoretical perspective. For example, Boczkowski and Peer (2011) found thematic gaps in the types of content (e.g., public affairs vs. non-public affairs) that were as low as 13% and as high as 51% in their analysis of *CNN*, the *Chicago Tribune*, the

Seattle Post-Intelligencer, and *Yahoo! News*. Similarly, Boczkowski et al. (2011) found thematic gaps as low as 8.7% and as high as 30.3% in their analyses of nine Latin American and Western European publications. The U.K.-focused work of Bright and Nicholls (2013) found that, on average, 12% of items that appeared *anywhere* on the homepage became popular at some point.[2] Drawing on these empirical findings, the following hypothesis is posited:

H1: A minority of prominent news items will be popular at some point in time.

For an item's popularity to exert any influence on its prominence, it must appear on the 'most viewed' list prior to its final appearance in an area of prominence. Put differently, if an item is removed from an area of prominence before it becomes popular, then its popularity simply cannot exert any effect on prominence. It is therefore important to assess the proportion of items that appeared on the 'most viewed' list before they disappeared from an area of prominence—that is, to identify the proportion of items for which popularity has the capacity to exert a practically meaningful effect on prominence. Unfortunately, the work of Boczkowski and colleagues (Boczkowski et al., 2011; Boczkowski & Peer, 2011) does not evaluate temporal relationships, and the work of Lee and colleagues (2014) and Bright and Nicholls (2013) offers little guidance in this regard. In the two latter cases, effects were found for an item's popularity on its subsequent

---

[2] An important methodological consideration of that study is that the different sites had various numbers of slots on the 'most read' list—their proxy for establishing popularity. *The Guardian*, for example, only had five slots whereas the *Daily Mail* had "up to" 20. As such, there was less opportunity for an item to be considered popular for *The Guardian* as it would need to garner more hits (relative to other *Guardian* stories) when compared to the *Daily Mail*.

prominence and visibility, but no information was given about the proportion of items that were popular before they were prominent. This is important because it impacts the consequence of those effects. That is, the effect is likely to be of limited consequence if it only applied to a relatively small number of items. In light of this, the following research question is posed:

RQ1: What proportion of prominent news items become popular before they are removed from an area of prominence?

One way to assess whether an item's popularity influences its subsequent prominence is to assess whether popular items have a lower risk of being removed from the prominent areas of the homepage at a later point in time relative to their non-popular items. Another way is to assess whether changes in popularity lead to subsequent changes in the prominence accorded to that item. In drawing from the literature, competing perspectives may be considered with regard to these effects. It may be reasoned that online editors will seek to ensure that content that is in high demand will remain in prominent areas and perhaps even be made more prominent in order to make it clear to news consumers that the organization has content relating to the topic or event that is of demonstrable interest. However, it is also possible that those individuals would reason that popular content will already have been seen by a large portion of readers—this may have been what made content more popular to begin with—and that such content may be removed from prominent areas or demoted to a less-prominent spot in order to make

space for new content.[3] It may also be reasoned that there should be *no* effect or minimal effects as online editors reject making short-term decisions based on information gleaned from audience metrics, relying instead primarily on 'gut' feelings or other considerations that are unrelated to audience preferences. In light of this theoretical divergence, the empirical findings from the modest body of work on metrics-related, content-level effects (Bright & Nicholls, 2013; Lee et al., 2014) guide the following research question and hypotheses:

H2: Items that are popular will have a lower risk of being removed from the prominent regions of the homepage at a later point in time than non-popular items.

RQ2: Among popular items, do more-popular items have a lower risk of being removed from the prominent regions of the homepage at a later point in time than less-popular items?

H3: An increase in an item's popularity will lead to a decrease in its prominence at a subsequent point in time.

Method

Sampling

[3] Although these rationales are introduced here in an oppositional fashion, it is theoretically possible that one rationale may be employed by certain online editors or for certain kinds of content, and another by different online editors or for different kinds of content. In such instances, minimal effects would be found within the organization due to the aggregative nature of the analysis. It is assumed in this study that there is a dominant and mostly uniform rationale that would develop through socialization and routinization.

To address the aforementioned questions and hypotheses, 14 news organizations were analyzed (see Table 2). These organizations were selected from a list of the 50 largest U.S. newspaper-producing companies, based on their circulation, compiled by the *Alliance for Audited Media* on September 26, 2014. The inclusion criteria were that the organization had to have a website with at least five distinct spots for content to appear in prominent areas of the homepage as well as a 'most viewed' list that served as a useful proxy for the current popularity of news items. This sampling frame was chosen because it represented organizations that are not only influential but are also likely to have the greatest amount of tension between traditional journalistic values and growing pressures to use audience metrics to inform editorial decision-making due to their comparably-high adherence to those values and the economic uncertainty faced by the newspaper industry (Soloski, 2013; Weaver et al., 2007). The selection criteria were also adopted to ensure comparability in the sources of data for the key variables of prominence and popularity. Specifically, the comparability of the 'most viewed' lists was assessed through a two-month analysis of the lists across two dimensions: the rate at which the lists changed over the course of the day and the median time it took a news item to appear on the list, relative to its first appearance on the homepage. For more information on this process, see Chapter IV.

Key Variables

The unit of analysis in this study was the individual news item, appearing on the homepage of a given news organization at a given time. A news item was operationalized

as a package that includes a headline and a hyperlink (Boczkowski et al., 2011; Boczkowski & Peer, 2011; Lim, 2012). The present study focused on two key sets of variables that deserve special attention: popularity and prominence.

Popularity. While there are several ways to operationalize the popularity of a news item (e.g., number of times it has been emailed or tweeted about), it was operationalized in this study as the number of times it has been viewed, which indicates its appeal. This operationalization was preferred because it is not only consistent with the relevant literature (e.g., Boczkowski & Peer, 2011; Bright & Nicholls, 2013; Lee et al., 2014) but also because page views has been repeatedly found to be the most salient metric in newsrooms (Anderson, 2011a; Groves & Brown, 2011; MacGregor, 2007; Usher, 2012, 2013). Like Bright and Nicholls (2013) and Lee et al. (2014), information on popularity was obtained through the proxy of a news item's presence and ranking on a news organization's list of most-viewed items, which itself is automatically generated using server data on page views (see also Boczkowski et al., 2011; Boczkowski & Mitchelstein, 2013; Boczkowski & Peer, 2011).

Because different news organizations offer varying numbers of most-read items through their public-facing lists, only the top five items were considered in this analysis in order to ensure the comparability of the lists. Popularity was assessed as a dichotomous variable (popular if appearing on the list, non-popular otherwise) for some of the research questions and as an ordinal variable for others. When treated as an ordinal variable, items were reverse coded so that the most popular item was assigned the value 5 and the least popular item the value 1, with items that did not appear on the list receiving

a value of 0 (for a similar application, see Lee et al., 2014).

Prominence. The prominence of an item refers to its relative position on the homepage, with items appearing in more noticeable spots deemed to be more prominent and those appearing in less noticeable spots deemed less prominent. Building off the traditions of its analog counterpart (the newspaper's front page), the homepage of a news organization is typically designed such that the most important content appears in the most prominent areas of the page (Boczkowski & Peer, 2011). However, there is no one area of the homepage that is inherently more prominent than another. Such values are inscribed by individuals (e.g., online editors) based on their perceptions of the norms and cultural practices of their constructed audience (e.g., what portions of the screen they typically look at first).

To assign a ranking of prominence for the distinct areas of a homepage, the conventions adopted by Lim (2010, 2012), Boczkowski and colleagues (Boczkowski et al., 2011; Boczkowski & Peer, 2011), and Lee et al. (2014) were also adopted in this study. Specifically, the prominence of a given region was determined by following an F-shape pattern that privileged distinct spots where items could be placed appearing left to right and then top to bottom. The present research departed from those scholars, however, by privileging areas of the homepage that were clearly intended to draw readers, such as those that included large pictures and larger font sizes.[4] The rationale for this decision is

[4] It is important to note that a different protocol for coding the prominence of the news items would likely yield different results. This process, detailed in Chapter III, aimed to be both consistent with prior work to

that those items are clearly distinguished from competing items, thus indicating the intent on the part of the staff to draw attention to them.[5]

Following this procedure, the five most prominent areas of a homepage (for a given design) that could contain a news item were identified and assigned a value from 1 (least prominent) to 5 (most prominent), and all news items that appeared in that area over the course of the collection period were assigned that value. If an item did not appear in a prominent area of the homepage, it was assigned a value of 0. Only a single item was coded for each area (the dominant headline) and "related items"—news items that appear as sub-units of an associated parent item—were excluded. The rationale for this exclusion is that such items are clearly assigned a lower value of newsworthiness by being treated as tangential and not worthy of a primary spot on the page.[6] As with popularity, this variable was treated as dichotomous to assess some of the research questions and hypotheses, and as an ordinal variable to assess others (1 for least prominent, 5 for most). For examples of how different websites were coded, see Figure 2, Figure 3, Figure 4.

Procedure

A computational content analysis was performed to assess the research questions

---

facilitate comparisons but also sensitive to developments in the scholarly understanding of how homepages are designed and the kinds of aesthetic artifacts that draw attention.

[5] This is sometimes made evident in a website's code, with a specific section (e.g., the central spot with a large accompanying picture) assigned an id or class element like 'top_story' or 'lead_item'.

[6] The links for these items were also typically in relatively small font sizes that would be easily missed by a reader who casually scans the page.

and hypotheses. A computational analysis was favored because of the large volume of data involved in the study and because the systematic nature of the source of data—the homepages—made a computational analysis less error-prone. For example, changes in the headline of a story would not lead to the miscoding of the item as a new, distinct unit since the URL would not typically change. The homepages of the 14 news organizations were systematically downloaded and stored every 15 minutes between October 18 and December 20, 2014 through the use of Python-based computer scripts developed by the author. Specifically, these scripts used the Selenium framework to simulate multiple browsing sessions with the Mozilla Firefox browser and store the page's source code—after all elements had been loaded—in an organized fashion. Because the U.S. mid-term elections—an exceptional and planned event that prompted the use of a special layouts across organizations and a focus on constantly updated voting results—occurred during this time period, all data collected on November 3, 4, and 5 were discarded.

A second set of computer scripts were then developed for each news organization. For each downloaded file, these scripts would use the BeautifulSoup library to detect the layout of the page; identify the location of the 'most viewed' list and extract the top five items; identify the five most prominent spots on the homepage and extract the dominant item in each spot; and store information about each item in a relational database, along with additional information like the name of the organization and the time of the snapshot

analyzed in the organization's time zone.[7] In order to ensure that only news items were coded (rather than links to static pages like section fronts), a URL pattern for news items was identified for each organization and only those items fitting the pattern were recorded. All scripts had extensive built-in error detection to help the researcher identify and diagnose instances where the algorithm failed to code content, and a subset of the final coded data was manually reviewed by the researcher using a customized electronic interface to ensure that the content was coded accurately. Data were then cleaned to ensure that there were no persistent entries (i.e., items that appeared in all snapshots) and that URLs with symbolic paths (i.e., singular items that could be accessed through different links) were reduced to their unique identifiers. For more information about this procedure, see Chapter III.

Results

Data were collected for a total of 29,465 distinct news items that appeared on either the top five spots of the 'most viewed' list or in one of the five most prominent areas of the homepages of the 14 news organizations over 61 days using 15-minute intervals. This yielded a total of 684,931 rows of data that included information about 10 variables pertaining to each unit (news item), such as the time of the snapshot in the local time zone and its popularity and prominence rankings at that point in time. As shown in Figure 9, there were notable differences in the volume of distinct items, with *The Denver*

[7] If there were multiple lists of most viewed items, the one covering the shortest time period (e.g., 'past hour') was selected.

Post (3,775), the *San Jose Mercury News* (3,166), and the *St. Paul Pioneer Press* (3,047) having the greatest number of news items. In contrast, the *Register* (864), the *Milwaukee Journal Sentinel* (1,192), and the *Fort Worth Star-Telegram* (1,288) had the lowest amount of items. That there is considerable variation in the amount of unique items is, of course, to be expected since some news organizations produced more original content or used more third-party (i.e., newswire) content than others and, more importantly, some organizations constantly added new material to the areas of prominence while others utilized those spots to highlight stories for an extended period of time.[8]

Content that is Popular and Prominent

The first hypothesis posited that a minority of prominent news items would be popular at some point in time. As shown in Figure 10, a minority of items appearing in an area of prominence or on the 'most viewed' list was both popular and prominent at some point in time in all but two cases. The overlap was greatest for the *Star Tribune* (54.9%), *The Star-Ledger* (50.6%), and the *Register* (43.3%). In contrast, the *Salt Lake Tribune* (12.1%), the *Fort Worth Star-Telegram* (15.4%), and the *Daily News* (20.4%) had the smallest proportion of items be both popular and prominent. The median proportion was 33.6%, indicating that, for the average news organization in this sample, only one-third of

[8] Although some organizations linked to the websites of wire services (e.g., *The Associated Press*), most incorporated wire copy into their content management systems by using the same URL pattern as their original copy, as well as a consistent aesthetic. Stories linked to outside of the primary domain used by the news organization (e.g., startibune.com) were generally excluded from the analysis, which would lead to a lower item count. Additionally, some organizations made it a point to focus to only include staff-produced content in prominent areas, which would also lead to a lower item count.

the items were both popular and prominent at some point in time. The first hypothesis was therefore supported.

The first research question inquired about the proportion of prominent news items that became popular before they were removed from an area of prominence. As shown in Figure 10, although a few organizations had a sizable proportion of items become popular prior to being removed from the areas of prominence, that proportion was often small. The proportion was highest for the *Star Tribune* (49.7%), *The Star-Ledger* (47.9%), and the *Plain Dealer* (41.3%). It was lowest for the *Salt Lake Tribune* (9.1%), *The Denver Post* (12.6%) and the *Fort Worth Star-Telegram* (12.9%). The median proportion was 21.8%, indicating that for the average news organization in this sample, just over one-fifth of the prominent items became popular before they were removed from an area of prominence.

To make the analysis parsimonious, only those organizations that had at least 20% of their news items appear on the 'most viewed' list prior to their last appearance in an area of prominence were deemed to have the potential for practically meaningful effects. That is, additional analysis of the organizations below that threshold would not detract from the conclusion that, for the vast majority of news items, their prominence is immune to their prior popularity based on the analytic strategy employed. The threshold, though artificial, is useful for establishing a minimal bound and drawing attention to only those organizations for which effects might be of general consequence.

Less than one-fifth of the news items became popular prior to being removed from the areas of prominence for the *Daily News*, the *Fort Worth Star-Telegram*, the *Salt*

Lake Tribune, the *San Jose Mercury News*, the *St. Paul Pioneer Press*, and *The Denver Post*. In contrast, the *Milwaukee Journal Sentinel*, the *Oregonian*, the *Plain Dealer*, the *Register*, the *Star Tribune*, *The Star-Ledger*, the *Wall Street Journal*, and the *Washington Post* all had a considerable proportion of news items become popular prior to being removed from the areas of prominence.

With regard to the first research question, it may therefore be surmised that just over half of the 14 organizations analyzed had more than one-fifth of their items—that is, a proportion large enough to potentially yield practically meaningful effects—become popular prior to being removed from the areas of prominence.

Effect of Popularity on Likelihood of Remaining in Prominent Area

In order to evaluate the second hypothesis, which posited that popular items would have a lower risk of being removed from the prominent regions of the homepage at a later point in time than their non-popular counterparts, and the second research question, which inquired as to whether more-popular items would have a lower risk of removal than less-popular items, Cox proportional hazards modeling (Andersen & Gill, 1982; Cox, 1972) was used to analyze the eight organizations that had at least one-fifth of their prominent news items become popular prior to being removed from an area of prominence. In this approach, a baseline hazard captures the effect of the time that elapsed since an item first appeared in a prominent area, and a set of covariates, such as whether it appears on a ‘most viewed’ list or not, can be added to evaluate their individual impact. Put differently, this technique allows for the evaluation of the

difference in the likelihood that an item will remain in an area of prominence 15 minutes later—as these were the measurement intervals—based on whether it appeared on a 'most viewed' list at that point in time.

Two sets of models were analyzed. The first set treated popularity as a dichotomous variable (i.e., popular or not popular), assessing the risk of removal relative to the reference category of non-popular items for all items that appeared in an area of prominence. The second set of models treated popularity as an ordinal variable (i.e., most popular to least popular), assessing the risk relative to the reference category of the least popular items for items that were both popular and appeared in an area of prominence. In both sets of models, prominence was treated as a dichotomous variable (present in one of the five areas of prominence or not). This allowed the evaluation of both general popularity as well as relative popularity.

As shown in the odd-numbered models in Table 3, it was the case for all but one of the organizations that appearing on the 'most viewed' list increased a news item's likelihood of remaining in an area of prominence 15 minutes later. The greatest effect was found for *The Star Ledger*, where the risk of removal was 62% lower than that of an item that did not appear on the list. It was followed by the *Star Tribune* (44% lower), the *Oregonian* (39% lower), the *Register* (32% lower), the *Plain Dealer* (29% lower), the *Wall Street Journal* (23% lower), and the *Washington Post* (22% lower). The lone exception was the *Milwaukee Journal-Sentinel*, where items appearing on the 'most viewed' list had a 21% higher risk of being removed from an area of prominence. In order to facilitate the interpretation of these results, the probabilities of removal over time

were estimated. As shown in Figure 11, these differences generally persisted over time and in some instances were quite substantial. For example, while a popular item on the homepage of *The Star-Ledger* had a roughly 65% probability of remaining in a prominent area of the page six hours later, a non-popular item had just a 25% probability. The second hypothesis was therefore supported.

With regard to the second research question, the ranking of an item within the list of 'most viewed' items had minimal effects on that item's risk of being removed from an area of prominence (see the even-numbered models in Table 3). Put differently, relative to the least popular item on that list, the differences in the risk of removal among the four higher rankings were usually statistically insignificant. In short, for most organizations, being popular enough to make it onto the top five spots of the 'most viewed' list had a notable impact on the item's visibility. However, the degree of popularity past that threshold did not.

Effect of Popularity on Subsequent Prominence

In order to assess the third hypothesis, which posited that an increase in an item's popularity would lead to a decrease in its prominence at a subsequent point in time, path analyses were conducted for the same eight organizations. This analytic strategy is useful because it allows the researcher to perform simultaneous regressions on multiple dependent variables—which is necessary for evaluating effects at multiple time points—and assess the directionality of those lagged effects by controlling for potential reciprocal effects among exogenous and endogenous variables. All analyses were performed using

maximum likelihood estimation.

A single theoretical model, shown in Figure 12, was utilized for all analyses. This model was adapted from the work of Lee and colleagues (2014), and assesses the effect of an item's popularity ranking at Time (t) on its prominence ranking at Time (t+1) across five distinct time points. In order to have a parsimonious model while assessing a sufficiently long period of time, a one-hour interval was used for this analysis. Time (0) therefore refers to when the item first became popular (i.e., the first instance in which an item appeared on the 'most viewed list') and Time (1) refers to that item's rankings one hour after that. This model controlled for the correlation between prominence at Time (t) and popularity at Time (t+1), the effect of popularity at Time (t) on popularity at Time (t+1), and the effect of prominence at Time (t) on prominence at Time (t+1).

As shown in Table 4, the theoretical model was a good statistical fit across the eight organizations analyzed. The comparative fit index exceeded the suggested minimum threshold of 0.93 (Byrne, 1994) for all models. Similarly, all models exceeded the recommended minimum threshold of 0.90 for the Tucker-Lewis index (Hu & Bentler, 1995). All models had a root mean square error of approximation below the recommended maximal threshold of 0.05 (Browne & Cudeck, 1993). The standardized root mean square residual for all models was below the suggested maximal threshold of 0.08 (Hu & Bentler, 1999). For some of the organizations, however, the chi-square test of model fit was statistically significant.

For three of the organizations analyzed—the *Milwaukee Journal-Sentinel*, the *Register*, and the *Washington Post*—there was no statistically significant effect of an

item's popularity at Time (t) on its prominence at Time (t+1) after controlling for potential reciprocal effects (see Table 4). That is, for the average item on those organizations' homepages, a change in its popularity ranking had no discernable statistical effect on its subsequent prominence ranking.

As shown in Table 4, for four of the organizations—the *Plain Dealer*, the *Star Tribune*, *The Star-Ledger*, and the *Oregonian*—the effects were typically negative. This indicates that, for the average item and after controlling for other effects, a one-rank advancement on the 'most viewed' list at Time (t) led to a decrease in that item's prominence ranking at Time (t+1). These effects were fairly consistent in terms of magnitude across these four publications, but they were strongest for the *Oregonian*: although there was no statistically significant lagged effect when the average item first became popular, there was a 0.13-unit decrease at Time (2), a 0.21-unit decrease at Time (3), a 0.19-unit decrease at Time (4), and finally a 0.06-unit *increase* at Time (5).

The lone organization not to have any negative effects was the *Wall Street Journal*. However, only one of those effects—four hours after the story first became popular—was statistically significant. Furthermore, this effect was quite small: a one-unit increase in the popularity ranking of that item at Time (4) led to a 0.08-unit increase in the prominence ranking at Time (5). The third hypothesis was therefore only partially supported, as only half of the news organizations exhibited the hypothesized negative relationship in a manner that was statistically significant.

Discussion

The purpose of this study was to assess the amount of overlap between the editorial and audience agendas as well as the effect that a news item's popularity had on its subsequent prominence on the homepage and on the likelihood that it would remain in a prominent area of the homepage at a later point in time. First, it was found that there remains an extensive gap between the editorial and audience agendas, as evidenced by the fact that, for the average organization in this study, only one-third of the items became both popular and prominent at some point in time. Second, it was found that for roughly half of the news organizations, less than one-fifth of prominent news items became popular prior to being removed from a region of prominence. This indicates that for many news organizations, only a relatively small set of items can be potentially influenced by the page view metric. Third, it was found that, among the set of organizations that had more than one-fifth of their prominent items become popular prior to being removed from a region of prominence, the effects of popularity on subsequent prominence were generally negative, though often statistically insignificant and invariably of a small magnitude. Fourth, it was found among those same organizations that popular items were less likely to be removed from the homepage 15 minutes later than items that were not popular.

The first finding is consistent with the findings from the work of Boczkowski and colleagues (Boczkowski et al., 2011; Boczkowski & Mitchelstein, 2013; Boczkowski & Peer, 2011), and offers additional evidence for Lee and Chyi's (2014) contention that readers rarely find content deemed to be newsworthy by journalists to be noteworthy. When viewed in conjunction with the second finding, one begins to see that the potential

effects of audience metrics on content placement are fairly muted for many organizations. That is, not only are the editorial and audience agendas very distinct in many cases, but by the time an item gains extensive popularity, it has often already been removed from a region of notable prominence.

When analyzing the subset of prominent items that can be affected by their popularity, one finds what would at first appear to be contradictory findings. Specifically, how can an item lose prominence as it gains popularity yet remain in a prominent area of the homepage longer? First, both findings are consistent with prior work—Bright and Nicholls found a 26% lower risk, on average, of removal for popular items and Lee et al. (2014) found an overall effect of -0.15 in their analysis of the lagged effect of popularity on prominence—thus easing fears that they are anomalies. Second, it is important to keep in mind that an item may maintain the same relative amount of popularity or prominence (i.e., the same ranking) for several consecutive points in time. Thus, an item may stay in a prominent region longer because it persists with the same amount of relative popularity. Third, and perhaps most important, is that the magnitude of the effect of a change in popularity on a change in prominence is invariably negligible for practical purposes: an item could increase from the lowest popularity ranking to the highest and not move down a single full prominence ranking.

Ultimately, the findings of this study, though similar in many ways to those of prior scholarship, lend themselves to a very different conclusion: that the influence of audience metrics on content may not be as great as assumed by many scholars, at least as it pertains to a particular, though key, editorial practice: placing content on the homepage.

This is not to say that audience metrics are of little consequence to journalistic work. A great deal of compelling ethnographic, survey, and interview-driven scholarship clearly indicates that they are becoming important discursive objects, and that journalists and editors alike certainly make use of those figures in a multitude of ways (e.g., Anderson, 2011a; Groves & Brown, 2011; MacGregor, 2007; Tandoc, 2014a; Usher, 2012, 2013). However, the findings of the present work indicate that a shift toward an "agenda of the audience" (Anderson, 2011a, p. 529) remains unrealized, thereby easing the fears among scholars and practitioners of a paradigm of journalism-by-the-numbers (Nguyen, 2013; Tandoc & Thomas, 2015).

More broadly, these findings lend support for the proposition that while online editors may be consulting audience metrics, they likely continue to rely primarily on other considerations, which may be more consistent with the occupational ideology and professional logic of journalism described by Deuze (2005) and Lewis (2012). That is, although this study did not evaluate the content of the news items and the extent to which they aligned with those traditional journalistic values, the extensive gap between prominent and popular items as well as the modest impact of relative popularity among popular items (i.e., most popular vs. least popular) suggests that assessments of newsworthiness and public service ideals may outweigh metrics-related considerations. It is, however, possible that metrics exert a strong influence for certain types of content, and not others. For example, a news organization may want to include at least one politics-related story in an area of prominence at all times, and look to the most popular politics-related stories to guide the selection of that content. Similarly, a news organization may

want to keep their areas of prominence free of certain kinds of 'soft' news, and thus ignore item popularity for that type of story. Given the potential mediating role of the subject matter of stories, scholars are encouraged to consider that variable in future scholarship.

The findings of the present study should also serve as a caution against overstating the role of the audience when it comes to particular gatekeeping decisions. Shoemaker and Vos (2009) argue that although scholars have a limited understanding of exactly how newsworkers use audience metrics to inform their editorial decisions, "we do know that the dotted line representing a weak audience feedback loop in mass communication models can now be made solid" (p. 7). The findings of this study highlight the importance of incorporating the word 'weak' as a qualifier for the influence of the audience channel, insofar as it pertains to the placement and de-selection of content on the homepage. Put differently, although audience feedback is indeed a part of the equation, it is important to keep in mind that when it comes to matters of gatekeeping, editorial newsworkers "are extremely reluctant to relinquish control over those decisions, despite the greatly increased visibility of user activity" (Singer, 2011, p. 630). Additionally, as Petre (2015) cautions, scholars should not assume that access to audience analytics will lead newsrooms to not only use analytics but use it in particular manners. It is important to consider the organizational context within which a given newsroom operates. For example, Petre (2015) found that Gawker's historical emphasis on metrics led to a focus on what had previously worked and the production of short stories at a fast pace. In contrast, the *New York Times*' organizational culture rhetorically and structurally

deemphasized the use of audience analytics and promoted the development of content that editors believed would be consistent with the newspaper's civic-minded mission. In particular, the *New York Times* restricted access to audience analytics to particular classes of newsworkers—a stark contrast to the Gawker websites—largely because of a fear that their metrics would be misinterpreted and perhaps misused (Petre, 2015).

While the present work found an overall limited impact of a particular metric on a specific editorial function, it did not investigate alternative uses of such metrics that may emerge in particular organizational contexts—like validating decisions guided by an intuitive understanding of the audience (see Petre, 2015; Usher, 2013). Additionally, it did not assess the role of the various factors at the organizational level and elsewhere (e.g., incentives for correctly placing content, as measured by the number of page views it ultimately receives, or access to particular audience analytics suites) that may explain uptake and use. A content analysis incorporating such potential sources of influences would offer a significant contribution to this stream of research.

It must also be noted that the specific findings pertaining to the extent of the divergence in the editorial and audience agendas and the effects of popularity on prominence should not be generalized to the entire news industry. First and foremost, the organizations analyzed have specific traditions (in newsprint) and cover only large news organizations. Moreover, lists of most-viewed items—a criterion for inclusion in the sample—are only present on the websites of certain news organizations, introducing particular sampling biases (see Chapter IV). Additionally, only the five most popular items were considered, using ordinal data that assumes equidistance between intervals

and that limits the potential variance. This is of particular import as it influences the findings by focusing strictly on the upper echelon of most popular items—that is, those that have gained exceptional popularity. This limits the potential for examining the overlap between the mid-level editorial and audience agendas and prevents the assessment of items that may have been sufficiently popular to catch the eye of an online editor, but not so popular as to join the elite group of 'most viewed' news items. A more comprehensive data source may yield results that are more sensitive to popularity gain, especially soon after an item is first published, and less influenced by the popularity of competing items. However, due to the commercial value and potential implications of such data, they would likely be difficult to obtain for a large number of organizations in order to engage in broad, comparative work. Scholars should nevertheless seek the cooperation of such organizations and communicate the potential benefits of developing partnerships. As Petre (2015), among others, has illustrated, partnerships between scholars and news organizations can be mutually beneficial, both in terms of offering a more comprehensive understanding of the general practice of journalism as well as reflective examinations of the particular processes and routines used by the specific industry partners.

In conclusion, the present study lends support for the contention that the presumed impact of audience metrics on a specific editorial function (the placement of content) may be overstated in the much of the scholarly literature. Specifically, the editorial agenda remains very distinct from the audience agenda and the effects of popularity on prominence for the relatively small subset of items that can be affected are

generally limited in both a statistical and practical sense. Therefore, although audience metrics may receive attention in placement-related decisions, it is quite possible that other considerations driven by the occupational ideology and professional logic of journalism take precedence.

CHAPTER VI:
CONCLUSION

The present work was driven by a desire to better understand the impact that audience metrics are having on the presentation of news content on the homepages of news organizations with a print heritage, and to offer methodological guidance for scholars interested in studying similar phenomena. It has demonstrated that algorithms can be effectively leveraged to computationally analyze certain aesthetics of a large volume of homepages; that the "most viewed" list can serve as a useful indicator of popularity, though it introduces some important limitations and should not be assumed to be comparable across organizations; and that the effect of an item's popularity on its subsequent placement on the homepage is fairly muted in the process of selection, though it is greater in the process of de-selection. In short, the present research indicates that the current effects of audience metrics—at least as it pertains to a particular editorial act—may be overstated in the literature, and offers pathways for further studying the relationship between audience metrics and the content produced by news organizations.

Having previously discussed the narrow set of implications of the findings, the following sections situate them among two broad questions of interest to the immediate context as well as the broader understanding and study of contemporary journalism: How might a computational social scientific paradigm contribute to the study of online journalism? And how might the proliferation of audience analytics and metrics affect the ability of journalism to fulfill its public service mission?

Online Journalism and Computational Social Science

The work of journalism has changed considerably over the past decade and a half as news organizations have shifted key resources toward digital and networked platforms (Agarwal & Barthel, 2013; Boczkowski, 2005; Soloski, 2013; Tang et al., 2011). For many large and mid-size newspapers, work is no longer oriented around an evening deadline; instead, a continuous deadline drives the process (Barnhurst, 2011). Stories are no longer filed when they are "complete"; rather, they are posted once enough basic information has been acquired, and updated over the course of the day (Mitchelstein & Boczkowski, 2009). Headlines are no longer crafted solely by experienced copy editors; they are often written on-the-fly by online editors and then subjected to real-time A/B testing (Soberman, 2013). Visual media attached to news items are no longer static objects, but often interactive and self-updating as in the case of dynamic data visualizations (Dick, 2013; Smit et al., 2014). Space is no longer viewed as a key limitation, and reporters are sometimes asked to contribute multiple short blog posts to supplement their reporting (Barnhurst, 2011).

In short, the volume and velocity of online news is far greater than its analog counterpart, and its liquid and interactive nature makes it quite distinct. From a methodological standpoint, this means that there are often more units to analyze, resulting in larger datasets; that units emerge on an irregular schedule, sometimes requiring data to be collected through a continuous process; and that they are not static, requiring at least one form to be 'frozen' in order to treat them as stable objects (Karlsson & Strömbäck, 2010; Karlsson, 2012; Sjøvaag & Stavelin, 2012).

In light of this, scholars have pointed to the importance of turning to

computational solutions to help scholars deal with these developments (Flaounas et al., 2013; Lewis et al., 2013; Shah et al., 2015; Zamith & Lewis, 2015). Computer programs are well-suited to serve as an aid for keeping up with and freezing content appearing on the homepages of news organizations, and can thus be leveraged to develop more rigorous study designs (Karlsson, 2012; Sjøvaag et al., 2015). For example, a mirroring tool can be used to capture all of the available content on a website, providing a population that may be evaluated as a census or randomly sampled from. Similarly, RSS feeds can be automatically accessed by a script in short intervals to ensure that units are always downloaded one minute after their introduction to a website, guaranteeing that they are frozen in a consistent manner. Moreover, automated processes enable shorter intervals to be assessed as there is no need to hire human beings to access and store content throughout the day.

Unsurprisingly, then, scholars have turned in recent years to specialized website mirroring tools like HTTrack and automated well-known programs like Wget to systematically capture and organize content (Hermida et al., 2014; Karlsson, 2012; Kiousis, Kim, McDewitt, & Ostrowksi, 2009; Sjøvaag et al., 2015). However, as websites have become more interactive—and in particular leveraged advanced browser features through the use of JavaScript code—these popular tools have become less useful due to their limited capabilities (e.g., their inability to handle dynamic objects, such as lists that change when a user clicks on a column header). Put differently, the study of contemporary online journalism in many instances demands new tools and processes for acquiring and storing such content.

The framework introduced in the present work offers a way to overcome these challenges, particularly through the use of Selenium to emulate a full, modern browser. The advantages of that procedure include not only the ability to process client-side instructions (i.e., JavaScript code to load certain elements on a page) but also in that it can simulate the experience of a particular user under specific circumstances. This may include accessing a page using different resolutions (a feature useful when assessing responsive designs) as well as taking a particular route to a news item or utilizing a particular browsing history (a feature useful for analyzing different representations of webpages that leverage data-driven algorithms to customize the user experience). This opens up new possibilities for researchers to explore novel questions pertaining to the presentation of online content and to work with content appearing in websites that leverage newer Web technologies. Furthermore, this framework enables the capturing of both the website code and full-page screenshots, which can be used for separate analyses or in combination (as was the case in the present work).

However, it is important to note that employing this framework comes at a considerable computational cost: modern browsers have become platforms in their own right, and thus require substantial resources to run (in the form of system memory and CPU cycles). Researchers intending to concurrently access multiple webpages or emulate different user experiences may therefore become limited by their hardware. Indeed, this was the case for the present work, where a server built with consumer-grade hardware could not access more than 21 homepages concurrently. Future work in this area should therefore explore lighter solutions that provide similar functionality, with Node.JS and

PhantomJS being alternatives worth exploring.

Beyond capturing and organizing particular forms of content, computational solutions also offer promise for analyzing that content. With datasets increasing in size—be it a function of analyzing more content from a single source or analyzing more sources in order to make comparisons—it becomes increasingly difficult for humans to drive the majority of the analytical work (Schwartz & Ungar, 2015; Shah et al., 2015). Although the use of computational tools to engage in analytical work has increased in recent years, there is a great deal of work that needs to be done in this area to create tools and frameworks that are both accessible and accurate (Hesse, Moser, & Riley, 2015; Lin, 2015). Indeed, in the context of assessing the liquidity of the homepages of news organizations, Karlsson (2012) has directly pointed to the lack of suitable computational solutions as a key barrier to comprehensive analyses.

The present work has pointed to the usefulness of technologies like BeautifulSoup, which may be leveraged to take advantage of the uniformity of certain web content—such as content appearing on websites that utilize content management systems (see also Sjøvaag et al., 2015). These technologies offer far greater consistency than most human coders while being exponentially faster and ostensibly more transparent (Zamith & Lewis, 2015). Indeed, a manual replication of this analysis would simply be unfeasible given the timeframe for the present work. Furthermore, the algorithms powering this analysis could be quickly adapted to evaluate other websites that use similar content management systems.

Beyond assessing the placement of content, the framework described in the

present work could also be adapted to assess other variables of interest to mass communication scholars. For example, this approach could be utilized to assess the presence and type of visual elements appearing alongside particular news content, calculate the number of comments posted by readers or specific sets of individuals, or identify the presence of non-traditional elements in news stories like lists of items. Moreover, the framework described could be utilized to extract specific portions of content that could then be fed to algorithms purposed for textual analyses, from keyword-based sentiment analysis to machine learning-driven frame analyses. In effect, this framework provides researchers with a solid foundation through which to engage in computational social scientific inquiry, both to evaluate traditional theoretical questions in a more systematic and reproducible fashion as well as to tackle novel theoretical questions that require computational work.

Fulfilling Journalism's Public Service Mission

Newswork is typically conducted with a constructed audience in mind (DeWerth-Pallmeyer, 1997; Gans, 1979; Pool & Shulman, 1959; Schlesinger, 1978). As Napoli (2011) has documented, news organizations have long attempted to measure their audience's preferences in order to inform those constructions, from the proliferation of readership reports in the 1930s to the emergence of news consultants in the 1970s. However, it appears that we are in the midst of a third wave toward the rationalization of audience understanding. This is evidenced by the tremendous growth in the prevalence of audience analytics: view the source code of any page on a news organization's website

and one is almost certain to find references to Chartbeat, Omniture, Parse.ly, or some other analytic platform.

Though journalists and editors, especially those working in newspaper organizations, have long rejected information about their audience, there is considerable evidence that this is changing (Anderson, 2011a; Groves & Brown, 2011; Tandoc, 2014b; Usher, 2012; Vu, 2014). Drawing from theories of organizational change, one can view this change as a rational response to the environmental uncertainty that has emerged from a precipitous drop in revenue and considerable growth of nontraditional competitors, resulting in a tighter coupling of editorial and business considerations (Gade, 2009a; Lowrey, 2011; Meyer et al., 1993; Snow et al., 2005). Unsurprisingly, then, these technologies and their output have gained considerable cultural and economic cachet as a way of better understanding and appealing to news consumers.

The result of the proliferation of audience analytics and metrics within newsrooms, occurring within the context of considerable economic uncertainty, has been a discussion that has largely framed the use of audience analytics or the employment of particular metrics to inform editorial decisions as being either good or bad. These discussions often begin—as they should—by considering the role news media should play in society. In the United States, the dominant theory emphasizes a watchdog role, wherein news media should watch powerful interests, prevent abuses of power, and hold them to account for their actions (Siebert et al., 1956). From another perspective, journalism should serve a communitarian role, facilitating public discourse and providing a space for dialogue on the common good, thereby bringing the public into existence as a

community (Christians, 1997; Croteau & Hoynes, 2001; Lee Plaisance, 2005). Though these theories are often unreflective of contemporary practice and instead largely a product of journalism's own myth-making (Hampton, 2010), they are nevertheless useful for discussing the ethical ideal-types for the field (Tandoc & Thomas, 2015). In particular, both of these theories emphasize the importance of having journalism serve the public interest in some fashion (Bennett et al., 2008; Carey, 1993; Habermas, 1989).

There has been a rhetoric of empowerment emerging around the inclusion of the audience in the process of news production, with proponents arguing that news consumers are largely rational and able to determine and identify the information they need to participate in a democratic society (Batsell, 2015; Henry, 2012; Lee & Chyi, 2014). That is, the proliferation of audience analytics is a good outcome because it turns the so-called audience from passive consumers to active agents capable of influencing the content presented to them. In contrast, others have argued that this phenomenon leads to a race toward the lowest common denominator, wherein news organizations are rewarded for catering to inconsequential interests, often at the expense of civically-important work (Nguyen, 2013). As Tandoc and Thomas (2015) put it, the confluence of market pressures and the increased availability of analytics may ultimately create "a media ecosystem that panders to, rather than enlightens and challenges, its audience, and thus poses a barrier to the … collective subscription to the success of democracy" (p. 249).

Such a discussion is, however, limiting. In particular, it considers audience analytics largely through a deterministic lens, and grants audience metrics exceptional power. The question should not be, is the use of audience analytics and metrics ethical?

Rather, a more useful question is, how can audience analytics and metrics be used ethically? Such a question acknowledges that technological affordances may be appropriated in different ways, and put to uses that are consistent with the values of their users (Gillespie, Boczkowski, & Foot, 2014; Leonardi, 2009; Siles & Boczkowski, 2012). That is, although technologies are developed with certain intentions and assumptions in mind, these intentions and assumptions may be deliberately ignored by users in the process of interacting with the technology, leading them to rationalize and utilize that technology in a manner that is consistent with their ideology and routines (Orlikowski, 2000; Williams & Edge, 1996). Similarly, such a question recognizes that the meaning of cultural objects like audience metrics is socially shaped and therefore malleable (Carey, 1989). Such a perspective enables a discussion that focuses on how to best make use of audience analytics and metrics to further the public-service mission of journalism.

According to Hindman (forthcoming), ethical journalistic practice in the contemporary media environment demands attention to data on audience behavior. Under an ethical model, audience analytics and metrics are used to gain a more nuanced understanding of what appeals to the audience in order to maximize the audience for civically valuable content. Put differently, this approach maintains the core aim of serving the public with the information that news professionals deem to be important but alters the practices adopted to inform those decisions (Hindman, forthcoming). As such, it implicitly treats editorial autonomy and sensitivity to audience data as a false dichotomy and explicitly encourages journalists and editors to use audience data to make civically valuable content more attractive, such as by seeking out patterns in the data in

order to identify story formats that resonate with news consumers, types of supplementary content that is of value to them in particular contexts, and even portions of the content they find uninteresting. Ultimately, this requires journalists and editors to look beyond the page view metric, which provides a limited and often problematic reflection of the kinds of content that users find appealing (Graves & Kelly, 2010). Instead, they should evaluate a broader cocktail of complementary measures (such as time spent on page, recirculation, and referrals) that can not only increase confidence in those emergent patterns by triangulating information but also provide greater nuance and detail about audience preferences. This would demand a substantial shift in how editorial newsworkers typically interact with such data as they often rely on a single metric (e.g., page views or concurrent visitors) —and sometimes develop strong emotional responses to that metric—despite attempts by the designers of systems like Chartbeat to call attention to alternative metrics (Petre, 2015). Furthermore, it would require journalists and editors to identify and be cognizant of the limitations of audience analytics and metrics, such as the dimensions of audience behaviors that cannot be accurately captured by contemporary systems. Put differently, it would require those individuals to recognize that not everything that can be counted counts, and that such data are not inherently objective nor complete (Petre, 2015).

The findings of the present work, which indicate a rather limited effect of an item's popularity on its subsequent prominence and visibility on the homepage, may in effect be pointing to the diminishing influence of the page view metric as well as a move toward a more ethical use of audience analytics and metrics—that is, an approach that

neither outright rejects the use of those systems and measures due to the perception that they inherently oppose key values in journalism nor utilizes them unreflectively to maximize perceived economic prospects, but rather identifies ways in which they can be used to enhance the organization's ability to fulfill its mission and goals.[1] From a methodological standpoint, such a shift would indicate a need to move away from treating page views as the sole indicator of popularity. This change would introduce the considerable challenge of acquiring data for multiple metrics, which would likely require the cooperation of the news organizations being studied as such data are unlikely to be found in any public-facing interface (e.g., a list of stories with the greatest amount of 'engagement'). Furthermore, it would require the researcher to accurately capture the calculus for weighing those different measures of appeal (e.g., page views and time spent on page), which is likely to vary considerably across individual editors and news organizations.

From a theoretical standpoint, a shift toward the ethical use of audience analytics and metrics would require that audience feedback not only be included in models of journalistic work, but be granted a privileged position. That is, under an ethical framework, the gatekeeping process would be more reflective of the latest version of the

---

[1] This contention is consistent with the findings of the author's ongoing research, which consists of interviews with online editors at several of the organizations examined in the present work. However, as Petre (2015) has noted, a shift toward a more ethical use of audience analytics and metrics would require a significant cultural shift among editorial newsworkers. Though Chartbeat—a popular audience analytics suite—has rhetorically emphasized alternative metrics, such as engagement time and recirculation, and prominently juxtaposed them against page views on their dashboard interface, Petre (2015) found that Chartbeat's speedometer-like dial that shows how many individuals are accessing a particular page at any given moment remains the product's most popular feature and has a considerable emotional impact.

gatekeeping model, put forth by Shoemaker and Vos (2009). Nevertheless, an ethical framework would continue to require that traditional inputs pertaining to journalism's occupational ideology and professional logic be taken into account. That is, while the role of certain inputs, such as the perceived need to be autonomous and in control of news information (Deuze, 2005; Lewis, 2012), would surely be diminished relative to earlier interpretations of the gatekeeping process, it would not be the case that editors cater to their audience in an unreflexive fashion.

Such a model would promote a middle way for editorial newsworkers: one that would enable them to be more responsive to what their audience wants without having to compromise the key values that define their semi-profession and enable them to perform an important public service. Future work should explore this development to help disentangle the beliefs oriented around the way audience analytics and metrics should be used—and the extent to which the ethical use of audience analytics and metrics is manifesting itself in contemporary practice. Until such a shift has received empirical support, however, scholars must be careful not to overstate the role of the audience when it comes to gatekeeping decisions, particularly those involving the placement and de-selection of content on the homepages of news organizations.

Final Remarks

The growth in the availability of audience analytics and the cachet of audience metrics is a phenomenon that is likely to persist. In particular, it marks a continuation of a decades-old effort to further quantify audiences and provide decision-makers with data

from which they can make decisions. However, in studying this phenomenon, scholars must be mindful not to overstate its consequence or give the technology powering it a deterministic character. In particular, though the phenomenon certainly has the potential drive journalism away from its public service mission, it also has the potential to be used in a manner that is entirely consistent with it. News professionals must be mindful of this, and act accordingly if journalism is to preserve its vaunted status and continue to make jurisdictional claims as the primary sense-maker of current events.

Though an impressive body of work has emerged around audience analytics and metrics, much remains unknown about it. In particular, more attention must be paid to the manner in which content is transforming, rather than presuming such transformations solely from observations of the process of journalism or through self-reported accounts. However, as the use of analytics by news practitioners becomes more complex, so must the scholarly endeavors that seek to understand it. The present work has offered a useful foundation to build upon, particularly in how to analyze the liquid online content that is of growing import. As the present work has illustrated, however, much work remains to be done for identifying pathways for acquiring and properly evaluating audience analytics data. Such work will almost certainly demand the cooperation of the organizations being studied (i.e., by providing access to private data), though tools and frameworks must yet be developed for turning that data into something that scholars can use. Despite these challenges, the growing pervasiveness of audience analytics and metrics makes it an area that demands continued work.

TABLES & FIGURES

Table 1

List of the 50 Largest U.S. Newspaper Organizations

Name	Location	Parent Company	Circulation	Analyzed
Arizona Republic	Phoenix, AZ	Gannett Company Inc.	290,653	No
Arkansas Democrat Gazette	Little Rock, AR	WEHCO Media Inc.	161,047	No *
Atlanta Journal-Constitution	Atlanta, GA	Cox Media Group	198,568	No
Boston Globe	Boston, MA	Boston Globe Media Partners	238,108	No
Buffalo News	Buffalo, NY	The Buffalo News	160,674	No *
Chicago Sun-Times	Chicago, IL	Wrapports, LLC	451,864	No
Chicago Tribune	Chicago, IL	Tribune Publishing Company	413,475	No
Cincinnati Enquirer	Cincinnati, OH	Gannett Company Inc.	130,968	No
Courier-Journal	Louisville, KY	Gannett Company Inc.	139,225	No
Daily News	New York, NY	New York Daily News	501,130	Yes
Dallas Morning News	Dallas, TX	A.H. Belo Corporation	409,696	No
Detroit News/Free Press	Detroit, MI	Gannett/MediaNews	331,005	No
El Vocero de Puerto Rico	San Juan, PR	El Vocero de Puerto Rico	216,723	No
Fort Worth Star-Telegram	Fort Worth, TX	McClatchy Company	186,625	Yes
Hartford Courant	Hartford, CT	Tribune Publishing Company	129,903	No
Honolulu Star-Advertiser	Honolulu, HI	Oahu Publications, Inc.	200,682	Yes
Houston Chronicle	Houston, TX	Hearst Newspapers	332,954	Yes
Indianapolis Star	Indianapolis, IN	Gannett Company Inc.	159,037	No
Kansas City Star	Kansas City, MO	McClatchy Company	186,350	Yes
Las Vegas Review Journal	Las Vegas, NV	Stephens Media Group	252,110	No *
Los Angeles Times	Los Angeles, CA	Tribune Publishing Company	647,723	No
Miami Herald	Miami, FL	McClatchy Company	191,426	Yes
Milwaukee Journal Sentinel	Milwaukee, WI	Journal Communications, Inc.	202,573	Yes
New York Post	New York, NY	News Corporation	547,508	No
New York Times	New York, NY	New York Times Company	1,852,698	Yes
Newsday	Long Isla., NY	Newsday Holdings LLC	427,721	No
Oregonian	Portland, OR	Oregonian Publishing Co.	226,566	Yes
Orlando Sentinel	Orlando, FL	Tribune Publishing Company	161,837	No
Philadelphia Inquirer	Philadelphia, PA	Philadelphia Media Network	301,639	No
Pittsburgh Post-Gazette	Pittsburgh, PA	Block Communications, Inc.	177,411	No *
Plain Dealer	Cleveland, OH	Plain Dealer Publishing Co.	292,302	Yes
Register	Santa Ana, CA	Freedom Communications	320,628	Yes
Sacramento Bee	Sacramento, CA	McClatchy Company	195,030	No *
Salt Lake Tribune	Salt Lake City, UT	Newspaper Agency Corp.	237,493	Yes
San Francisco Chronicle	San Franc., CA	Hearst Newspapers	223,225	No
San Jose Mercury News	San Jose, CA	MediaNews Group, Inc.	232,272	Yes
Seattle Times	Seattle, WA	Seattle Times Company	259,138	Yes
South Florida Sun-Sentinel	Fort Laud., FL	Tribune Publishing Company	161,933	No
St. Louis Post-Dispatch	St. Louis, MO	Lee Enterprises, Incorporated	169,352	Yes
St. Paul Pioneer Press	St. Paul, MN	MediaNews Group, Inc.	236,279	Yes
Star Tribune	Minneapolis, MN	Star Tribune Media	303,929	Yes
Sun	Baltimore, MD	Tribune Publishing Company	171,614	No
Tampa Bay Times	St. Petersburg, FL	Times Publishing Company	246,240	No
The Denver Post	Denver, CO	MediaNews Group, Inc.	414,673	Yes
The Star-Ledger	Newark, NJ	Advance Publications, Inc.	305,903	Yes
Tribune Review	Pittsburgh, PA	Trib Total Media	200,502	No
U-T San Diego	San Diego, CA	San Diego Union-Tribune	225,189	No *
USA Today	Washing., DC	Gannett Company Inc.	1,739,338	No
Wall Street Journal	New York, NY	Dow Jones/News Corp.	2,320,915	Yes
Washington Post	Washing., DC	Nash Holdings, LLC	454,938	Yes

All figures and names according to the *Alliance for Audited Media* on Sept. 26, 2014. Items with an asterisk had a list of 'most viewed' items, but were not analyzed due to resource limitations.

Table 2

List of News Organizations Analyzed in Chapter V

Organization	Location	Parent Company	Circulation
Daily News	New York City, NY	New York Daily News	501,130
Fort Worth Star-Telegram	Fort Worth, TX	McClatchy Company	186,625
Milwaukee Journal Sentinel	Milwaukee, WI	Journal Communications, Inc.	202,573
Oregonian	Portland, OR	Oregonian Publishing Company	226,566
Plain Dealer	Cleveland, OH	Plain Dealer Publishing Co.	292,302
Register	Santa Ana, CA	Freedom Communications, Inc.	320,628
Salt Lake Tribune	Salt Lake City, UT	Newspaper Agency Corporation	237,493
San Jose Mercury News	San Jose, CA	MediaNews Group, Inc.	232,272
St. Paul Pioneer Press	St. Paul, MN	MediaNews Group, Inc.	236,279
Star Tribune	Minneapolis, MN	Star Tribune Media	303,929
The Denver Post	Denver, CO	MediaNews Group, Inc.	414,673
The Star-Ledger	Newark, NJ	Advance Publications, Inc.	305,903
Wall Street Journal	New York, NY	Dow Jones/News Corp.	2,320,915
Washington Post	Washington, D.C.	Nash Holdings, LLC	454,938

Table 3

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Milwaukee Journal Sentinel					
	Model 1			Model 2		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	0.19 *	0.09	1.21			
Rank 2				0.19	0.30	1.21
Rank 3				0.19	0.28	1.21
Rank 4				0.30	0.27	1.35
Rank 5				-0.11	0.29	0.9
N		22,868			3,234	
Wald		4.19 *			3.31	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Oregonian					
	Model 3			Model 4		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.50 ***	0.07	0.61			
Rank 2				-0.05	0.20	0.95
Rank 3				-0.22	0.20	0.80
Rank 4				-0.34	0.20	0.71
Rank 5				-0.08	0.19	0.92
N		25,871			7,232	
Wald		54.51			3.89	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Plain Dealer					
	Model 5			Model 6		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.34 ***	0.06	0.71			
Rank 2				0.14	0.19	1.15
Rank 3				0.21	0.18	1.23
Rank 4				-0.05	0.19	0.95
Rank 5				0.01	0.18	1.01
N		25,872			8,061	
Wald		28.94 ***			3.32	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Register					
	Model 7			Model 8		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.39 ***	0.11	0.68			
Rank 2				-0.29	0.32	0.75
Rank 3				-0.21	0.30	0.81
Rank 4				-0.33	0.30	0.72
Rank 5				-0.40	0.31	0.67
N		17,101			3,855	
Wald		12.56 ***			1.95	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Star Tribune					
	Model 9			Model 10		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.59 ***	0.06	0.56			
Rank 2				0.09	0.16	1.09
Rank 3				-0.16	0.16	0.85
Rank 4				-0.34 *	0.16	0.71
Rank 5				-0.48 **	0.16	0.62
N	24,452			10,865		
Wald	93.29 ***			18.32 ***		

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	The Star-Ledger					
	Model 11			Model 12		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.97 ***	0.07	0.38			
Rank 2				-0.06	0.21	0.95
Rank 3				-0.25	0.21	0.78
Rank 4				-0.57 *	0.22	0.56
Rank 5				-0.23	0.21	0.80
N	24,861			9,343		
Wald	172.30 ***			7.89		

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Wall Street Journal					
	Model 13			Model 14		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.26 *	0.10	0.77			
Rank 2				-0.53	0.31	0.59
Rank 3				-0.55	0.31	0.58
Rank 4				-0.57	0.31	0.57
Rank 5				-0.48	0.28	0.62
N		22,710			2,307	
Wald		6.40 *			5.72	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 3 *(continued)*

Effect of Being Popular at Time (t) on Visibility at Time (t+1)

	Washington Post					
	Model 15			Model 16		
Predictor	*B*	*SE*	*Exp(B)*	*B*	*SE*	*Exp(B)*
Popular (Yes)	-0.24 ***	0.10	0.78			
Rank 2				-0.22	0.28	0.80
Rank 3				0.26	0.26	1.29
Rank 4				-0.48	0.31	0.62
Rank 5				-0.31	0.26	0.73
N		21,864			4,154	
Wald		6.55 ***			9.26	

Note: Odd-numbered models evaluate popularity as a dichotomous variable (popular/not popular), whereas even-numbered models evaluate it as a continuous variable (most popular to least popular). Even-numbered models only include the items that appeared on the 'most viewed' list at least once. The reference category for the even-numbered models is Rank 1 on the 'Most Viewed' list. All rankings were reverse coded so that higher rankings indicate greater popularity. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$.

Table 4

Effect of Item's Popularity Ranking at Time (t) on its Prominence Ranking at Time (t+1)

	Milwaukee Journal Sentinel		*Oregonian*		*Plain Dealer*	
Time (t)	*B*	*SE*	*B*	*SE*	*B*	*SE*
Pop. @ T(1)	0.01	0.09	-0.09	0.06	-0.05	0.06
Pop. @ T(2)	0.01	0.09	-0.13 **	0.05	-0.16 ***	0.05
Pop. @ T(3)	0.01	0.09	-0.21 ***	0.04	-0.27 ***	0.04
Pop. @ T(4)	0.16	0.12	-0.19 ***	0.05	-0.12 **	0.04
Pop. @ T(5)	0.01	0.04	0.06 *	0.03	0.04	0.02
N	206		553		566	
χ^2	37.77		47.89 *		40.11	
RMSEA	0.03		0.03		0.02	
CFI	1.00		1.00		1.00	
TLI	1.00		0.99		1.00	
SRMR	0.03		0.03		0.04	

Note: Estimates represent unstandardized coefficients. All models have 32 degrees of freedom. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. All rankings were reverse coded; unit increases thus indicate greater popularity or prominence.

Table 4 *(continued)*

Effect of Item's Popularity Ranking at Time (t) on its Prominence Ranking at Time (t+1)

	Register		*Star Tribune*		*The Star-Ledger*	
Time (t)	*B*	*SE*	*B*	*SE*	*B*	*SE*
Pop. @ T(1)	-0.01	0.02	-0.10 *	0.04	-0.13 *	0.06
Pop. @ T(2)	0.02	0.03	-0.08 *	0.04	-0.07	0.05
Pop. @ T(3)	0.05	0.04	-0.15 ***	0.03	-0.17 ***	0.05
Pop. @ T(4)	0.01	0.04	-0.06	0.04	-0.15 **	0.05
Pop. @ T(5)	-0.01	0.03	0.01	0.02	0.05 *	0.03
N	239		644		581	
χ^2	50.94 *		36.25		76.63 ***	
RMSEA	0.05		0.01		0.05	
CFI	1.00		1.00		0.99	
TLI	0.99		1.00		0.98	
SRMR	0.03		0.02		0.05	

Note: Estimates represent unstandardized coefficients. All models have 32 degrees of freedom. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. All rankings were reverse coded; unit increases thus indicate greater popularity or prominence.

Table 4 *(continued)*

Effect of Item's Popularity Ranking at Time (t) on its Prominence Ranking at Time (t+1)

	Wall Street Journal		*Washington Post*	
Time (t)	*B*	*SE*	*B*	*SE*
Pop. @ T(1)	0.05	0.06	-0.10	0.06
Pop. @ T(2)	0.02	0.05	-0.03	0.06
Pop. @ T(3)	0.09	0.07	0.03	0.06
Pop. @ T(4)	0.12	0.06	0.02	0.06
Pop. @ T(5)	0.08 *	0.03	0.03	0.02
N	293		276	
χ^2	32.84		27.22	
RMSEA	0.01		0.02	
CFI	1.00		1.00	
TLI	1.00		1.00	
SRMR	0.03		0.02	

Note: Estimates represent unstandardized coefficients. All models have 32 degrees of freedom. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. All rankings were reverse coded; unit increases thus indicate greater popularity or prominence.

```
##### Locate the 'Most Viewed' Items and Store in List
try:
    mostviewed_1to5 = [link.find("a", href=link_pattern) for link in document_soup.find("div", class_="tab-content most-viewed").find_all("li")]
    mostviewed_linklist = []
    for link in mostviewed_1to5:
        try:
            link = link.get("href")
            mostviewed_linklist = parserfunctions.linklist_actions(link, mostviewed_linklist)
        except:
            pass
except:
    message = "Failed to retrieve the list of 'most viewed' links"
    seriousness = 2
    parserfunctions.error_log_entry(cur, conn, mysql_log_name, curr_time, pubshort, homepage, seriousness, message)
```

Figure 1. A sample of the code used to identify the ‘most viewed’ list on the homepage of the *New York Times*. A selector is used to identify the first ‘div’ element with the ‘class’ attribute of ‘tab-content most-viewed’ and then identify all ‘li’ child elements that have an ‘a’ element with an ‘href’ attribute matching the organization’s link pattern. If the selector fails to identify those elements, the error is logged in a database.

Figure 2. A screenshot of *The Denver Post*'s homepage on November 4, 2014 at 04:45 UTC. The most prominent area is the center piece, accompanied by a large picture. The subsequent areas all appear on the left bar. Note that although this area says, "Latest News," the items are not ordered chronologically.

Figure 3. A screenshot of the *New York Times*' homepage on December 9, 2014 at 00:15 UTC. The most prominent area is the top-left piece due to its large font size, followed by the middle item with its large accompanying picture, and then the items on the left bar due to their comparatively larger font sizes as well.

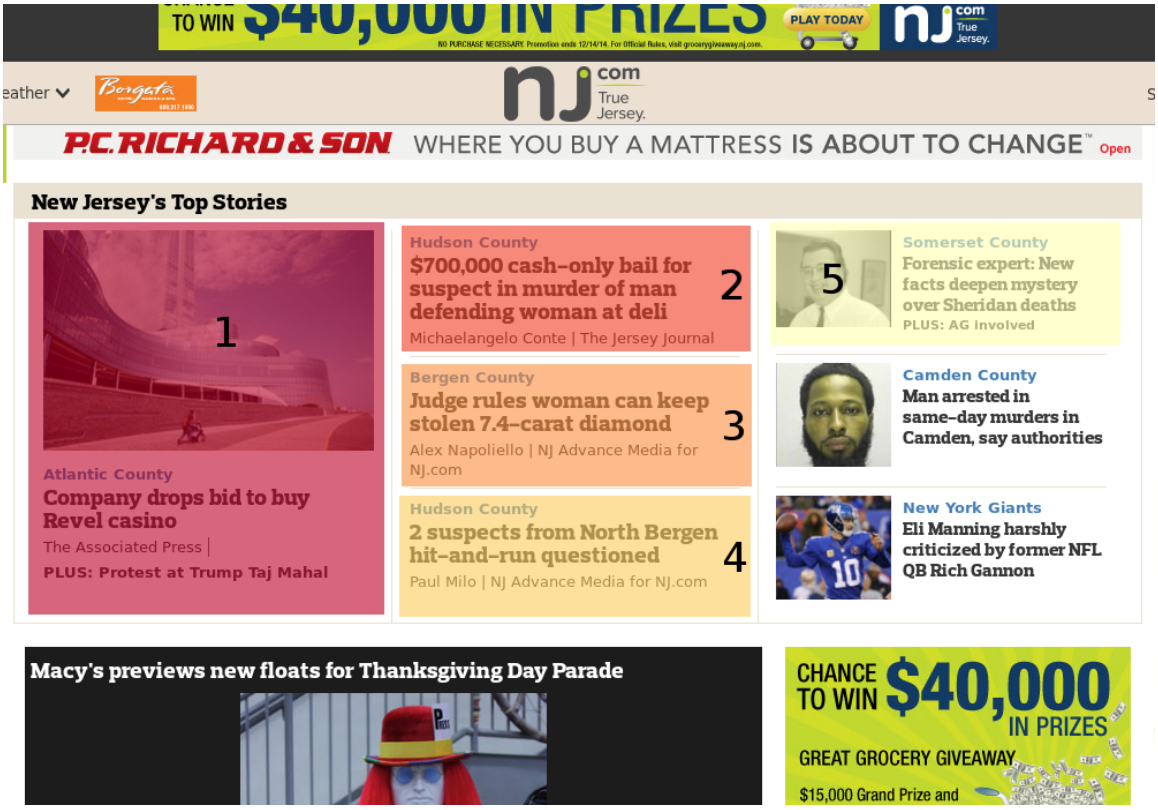

Figure 4. A screenshot of *The Star-Ledger*'s homepage on November 20, 2014 at 12:45 UTC. The most prominent area is left-most, as evidenced by the large accompanying picture. The second to fourth most prominent areas are in the middle, as they also have larger font sizes. The fifth most prominent area is the top-most item on the right bar; though it has a small picture, it has a smaller font size than the other items and appears to their right.

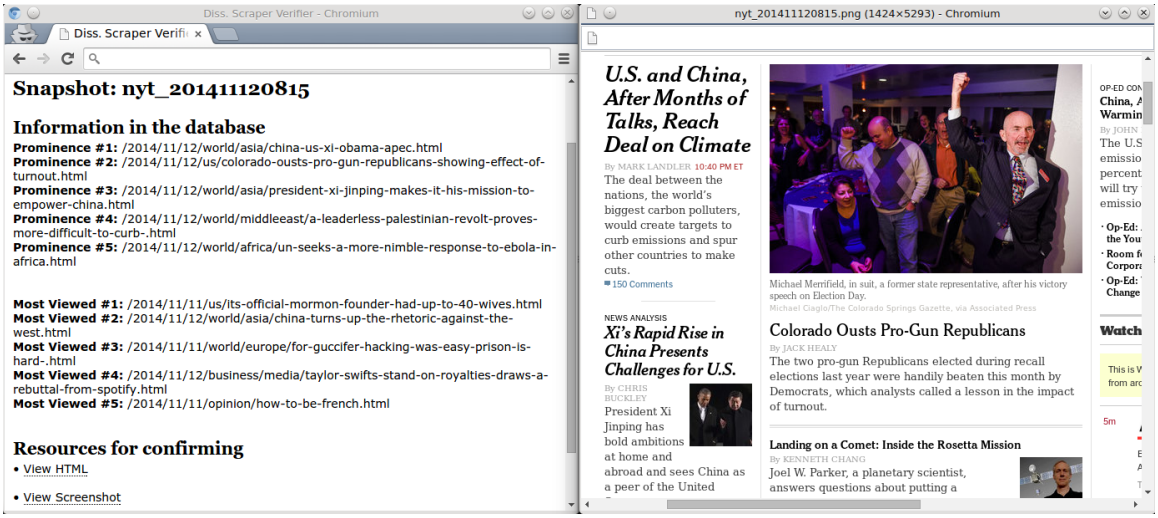

Figure 5. A screenshot of the researcher's interface for verifying the algorithm's coding decisions. On the left side, an electronic interface displaying the database information for the *New York Times* snapshot for November 12, 2014 at 8:15 UTC and links to the stored items. On the right side, a screenshot of that snapshot.

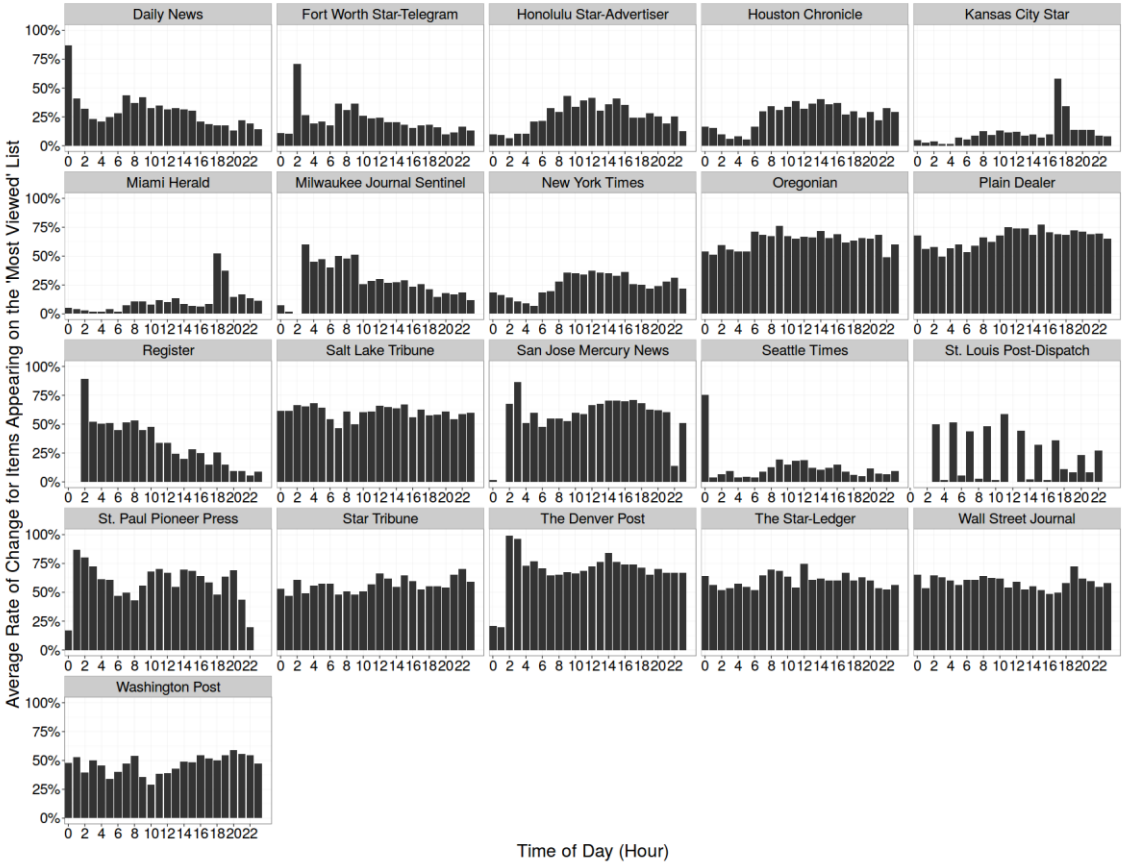

Figure 6. The average rate of change for the items appearing in the top five spots of the 'most viewed' list over the course of 61 days.

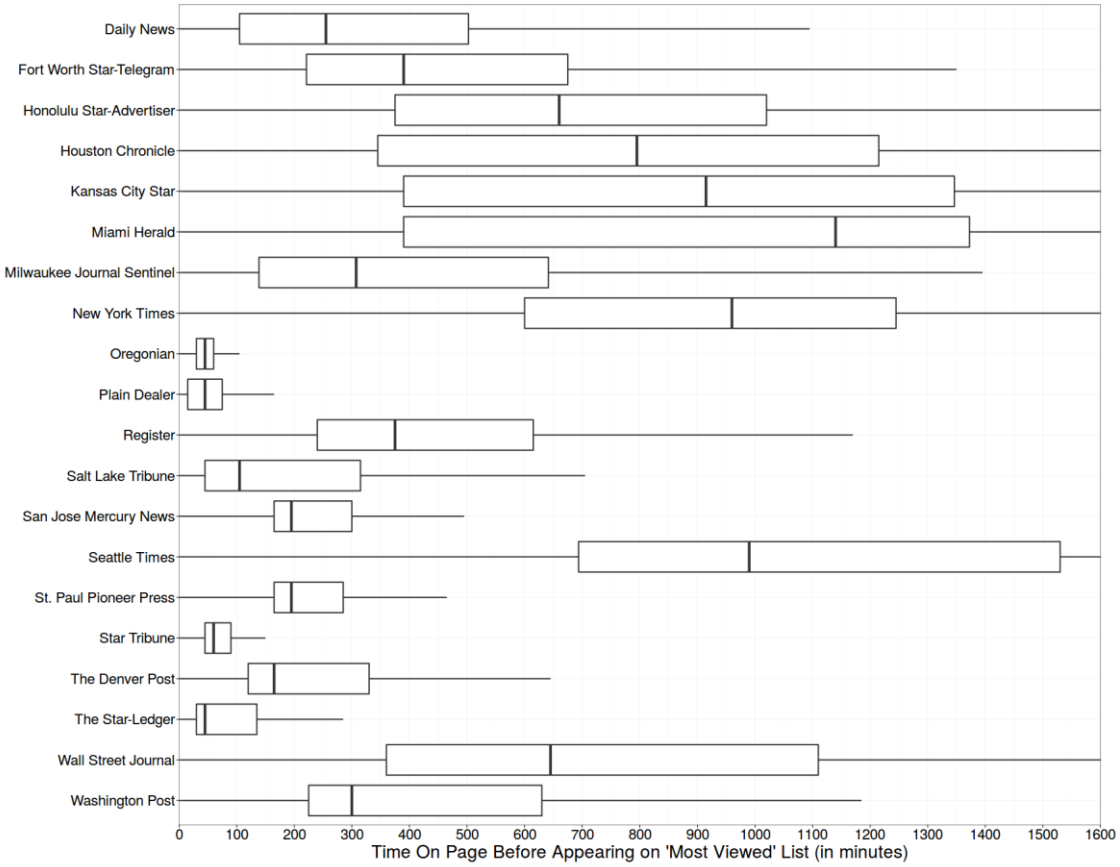

Figure 7. The amount of time it takes news items to appear on the ‘most viewed’ list.

	High Rate of Change	Low Rate of Change
High Median Time	• *Wall Street Journal*	• *Honolulu Star-Advertiser* • *Houston Chronicle* • *Kansas City Star* • *Miami Herald* • *New York Times* • *Seattle Times*
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*	• *Fort Worth Star-Telegram* • *Milwaukee Journal-Sentinel* • *Daily News* • *Register* • *Washington Post*

Figure 8. The aggregation of 'most viewed' lists into comparable clusters. Organizations in the dark, bottom-left cluster have lists that are good proxies for what is currently popular on the homepage. Organizations in the light, top-right cluster have lists that are poor proxies. Items in the mid-tone, upper-left and bottom-right clusters have lists that are of an intermediate quality.

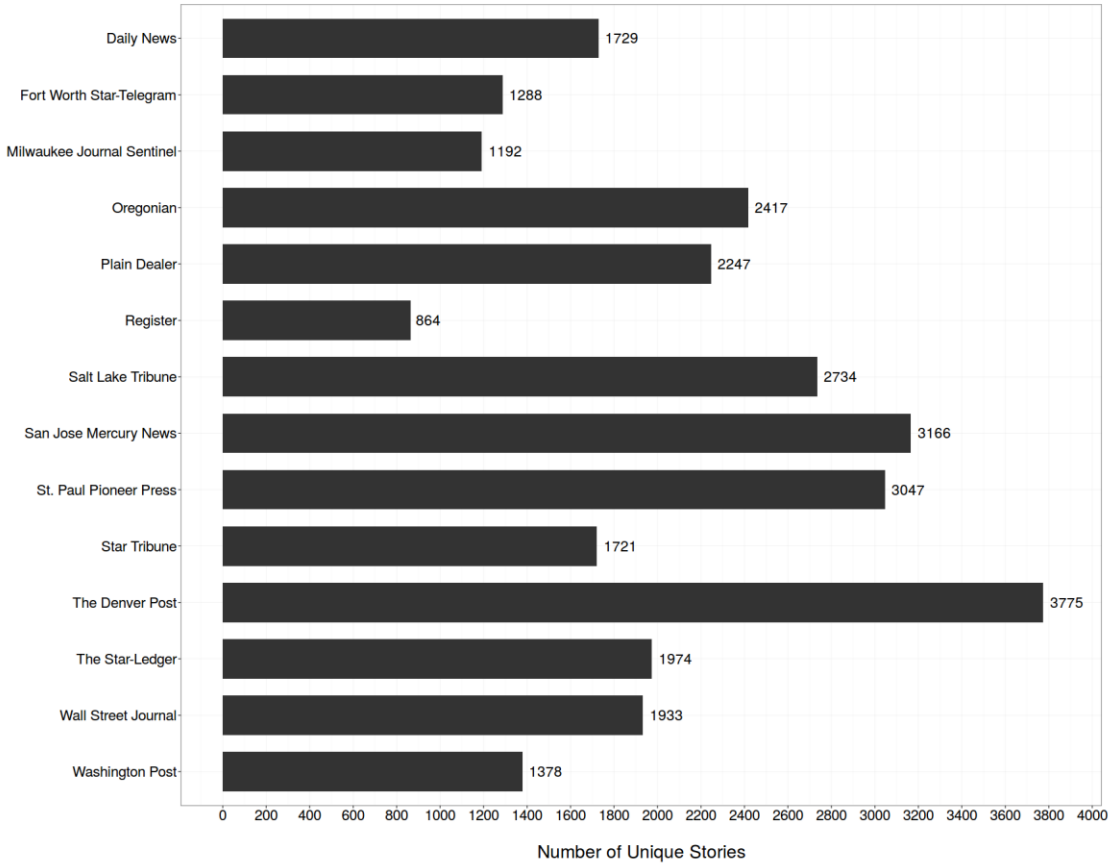

Figure 9. The number of distinct news items appearing on the top five spots of the 'most viewed' list or in one of the five top areas of prominence on the homepages of 14 news organizations between October 18, 2014, and December 20, 2014.

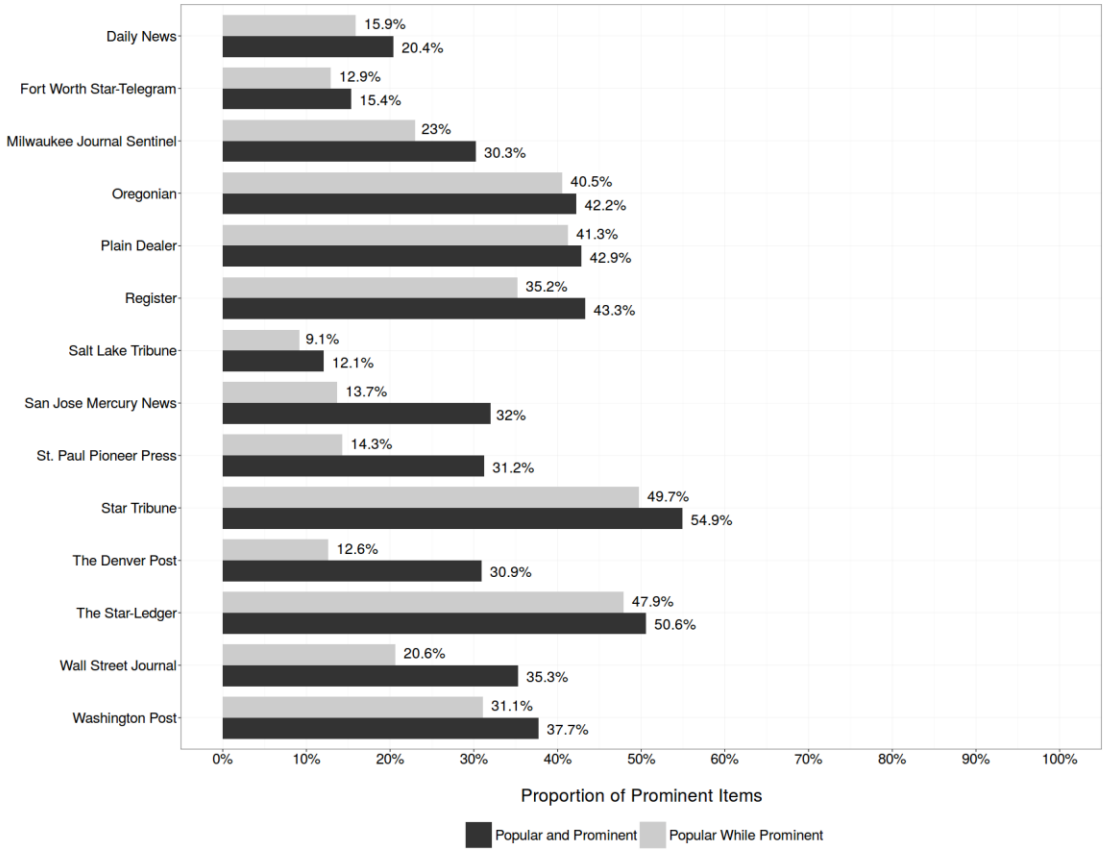

Figure 10. The proportion of prominent news items that were both popular and prominent at some point in time (dark tone) as well as popular prior to being removed from area of prominence (lighter tone).

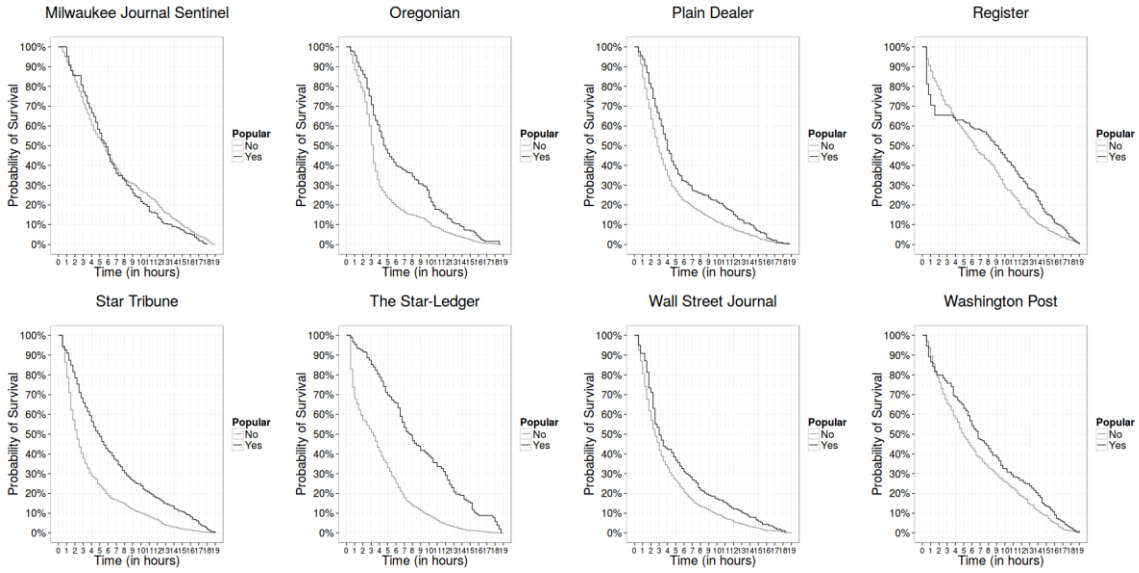

Figure 11. Fitted Cox Proportional Hazards models for eight news organizations predicting the probability a news item will remain in an area of prominence based on whether it is popular or not. The dark tone line reflects that of an item appearing on the 'most viewed' list. A mid-tone line reflects that of an item that does not appear on that list.

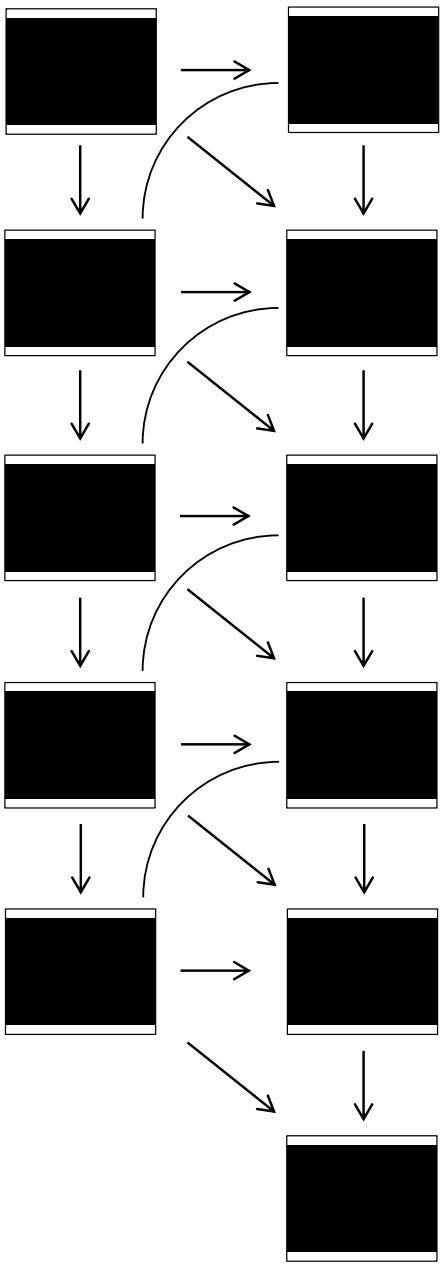

Figure 12. Path diagram illustrating the presumed relationships between an item's ranking on the 'most viewed' list and its prominence ranking, at five different points in time. For visual simplicity, the disturbances for exogenous variables are not displayed.

REFERENCES

Abbott, A. (1988). *The system of professions: An essay on the division of expert labor*. Chicago: University of Chicago Press.

Achtenhagen, L., & Raviola, E. (2009). Balancing tensions during convergence: Duality management in a newspaper company. *International Journal on Media Management*, *11*(1), 32–41. doi:10.1080/14241270802518505

Agarwal, S. D., & Barthel, M. L. (2013). The friendly barbarians: Professional norms and work routines of online journalists in the United States. *Journalism*. Advance online publication. doi:10.1177/1464884913511565

Allen, C. (2007). News directors and consultants: RTNDA's endorsement of TV journalism's "greatest tool." *Journal of Broadcasting & Electronic Media*, *51*(3), 424–437. doi:10.1080/08838150701457487

Alliance for Audited Media. (2013, June 10). Top 25 U.S. newspapers for March 2013. Retrieved May 10, 2014, from http://www.auditedmedia.com/news/research-and-data/top-25-us-newspapers-for-march-2013.aspx

Alves, R., & Weiss, A. S. (2004). Many newspaper sites still cling to once-a-day publish cycle. *Online Journalism Review*. Retrieved from http://www.ojr.org/ojr/workplace/1090395903.php

Andersen, P. K., & Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, *10*(4), 1100–1120.

Anderson, C. W. (2006). *The long tail: Why the future of business is selling less of more*. New York: Hyperion.

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

Anderson, C. W. (2011b). Deliberative, agonistic, and algorithmic audiences: Journalism's vision of its public in an age of audience transparency. *International Journal of Communication*, *5*, 1–19.

Anderson, C. W. (2013). *Rebuilding the news: Metropolitan journalism in the digital age*. Philadelphia: Temple University Press.

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

Andrejevic, M. (2013). *Infoglut: How too much information is changing the way we think and know*. New York: Routledge.

Attaway-Fink, B. (2005). Market-driven journalism: Creating special sections to meet reader interests. *Journal of Communication Management*, *9*(2), 145–154. doi:10.1108/13632540510621335

Baek, Y. M., Cappella, J. N., & Bindman, A. (2011). Automating content analysis of open-ended responses: Wordscores and affective intonation. *Communication Methods and Measures*, *5*(4), 275–296. doi:10.1080/19312458.2011.624489

Barnes, B. E., & Thomson, L. M. (1988). The impact of audience information sources on media evolution. *Journal of Advertising Research*, *28*(5), RC9–RC14.

Barnett, W. P., Greve, H. R., & Park, D. Y. (1994). An evolutionary model of organizational performance. *Strategic Management Journal*, *15*(S1), 11–28.

Barnhurst, K. G. (2011). The problem of modern time in American journalism. *KronoScope*, *11*(1/2), 98–123. doi:10.1163/156852411X595297

Batsell, J. (2015). *Engaged journalism: connecting with digitally empowered news audiences*. New York: Columbia University Press.

Beam, R. A. (1996). How perceived environmental uncertainty influences the marketing orientation of U.S. daily newspapers. *Journalism & Mass Communication Quarterly*, *73*(2), 285–303.

Beam, R. A. (2001). Does it pay to be a market-oriented daily newspaper? *Journalism & Mass Communication Quarterly*, *78*(3), 466–483. doi:10.1177/107769900107800305

Beam, R. A. (2003). Content differences between daily newspapers with strong and weak market orientations. *Journalism & Mass Communication Quarterly*, *80*(2), 368–390. doi:10.1177/107769900308000209

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom*. New Haven: Yale University Press.

Bennett, W. L., Lawrence, R. G., & Livingston, S. (2008). *When the press fails: Political power and the news media from Iraq to Katrina*. Chicago: University of Chicago Press.

Benton, J. (2015, January 12). Gawker, ever restless in restructuring its infrastructure, is stepping back from the stream. *Nieman Lab*. Retrieved February 10, 2015, from http://www.niemanlab.org/2015/01/gawker-ever-restless-in-restructuring-its-infrastructure-is-stepping-back-from-the-stream/

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Garden City: Doubleday.

Boczkowski, P. J. (2004). The processes of adopting multimedia and interactivity in three online newsrooms. *Journal of Communication*, *54*(2), 197–213. doi:10.1111/j.1460-2466.2004.tb02624.x

Boczkowski, P. J. (2005). *Digitizing the news: Innovation in online newspapers*. Cambridge: MIT Press.

Boczkowski, P. J. (2010). *News at work: Imitation in an age of information abundance*. Chicago: The University of Chicago Press.

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. doi: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: 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. doi: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. doi:10.1111/j.1460-2466.2011.01582.x

Breed, W. (1955). Newspaper "opinion leaders" and processes of standardization. *Journalism & Mass Communication Quarterly*, *32*(3), 277–328. doi:10.1177/107769905503200302

Bright, J., & Nicholls, T. (2013). The life and death of political news: Measuring the impact of the audience agenda using online data. *Social Science Computer Review*. Advance online publication. doi:10.1177/0894439313506845

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills: Sage.

Bruns, A. (2005). *Gatewatching: Collaborative online news production*. New York: Peter Lang.

Bruns, A. (2008). *Blogs, Wikipedia, Second life, and Beyond: From production to produsage*. New York: Peter Lang.

Bruns, A. (2013). Faster than the speed of print: Reconciling "big data" social media analysis and academic scholarship. *First Monday*, *18*(10). doi:10.5210/fm.v18i10.4879

Bruns, A., & Burgess, J. (2012). Researching news discussion on Twitter. *Journalism Studies*. Advance online publication. doi:10.1080/1461670X.2012.664428

Burgess, J., & Bruns, A. (2012). (Not) the Twitter election: The dynamics of the #ausvotes conversation in relation to the Australian media ecology. *Journalism Practice*, *6*(3), 384–402. doi:10.1080/17512786.2012.663610

Buzzard, K. S. (2003). James W. Seiler of the American Research Bureau. *Journal of*

Radio Studies, *10*(2), 186–201. doi:10.1207/s15506843jrs1002_4

Byrne, B. M. (1994). *Structural equation modeling with EQS and EQS/Windows*. Thousand Oaks: Sage Publications.

Carey, J. W. (1989). *Communication as culture: Essays on media and society*. Boston: Unwin Hyman.

Carey, J. W. (1993). The mass media and democracy. *Journal of International Affairs*, *47*(1), 1–21.

Christians, C. G. (1997). The common good and universal values. In J. Black (Ed.), *Mixed news: The public/civic/communitarian journalism debate* (pp. 18–35). Mahwah: Lawrence Erlbaum Associates.

Coddington, M. (2013). Defending judgment and context in "original reporting": Journalists' construction of newswork in a networked age. *Journalism*. Advance online publication. doi:10.1177/1464884913501244

Coddington, M. (2015). The wall becomes a curtain. In M. Carlson & S. C. Lewis (Eds.), *Boundaries of journalism* (pp. 67–82). New York: Routledge.

Cohen, E. L. (2002). Online journalism as market-driven journalism. *Journal of Broadcasting & Electronic Media*, *46*(4), 532–548. doi:10.1207/s15506878jobem4604_3

Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, *83*(1), 186–200. doi:10.1177/107769900608300112

Couldry, N., & Turow, J. (2014). Advertising, big data and the clearance of the public

realm: Marketers' new approaches to the content subsidy. *International Journal of Communication*, *8*, 1710–1726.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

Cranberg, G., Bezanson, R. P., & Soloski, J. (2001). *Taking stock: Journalism and the publicly-traded newspaper company*. Ames: Iowa State University Press.

Crawford, K., Gray, M. L., & Miltner, K. (2014). Critiquing big data: Politics, ethics, epistemology. *International Journal of Communication*, *8*, 1663–1672.

Croteau, D. (2006). The growth of self-produced media content and the challenge to media studies. *Critical Studies in Media Communication*, *23*(4), 340–344. doi:10.1080/07393180600933170

Croteau, D., & Hoynes, W. (2001). *The business of media: Corporate media and the public interest*. Thousand Oaks: Pine Forge Press.

De Maeyer, J. (2013). Towards a hyperlinked society: A critical review of link studies. *New Media & Society*, *15*(5), 737–751. doi:10.1177/1461444812462851

Denham, B. E. (2014). Intermedia attribute agenda setting in the New York Times: The case of animal abuse in U.S. horse racing. *Journalism & Mass Communication Quarterly*, *91*(1), 17–37. doi:10.1177/1077699013514415

Dennis, E. E., & Merrill, J. C. (1984). *Basic issues in mass communication: a debate*. New York: Macmillan.

Deuze, M. (2001). Understanding the impact of the Internet: On new media professionalism, mindsets and buzzwords. *EJournalist (On-line)*, *1*(1). Retrieved

October 13, 2014, from http://dare.uva.nl/record/99278

Deuze, M. (2003). The Web and its Journalisms: Considering the Consequences of Different Types of Newsmedia Online. *New Media & Society*, *5*(2), 203–230. doi:10.1177/1461444803005002004

Deuze, M. (2005). What is journalism? Professional identity and ideology of journalists reconsidered. *Journalism*, *6*(4), 442–464. doi:10.1177/1464884905056815

Deuze, M. (2007). *Media Work*. Cambridge: Polity Press.

Deuze, M. (2008). The changing context of news work: Liquid journalism for a monitorial citizenry. *International Journal of Communication*, *2*, 848–865.

DeWerth-Pallmeyer, D. (1997). *The audience in news*. Mahwah: Lawrence Erlbaum Associates.

Dick, M. (2013). Interactive infographics and news values. *Digital Journalism*, *2*(4), 490–506. doi:10.1080/21670811.2013.841368

Doctor, K. (2010). *Newsonomics: Twelve new trends that will shape the news you get*. New York: St. Martin's Press.

Engebretsen, M. (2006). Shallow and static or deep and dynamic? *NORDICOM Review*, *27*(1), 3–16.

Esper, T. L., Ellinger, A. E., Stank, T. P., Flint, D. J., & Moon, M. (2010). Demand and supply integration: A conceptual framework of value creation through knowledge management. *Journal of the Academy of Marketing Science*, *38*(1), 5–18. doi:10.1007/s11747-009-0135-3

Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere:

Politics and dissent online. *New Media & Society*, *12*(8), 1225–1243. doi:10.1177/1461444810385096

Ettlie, J. E., & Reza, E. M. (1992). Organizational integration and process innovation. *The Academy of Management Journal*, *35*(4), 795–827. doi:10.2307/256316

Fengler, S., & Ruß-Mohl, S. (2008). Journalists and the information-attention markets: Towards an economic theory of journalism. *Journalism*, *9*(6), 667–690. doi:10.1177/1464884908096240

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2013). Research methods in the age of digital journalism. *Digital Journalism*, *1*(1), 102–116. doi:10.1080/21670811.2012.714928

Freidson, E. (2001). *Professionalism, the third logic: On the practice of knowledge*. Cambridge: Polity Press.

Gade, P. J. (2004). Newspapers and organizational development: Management and journalist perceptions of newsroom cultural change. *Journalism & Communication Monographs*, *6*(1), 3–55. doi:10.1177/152263790400600101

Gade, P. J. (2009a). Integration of news and news of integration: A structural perspective on news media changes. *Journal of Media Business Studies*, *6*(1), 87–111.

Gade, P. J. (2009b). The structural integration of news media organizations: Opening the processes of journalism beyond the newsroom. Presented at the Future of Journalism Conference, Cardiff, UK. Retrieved from http://schools.caerdydd.ac.uk/jomec/resources/foj2009/foj2009-Gade.pdf

Gans, H. J. (1979). *Deciding what's news: A study of CBS Evening News, NBC Nightly*

News, Newsweek, and TIME. New York: Pantheon Books.

Gant, S. E. (2007). *We're all journalists now: The transformation of the press and reshaping of the law in the Internet age*. New York: Free Press.

García Avilés, J. A., León, B., Sanders, K., & Harrison, J. (2004). Journalists at digital television newsrooms in Britain and Spain: Workflow and multi-skilling in a competitive environment. *Journalism Studies*, *5*(1), 87–100. doi:10.1080/1461670032000174765

Gieber, W. (1956). Across the desk: A study of 16 telegraph editors. *Journalism & Mass Communication Quarterly*, *33*(4), 423–432. doi:10.1177/107769905603300401

Gillespie, T., Boczkowski, P. J., & Foot, K. A. (2014). *Media technologies: Essays on communication, materiality, and society*. Cambridge: MIT Press.

Gillin, P. (2007, March 5). Newspaper death watch. Retrieved May 12, 2014, from http://newspaperdeathwatch.com/

Golding, P., & Elliott, P. R. C. (1979). *Making the news*. London: Longman.

Grant, R. M. (1996). Prospering in dynamically-competitive environments: Organizational capability as knowledge integration. *Organization Science*, *7*(4), 375–387. doi:10.1287/orsc.7.4.375

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

Green, M. (2002). Mobilising readers: Newspapers, copy-tasters, and readerships. In M. Balnaves, T. O'Regan, & J. Sternberg (Eds.), *Mobilising the audience* (pp. 213–234). St. Lucia: University of Queensland Press.

Greenwood, E. (1957). Attributes of a profession. *Social Work*, *2*(3), 45–55. doi:10.1093/sw/2.3.45

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1–35. doi:10.1093/pan/mpp034

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

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.

Habermas, J. (1989). *The structural transformation of the public sphere: An inquiry into a category of bourgeois society*. Cambridge: MIT Press.

Hampton, M. (2010). The fourth estate ideal in journalism history. In S. Allan (Ed.), *The Routledge companion to news and journalism* (pp. 3–12). New York: Routledge.

Harris, L. C. (1998). Cultural domination: The key to market-oriented culture? *European Journal of Marketing*, *32*(3/4), 354–373. doi:10.1108/03090569810204643

Henry, E. (2012, May 23). 10 reasons why online journalists are better journalists (in theory). *Online Journalism Review*. Retrieved from http://www.ojr.org/p2073/

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.

doi:10.1111/jcc4.12074

Herring, S. C. (2010). Web content analysis: Expanding the paradigm. In J. Hunsinger, L. Klastrup, & M. Allen (Eds.), *International handbook of internet research* (pp. 233–249). New York: Springer.

Hesse, B. W., Moser, R. P., & Riley, W. T. (2015). From big data to knowledge in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 16–32. doi:10.1177/0002716215570007

Hindman, M. (forthcoming). Journalism ethics and digital audience data.

Holcomb, J., Boyles, J. L., Guskin, E., Jurkowitz, M., Matsa, K.-E., & Mitchell, A. (2014). *The changing revenue picture for American journalism*. Washington, D.C.: Pew Research Center. Retrieved from http://www.journalism.org/2014/03/26/the-revenue-picture-for-american-journalism-and-how-it-is-changing/

Hu, L., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks: Sage.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. doi:10.1080/10705519909540118

Jenkins, H. (2006). *Convergence culture: Where old and new media collide*. New York: New York University Press.

Kaplan, R. L. (2006). The news about new institutionalism: Journalism's ethic of objectivity and its political origins. *Political Communication*, *23*(2), 173–185. doi:10.1080/10584600600629737

Karlsson, M. (2011). The immediacy of online news, the visibility of journalistic processes and a restructuring of journalistic authority. *Journalism*, *12*(3), 279–295. doi:10.1177/1464884910388223

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

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. doi:10.1080/14616700903119784

Karpf, D. (2012). Social science research methods in Internet time. *Information, Communication & Society*, *15*(5), 639–661. doi:10.1080/1369118X.2012.665468

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

Kiousis, S., Kim, S.-Y., McDewitt, M., & Ostrowksi, A. (2009). Competing for attention: Information subsidy influence in agenda building during election campaigns. *Journalism & Mass Communication Quarterly*, *86*(3), 545–562.

Kirilenko, A. P., & Stepchenkova, S. O. (2012). Climate change discourse in mass media: Application of computer-assisted content analysis. *Journal of Environmental Studies and Sciences*, *2*(2), 178–191. doi:10.1007/s13412-012-0074-z

Koster, F., Stokman, F., Hodson, R., & Sanders, K. (2007). Solidarity through networks: The effects of task and informal interdependence on cooperation within teams. *Employee Relations*, *29*(2), 117–137. doi:10.1108/01425450710719978

Kreuter, F., Presser, S., & Tourangeau, R. (2008). Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly*, *72*(5), 847–865. doi:10.1093/poq/nfn063

Krippendorff, K. (2013). *Content analysis: An introduction to its methodology*. Thousand Oaks: Sage Publications.

Lee, A. M., & Chyi, H. I. (2014). When newsworthy is not noteworthy. *Journalism Studies*. Advance Online Publication. doi: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. doi:10.1177/0093650212467031

Lee Plaisance, P. (2005). The mass media as discursive network: Building on the implications of Libertarian and Communitarian claims for news media ethics theory. *Communication Theory*, *15*(3), 292–313. doi:10.1111/j.1468-2885.2005.tb00337.x

Leetaru, K. H. (2012). *Data mining methods for the content analyst: An introduction to the computational analysis of content*. New York: Routledge.

Leonardi, P. M. (2009). Crossing the implementation line: The mutual constitution of technology and organizing across development and use activities. *Communication Theory*, *19*(3), 278–310. doi:10.1111/j.1468-2885.2009.01344.x

Lewin, K. (1947). Frontiers in group dynamics: II. Channels of group life; social planning and action research. *Human Relations*, *1*(2), 143–153. doi:10.1177/001872674700100201

Lewis, S. C. (2012). The tension between professional control and open participation. *Information, Communication & Society*, *15*(6), 836–866. doi:10.1080/1369118X.2012.674150

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. doi:10.1080/08838151.2012.761702

Lim, J. (2010). Convergence of attention and prominence dimensions of salience among major online newspapers. *Journal of Computer-Mediated Communication*, *15*(2), 293–313. doi:10.1111/j.1083-6101.2010.01521.x

Lim, J. (2012). The mythological status of the immediacy of the most important online news. *Journalism Studies*, *13*(1), 71–89. doi:10.1080/1461670X.2011.605596

Lin, J. (2015). On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 33–47. doi:10.1177/0002716215569174

Lowrey, W. (2006). Mapping the journalism-blogging relationship. *Journalism*, *7*(4), 477–500. doi:10.1177/1464884906068363

Lowrey, W. (2011). Institutionalism, news organizations and innovation. *Journalism Studies*, *12*(1), 64–79. doi:10.1080/1461670X.2010.511954

Lowrey, W., & Gade, P. J. (2011). *Changing the news: The forces shaping journalism in uncertain times*. New York: Routledge.

Lowrey, W., & Woo, C. W. (2010). The news organization in uncertain times: Business or institution? *Journalism & Mass Communication Quarterly*, *87*(1), 41–61. doi:10.1177/107769901008700103

Luther, C. A., & Radovic, I. (2014). Newspapers frame Julian Assange differently. *Newspaper Research Journal*, *35*(1), 64–81.

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

Madden, M., Fox, S., Smith, A., & Vitak, J. (2007). *Digital footprints: Online identity management and search in the age of transparency*. Washington, D.C.: Pew Internet & American Life Project. Retrieved October 18, 2014, from http://www.pewinternet.org/2007/12/16/digital-footprints/

Mahrt, M., & Scharkow, M. (2013). The value of big data in digital media research. *Journal of Broadcasting & Electronic Media*, *57*(1), 20–33. doi:10.1080/08838151.2012.761700

Martin, H. J. (1998). Measuring newspaper profits: Developing a standard of comparison. *Journalism & Mass Communication Quarterly*, *75*(3), 500–517. doi:10.1177/107769909807500306

Massey, B. L., & Levy, M. R. (1999). 'Interactive' online journalism at English-language web newspapers in Asia: A dependency theory analysis. *International Communication Gazette*, *61*(6), 523–538. doi:10.1177/0016549299061006005

Matheson, D. (2004). Weblogs and the epistemology of the news: Some trends in online journalism. *New Media & Society*, *6*(4), 443–468. doi:10.1177/146144804044329

Mayer, J. (2011, July 7). 2011 journalist engagement survey. *Reynolds Journalism Institute*. Retrieved June 3, 2014, from http://rjionline.org/news/2011-journalist-engagement-survey

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *The Public Opinion Quarterly*, *36*(2), 176–187.

McGahan, A. M., & Porter, M. E. (1997). How much does industry matter, really? *Strategic Management Journal*, *18*(S1), 15–30.

McManus, J. H. (1994). *Market-driven journalism: Let the citizen beware?* Thousand Oaks: Sage Publications.

McNelly, J. T. (1959). Intermediary communicators in the international flow of news. *Journalism & Mass Communication Quarterly*, *36*(1), 23–26. doi:10.1177/107769905903600103

Meyer, A. D., Tsui, A. S., & Hinings, C. R. (1993). Configurational approaches to organizational analysis. *Academy of Management Journal*, *36*(6), 1175–1195. doi:10.2307/256809

Mitchell, A., Olmstead, K., Jurkowitz, M., Matsa, K.-E., Anderson, M., Boyles, J. L., … Holcomb, J. (2014). *Key indicators*. Washington, D.C.: Pew Research Center. Retrieved October 15, 2014, from http://www.journalism.org/2014/03/26/state-of-the-news-media-2014-key-indicators-in-media-and-news/

Mitchelstein, E., & Boczkowski, P. J. (2009). Between tradition and change: A review of

recent research on online news production. *Journalism*, *10*(5), 562–586. doi:10.1177/1464884909106533

Mullarkey, G. W. (2004). Internet measurement data—practical and technical issues. *Marketing Intelligence & Planning*, *22*(1), 42–58. doi:10.1108/02634500410516904

Napoli, P. M. (2003). *Audience economics: Media institutions and the audience marketplace*. New York: Columbia University Press.

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

Nguyen, A. (2013). Online news audiences: The challenges of web metrics. In S. Allan & K. Fowler-Watt (Eds.), *Journalism: New challenges* (pp. 146–161). Poole: Centre for Journalism & Communication Research, Bournemouth University.

Nooteboom, B. (2000). *Learning and innovation in organizations and economies*. Oxford: Oxford 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. doi:10.1287/orsc.11.4.404.14600

Pavlik, J. V. (2000). The impact of technology on journalism. *Journalism Studies*, *1*(2), 229–237. doi:10.1080/14616700050028226

Pavlik, J. V. (2001). *Journalism and new media*. New York: Columbia University Press.

Petre, C. (2015). *The Traffic Factories: Metrics at Chartbeat, Gawker Media, and The New York Times*. New York City: The Tow Center for Digital Journalism.

Retrieved from http://towcenter.org/research/traffic-factories/

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441. doi:10.1177/030631284014003004

Polanyi, M. (1966). *The tacit dimension*. Garden City: Doubleday.

Pool, I. de S., & Shulman, I. (1959). Newsmen's fantasies, audiences, and newswriting. *The Public Opinion Quarterly*, *23*(2), 145–158.

Raisch, S., & Birkinshaw, J. (2008). Organizational ambidexterity: Antecedents, outcomes, and moderators. *Journal of Management*, *34*(3), 375–409. doi:10.1177/0149206308316058

Raviola, E., & Hartmann, B. (2009). Business perspectives on work in news organizations. *Journal of Media Business Studies*, *6*(1), 7–36.

Reich, Z. (2013). The impact of technology on news reporting: A longitudinal perspective. *Journalism & Mass Communication Quarterly*, *90*(3), 417–434. doi:10.1177/1077699013493789

Rhodes, J., Hung, R., Lok, P., Lien, B. Y.-H., & Wu, C.-M. (2008). Factors influencing organizational knowledge transfer: Implication for corporate performance. *Journal of Knowledge Management*, *12*(3), 84–100. doi:10.1108/13673270810875886

Riffe, D., Lacy, S. R., & Fico, F. G. (2014). *Analyzing media messages: Using quantitative content analysis in research* (3rd ed.). New York: Routledge.

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.

Schlesinger, P. (1978). *Putting "reality" Together: BBC News*. London: Methuen.

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schudson, M. (1978). *Discovering the news: A social history of American newspapers*. New York: Basic Books.

Schudson, M. (1995). *The power of news*. Cambridge: Harvard University Press.

Schudson, M. (2001). The objectivity norm in American journalism. *Journalism*, *2*(2), 149–170. doi:10.1177/146488490100200201

Schudson, M. (2003). *The sociology of news*. New York: W. W. Norton.

Schwartz, H. A., & Ungar, L. H. (2015). Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 78–94. doi:10.1177/0002716215569197

Shah, D. V., Cappella, J. N., & Neuman, W. R. (2015). Big data, digital media, and computational social science possibilities and perils. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 6–13. doi:10.1177/0002716215572084

Shah, D. V., McLeod, D. M., Gotlieb, M. R., & Lee, N. (2009). Framing and agenda setting. In R. L. Nabi & M. B. Oliver (Eds.), *Sage handbook of media processes*

and effects (pp. 83–98). London: Sage Publications.

Shirky, C. (2008). *Here comes everybody: How digital networks transform our ability to gather and cooperate*. New York: Penguin Press.

Shoemaker, P. J. (1991). *Gatekeeping*. Newbury Park: Sage Publications.

Shoemaker, P. J., Danielian, L. H., & Brendlinger, N. (1991). Deviant acts, risky business and U.S. interests: The newsworthiness of world events. *Journalism & Mass Communication Quarterly*, *68*(4), 781–795. doi:10.1177/107769909106800419

Shoemaker, P. J., Johnson, P. R., Seo, H., & Wang, X. (2010). Readers as gatekeepers of online news: Brazil, China, and the United States. *Brazilian Journalism Research*, *6*(1), 55–77.

Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. New York: Routledge.

Siebert, F. S., Peterson, T., & Schramm, W. (1956). *Four theories of the press: The Authoritarian, Libertarian, Social Responsibility, and Soviet Communist concepts of what the press should be and do*. Urbana: University of Illinois Press.

Siles, I., & Boczkowski, P. (2012). At the intersection of content and materiality: A texto-material perspective on the use of media technologies. *Communication Theory*, *22*(3), 227–249. doi:10.1111/j.1468-2885.2012.01408.x

Singer, J. B. (1997). Still guarding the gate? The newspaper journalist's role in an on-line world. *Convergence: The International Journal of Research into New Media Technologies*, *3*(1), 72–89. doi:10.1177/135485659700300106

Singer, J. B. (1998). Online journalists: Foundations for research into their changing roles. *Journal of Computer-Mediated Communication*, *4*(1). doi:10.1111/j.1083-

6101.1998.tb00088.x

Singer, J. B. (2005). The political j-blogger: "Normalizing" a new media form to fit old norms and practices. *Journalism*, *6*(2), 173–198. doi:10.1177/1464884905051009

Singer, J. B. (2006). Stepping back from the gate: Online newspaper editors and the co-production of content in Campaign 2004. *Journalism & Mass Communication Quarterly*, *83*(2), 265–280. doi:10.1177/107769900608300203

Singer, J. B. (2011). Community service: Editor pride and user preference on local newspaper websites. *Journalism Practice*, *5*(6), 623–642. doi:10.1080/17512786.2011.601938

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. doi: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. doi:10.1177/1354856511429641

Sjøvaag, H., Stavelin, E., & Moe, H. (2015). Continuity and change in public service news online. *Journalism Studies*. Advance Online Publication. doi:10.1080/1461670X.2015.1022204

Smit, G., de Haan, Y., & Buijs, L. (2014). Visualizing news. *Digital Journalism*, *2*(3), 344–354. doi:10.1080/21670811.2014.897847

Snider, P. B. (1967). "Mr. Gates" revisited: A 1966 version of the 1949 case study.

Journalism & Mass Communication Quarterly, *44*(3), 419–427. doi:10.1177/107769906704400301

Snow, C. C., Miles, R. E., & Miles, G. (2005). A configurational approach to the integration of strategy and organization research. *Strategic Organization*, *3*(4), 431–439. doi:10.1177/1476127005057965

Soberman, J. (2013, August 15). Designing from data: How news organizations use A/B testing to increase user engagement. *Knight Lab*. Retrieved February 18, 2015, from http://knightlab.northwestern.edu/2013/08/15/designing-from-data-how-news-organizations-use-ab-testing-to-increase-user-engagement/

Soloski, J. (2013). Collapse of the US newspaper industry: Goodwill, leverage and bankruptcy. *Journalism*, *14*(3), 309–329. doi:10.1177/1464884912472016

Stepp, C. S. (2000, August). Reader friendly: Their futures uncertain, newspapers are undergoing a profound change in the way they carry out their missions. *American Journalism Review*, *22*(6), 22–43.

Steuer, J. (1992). Defining virtual reality: Dimensions determining telepresence. *Journal of Communication*, *42*(4), 73–93. doi:10.1111/j.1460-2466.1992.tb00812.x

Stuart, A. (2010). *News Culture*. McGraw-Hill International.

Tandoc, E. C. (2014a). Journalism is twerking? How web analytics is changing the process of gatekeeping. *New Media & Society*, 1461444814530541. doi:10.1177/1461444814530541

Tandoc, E. C. (2014b). Why web analytics click: Factors affecting the ways journalists use audience metrics. *Journalism Studies*. Advance Online Publication.

doi:10.1080/1461670X.2014.946309

Tandoc, E. C., & Thomas, R. J. (2015). The ethics of web analytics. *Digital Journalism*, *3*(2), 243–258. doi:10.1080/21670811.2014.909122

Tang, Y. (Elina), Sridhar, S. (Hari), Thorson, E., & Mantrala, M. K. (2011). The bricks that build the clicks: Newsroom investments and newspaper online performance. *International Journal on Media Management*, *13*(2), 107–128. doi:10.1080/14241277.2011.568420

Tenenboim, O., & Cohen, A. A. (2013). What prompts users to click and comment: A longitudinal study of online news. *Journalism*, 1464884913513996. doi:10.1177/1464884913513996

The New York Times Company. (2014). *Innovation*. New York: The New York Times.

Thorson, E. (2008). Changing patterns of news consumption and participation. *Information, Communication & Society*, *11*(4), 473–489. doi:10.1080/13691180801999027

Tremayne, M., Weiss, A. S., & Alves, R. C. (2007). From product to service: The diffusion of dynamic content in online newspapers. *Journalism & Mass Communication Quarterly*, *84*(4), 825–839. doi:10.1177/107769900708400411

Tuchman, G. (1978). *Making news: A study in the construction of reality*. New York: Free Press.

Tumber, H. (2001). Democracy in the information age: The role of the Fourth Estate in cyberspace. *Information, Communication & Society*, *4*(1), 95–112. doi:10.1080/13691180122542

Tunstall, J. (1971). *Journalists at work.* London: Constable.

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. doi:10.1177/0002716204270469

Tushman, M., & Romanelli, E. (1985). Organizational evolution: A metamorphosis model of convergence and reorientation. *Research in Organizational Behavior*, *7*, 171–222.

Underwood, D. (1993). *When MBAs rule the newsroom: How the marketers and managers are reshaping today's media*. New York: Columbia University Press.

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. doi:10.1080/21670811.2013.801690

Ven, A. H. V. D., Delbecq, A. L., & Koenig, R., Jr. (1976). Determinants of coordination modes within organizations. *American Sociological Review*, *41*(2), 322–338. doi:10.2307/2094477

Vu, H. T. (2014). The online audience as gatekeeper: The influence of reader metrics on news editorial selection. *Journalism*, *15*(8), 1094–1110. doi:10.1177/1464884913504259

Wanta, W. (1997). *The public and the national agenda: How people learn about important issues*. Mahwah, NJ: Lawrence Erlbaum Associates.

Weaver, D. H., Beam, R. A., Brownlee, B. J., Voakes, P. S., & Wilhoit, G. C. (2007). *The*

American journalist in the 21st Century: U.S. news people at the dawn of a new millennium. Mahwah: Lawrence Erlbaum Associates.

Webster, J. G. (2008). Developments in audience measurement and research. In B. J. Calder (Ed.), *Kellogg on advertising & media: The Kellogg School of Management* (pp. 123–138). Hoboken: John Wiley & Sons.

Webster, J. G., Phalen, P. F., & Lichty, L. W. (2000). *Ratings analysis the theory and practice of audience research*. Mahwah: Lawrence Erlbaum Associates.

Westley, B. H., & MacLean, M. S. (1957). A conceptual model for communications research. *Journalism & Mass Communication Quarterly*, *34*(1), 31–38. doi:10.1177/107769905703400103

White, D. M. (1950). The "Gate Keeper": A case study in the selection of news. *Journalism Quarterly*, *27*(4), 383–391.

Williams, R., & Edge, D. (1996). The social shaping of technology. *Research Policy*, *25*(6), 865–899. doi:10.1016/0048-7333(96)00885-2

Witschge, T., & Nygren, G. (2009). Journalism: A profession under pressure? *Journal of Media Business Studies*, *6*(1), 37–59.

Xu, K. (2013). Framing Occupy Wall Street: A content analysis of The New York Times and USA Today. *International Journal of Communication*, *7*, 21.

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. doi:10.1177/0002716215570576