NORTHWESTERN UNIVERSITY

# Computational Study of News Systems: Embracing the Complexity Paradigm

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Communication Studies

By

Nick Hagar

EVANSTON, ILLINOIS

June 2023

## Abstract

Computational Study of News Systems: Embracing the Complexity Paradigm

Nick Hagar

The production and spread of digital news involves a wide range of actors: journalists and the organizations that employ them, social media platforms, audiences, and myriad commentators, citizen journalists, bloggers, and other actors who contribute to the news ecosystem without inhabiting an official role. These actors interact in flexible, often unexpected ways. Because of the range of actors involved, the dynamic nature of their activity, and the ways in which they interact, the disparate parts constituting digital news media are difficult to encapsulate under one unified framework.

In this work, I argue for an approach to studying news media at the system level. Building on existing theories of digital journalism, I advocate for a paradigm that embraces the networked nature of news. In this view, it is counterproductive to examine news actors or processes in isolation, as it is their connections to other parts of the media system that make them consequential. To motivate this view, I draw on concepts from the study of complex systems, with a particular focus on interconnectedness and emergence. I demonstrate how these concepts can be integrated with ongoing theoretical developments in the field of communication, providing a generative toolset for understanding the complexity of digital news.

This dissertation includes four studies that embrace this complexity-oriented paradigm in empirical settings. In each, I identify a subset of actors and the relationships among them, then demonstrate how those relationships impact other parts of the larger media system. Taken together, these

studies offer insight into the myriad indirect influences that shape news production, distribution, and consumption. They also offer guidance on potential strategies for navigating unpredictability in empirical processes. Finally, they outline key methodological considerations for the large-scale study of digital systems. I end by discussing implications of this work and its potential applications in future research.

## Acknowledgments

I'm grateful to be surrounded by such a generous, talented academic community. I'm thankful to my advisor, Dr. Nicholas Diakopoulos, for years of guidance and collaboration, and for building a space where this kind of research is possible. My committee members, Drs. Emőke-Ágnes Horvát Horvat and Aaron Shaw, have been instrumental in shaping how I think about research. I received invaluable feedback on parts of this dissertation from so many sources: members of the Computational Journalism Lab and the Community Data Science Collective; Drs. James Webster, Joshua Becker, and Harsh Taneja; the data science team at Patreon; Drs. Laila Wahedi and Eric Dunford at Meta; and many others. This work is a collective effort, and I owe thanks to everyone who helped hone these projects.

Finally, I'm thankful to my family for their support over this long journey. My parents ingrained in me the curiosity to do this work, and the confidence to see it through. And my partner Geordan has been endlessly supportive of my chosen path, every step of the way.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1
# INTRODUCTION AND BACKGROUND

In a study of the local news ecosystem in Philadelphia, C. Anderson (2010) traces the dissemination of one news story, about the eviction and arrest of a group of homeowners, during one week in June 2008. This study traces how the story traveled—from a press release, through independent blogs and local newsrooms, and ultimately into the national spotlight—and what decisions shaped its presentation. The painstaking analysis represented an intensive effort to understand the pathways of information circulation in an increasingly fractured media environment, among a wide array of news actors.

This study leveraged conceptual and methodological approaches to digital media research that, while not novel on their own, pointed toward a new approach in combination. First, it cast a wide net when considering which news actors to include. C. Anderson (2010) argues that "Research on changing patterns of newswork... has focused on news production within institutions rather than the circulation of news in ecosystems" (p. 291). By adopting an ecosystem-level view, the study can highlight the role that activists and bloggers played in shaping the flow of news. This allows Anderson to illustrate the contextual fluidity of news: Journalists produce journalism, but so do a slew of other actors for whom a given story is important.

Second, it considered the ways in which those producers interacted in their decision making processes. Relationships are at the forefront of how this story evolved, as actors assessed each others' credibility and motivations. Within newsrooms, conversations around intended audience, brand awareness, and available reporting resources steered the shape of coverage. These interactions create a backdrop to the published reporting, one rich with unseen motivations, that actively

shapes coverage. Without that backdrop, the study argues, our understanding of how the story circulates is severely impoverished.

Finally, this study weaved the relationships among actors together using a network paradigm. Anderson sketches a *structure* of local news coverage from the interactions he observes, and he further situates actors' *positions* within that structure. While not grounded in formal network analysis, this study explores the roles actors take—those of seeding information, or of bridging between parts of the system—and how they contribute to the diffusion process at work in a story's spread. And while this structure is well defined, it also formed in a particular fashion because of the actions of activists and reporters. It was complex and contextual, "categorized and recategorized by these actors differently, depending on their own position within the media system" (C. Anderson, 2010, p. 306).

This case study took an approach to studying digital news systems that adopts a broad definition of news producers, that considers how those actors' interactions shape information flow, and that builds an underlying structure out of those interactions. It also involved intensive data collection and analysis on a local scale. C. Anderson (2010) combined hyperlink crawling, 60 semistructured interviews, and over 300 hours of newsroom observation to study one particular set of stories, within a subset of a particular city's media environment, over a narrow time window. Scaling such an effort to larger media systems is infeasible. Fortunately, researchers have made methodological and theoretical advancements toward easing systemic study of digital media. On the methodological front, this innovation took the form of novel open source tools, adaptations of statistical methods into social science applications, and open datasets. On the theoretical front, researchers have moved steadily toward a "news-as-system" view, one that encompasses a broad range of actors and relationships into our understanding of news phenomena.

## 1.1 Computational methods in digital media

Much of our ability to encapsulate the kind of news system being explored here stems from an embrace of computational methods in the social sciences. Under the label of "computational social science", Lazer et al. (2009) argue for a study of social phenomena that leverage large datasets, with a particular focus on unraveling social structures. From this view, unprecedented access to digital trace data would elevate our understanding of human behavior by way of new insight into social activities. This pursuit of so-called "Big Data" for its own sake has been met with criticism, for its lack of social and political nuance (boyd & Crawford, 2012). However, the combination of accessible digital trace data with computational analysis techniques has proven fruitful in studying social systems. As Hindman (2015) argues,

> One overlooked story about big data's impact, though, has to do with new methods of data analysis. Developed explicitly for big data problems, this new research goes by many different labels: machine learning, data mining, statistical learning, applied mathematics, data science... But whichever label one uses, these new techniques allow social scientists to be better at quantifying uncertainty and making accurate predictions, even—or rather especially—when their datasets are small.

This construction gets at a fuller picture of computational social science's benefits: Samples of data representing social phenomena, paired with novel analysis techniques and powered by abundant computational resources, allow researchers to generate more representative models than traditional approaches. Critical to this view of social science research is access: to data, to cutting-edge methods, and to computing resources (Lazer et al., 2009). There is, at present, no shortage of such access for communication researchers, in the form of academic initiatives, open-source tools, and corporate offerings. Increased access to large-scale digital datasets allows researchers to col-

lect fine-grained artifacts like social media posts (Bruns & Burgess, 2012) or news articles (Roberts et al., 2021; Schrodt, 2010). Because of the increasing availability of computing resources, we can automatically process some aspects of these artifacts with tools like natural language processing, sentiment analysis, or automated image recognition (Günther & Quandt, 2018; Hutto & Gilbert, 2014; Joo & Steinert-Threlkeld, 2022). With those characteristics processed and standardized, we can then model underlying relationships and structures to quantify their impact on some process of interest (Margolin, 2019). These tools, argue van Atteveldt et al. (2019), help to address several challenges in communication research: providing external and ecological validity, overcoming concerns about internal validity in observational study, and improving statistical power.

In the context of digital journalism, researchers have demonstrated these benefits in the study of a variety of news phenomena. In the realm of news production, natural language processing has proven valuable in automatically parsing headlines and articles. Researchers use this processing capability to examine news values, framing, and the semantics of headline construction at scale (di Buono et al., 2017; Kuiken et al., 2017). In examining distribution, access to large-scale social media and news publisher datasets enables detailed examination of how news travels (Buhl et al., 2019) and, coupled with computational analysis techniques, the structure of dissemination (Castaldo et al., 2022; Kaiser et al., 2019). And in studying consumption, researchers are able to more expressively model the reading habits of news audiences (Makhortykh et al., 2021; Vermeer et al., 2020).

This list is not comprehensive, but it demonstrates the extent to which computational approaches have permeated studies of digital journalism. The work in this dissertation continues in this vein, leveraging a mix of large-scale digital data collection, machine learning-driven data processing, and modeling and simulation techniques to scrutinize digital media actors. However, while many of the studies mentioned here focus on particular facets of digital news, this disserta-

tion aims to take a system-level view of a broader set of media actors. To motivate this position, we require a system-level theoretical framework.

## 1.2 Toward a system-level theory of news

Many traditional theories of journalism center on the role of news organizations, and of the newsroom. As news shifted to a largely digital enterprise, theories had to adapt to not only a change in medium, but also a radical reconfiguration of the actors at play in shaping journalism. Along with that shift came a need to encompass these many and varied actors under one framework, and to describe the ways in which their relationships blended the previously rigid distinctions among producers and consumers. Among many efforts to devise such a framework, theorists have begun to incorporate concepts from complexity science.

### 1.2.1 Theories of the newsroom

Throughout the 20th century, many theories of how news worked focused on the organizations and journalists responsible for producing it. To give one representative example, gatekeeping theory describes a process by which news takes form (White, 1950). It theorizes that journalists, editors, and organizations make decisions—what events to cover, what pieces of information to include in articles, how to frame news stories—that ultimately shape the news (Shoemaker & Vos, 2009). This view gives primacy to the newsroom as the determiner of news output. It also positioned news audiences largely as recipients of this decision making: Readers are scarcely mentioned in White (1950), other than as the people for whom an editor is making choices.

This example is emblematic of a paradigm seen across newsroom-centric theory of this time: The process of news happens within the newsroom, and journalism flows unidirectionally outward to audiences from there. In the theory of agenda setting, newsrooms played a pivotal role in

shaping political and social reality for their readers (McCombs & Shaw, 1972). The concept of newsworthiness shaped by Galtung and Ruge (1965) imagines the evaluation of a news story's relevance starting with journalists and moving out, in a linear chain, to other journalists, official agencies, and readers. These models made little room for non-newsroom actors.

### 1.2.2 Proliferating news actors in digital media

As news became an increasingly digital venture, and as platforms rose to power within the digital ecosystem, the role of the newsroom shifted. Many entities—citizen journalists, bloggers, social media platforms, and others—cropped up alongside traditional news organizations, playing novel roles in news production (Bruns, 2003; Lowrey, 2006). To account for this shift, novel theories of news media began to embrace a more heterogeneous set of consequential actors, and to grapple with how influence flowed among them.

Curated flows addresses news diffusion in digital environments (Thorson & Wells, 2016). It does away with any rigid division between news producer and consumer, instead focusing on an intermingled set of actors engaged in the process of *curation*. Journalists curate information into published news. Algorithms curate content into personalized feeds. Individuals curate what they encounter, via their behaviors and social networks. These flows are constantly interacting; in the absence of any one dominant curator, they create a complex underlying structure determining who sees what. In this view, our understanding of how information travels "increasingly depends on accurately mapping individuals' positions within these multiple curated flows" (Thorson & Wells, 2016, p. 310).

Researchers have deployed this framework to motivate empirical studies of digital intermediaries between news producers and consumers. These studies explore the idea of "proliferating contingencies" from Thorson and Wells (2016)—the rapidly expanding, contextually defined ways

that individual media experiences might vary. In some instances, social media firms exercise curation via product design or corporate strategy, broadly shaping encounters with news media (Kreiss & Mcgregor, 2018; Trielli & Diakopoulos, 2019a). In others, they narrowly govern news exposure via algorithmic recommendations (Thorson et al., 2021). Within the context of social media, other users also shape our news habits (Anspach, 2017). And finally, users themselves develop personal theories of how these processes unfold, and they alter their behavior in response (DeVito et al., 2018). These curating behaviors all unfold in conjunction, and they vary in importance from person to person. As such, this research highlights the complex, dynamic behaviors that make up the system of news dissemination.

At the same time, the curated flows framework de-emphasizes the processes underpinning news production. Its focus is on the processes that shape and transport news once it has taken form, not on the process of forming events, perspectives, and data into news in the first place. That production work is still consequential, and it involves its own set of interactions, priorities, and contextual decision making (Mitchelstein & Boczkowski, 2009).

To examine these news production dynamics, researchers have applied an actor-network theory lens to publishers. Actor-network theory calls for an understanding of the human actors that produce journalism, the technologies that work alongside them, and their processes (Primo & Zago, 2015). This approach has been critical in conceptualizing technology's growing role in journalism, both as a tool deployed by human journalists and as a news producer in its own right (Diakopoulos, 2019; Primo & Zago, 2015). It also invites an examination of actors outside the newsroom, allowing researchers to consider interplays among business concerns, editorial priorities, and technological advancements (Hagar & Diakopoulos, 2019; Lewis & Westlund, 2015).

Actor-network theory, then, might provide a production-focused complement to curated flows. In both cases, defining a set of actors and the interactions among them is key to understanding

news processes. Those relationships form networked structures within which processes unfold. And those structures are contextual, shifting in response to actors' priorities.

The parallels in these theoretical approaches raises a question of combination: Can we view the entirety of digital news—encompassing production, distribution, and consumption—through one systemic lens? Is it necessary, or even productive, to distinguish actors by their position in the digital media system, when their roles and behaviors are increasingly blurred (Hernández-Serrano et al., 2017)?

Theories in this vein have attempted to construct a more holistic view of digital media. Media ecology theory uses evolutionary competition to provide an analogy through which we can incorporate disparate actors. News producers compete for audiences; their success depends on the context in which they work and the strategies they employ (Scolari, 2012). Studies adopting this perspective are largely concerned with examining the interrelation between actors and environments, and how changes in one impact the behaviors of others. Using characteristics like scarcity, competition, and specialization, they chart the dynamics of a particular set of actors in context to explain observed changes. In an examination of news production by blogs, Lowrey (2012) constructs a model of outlet competition, growth, and specialization from a population view. Weber (2012) posits a model of collaboration among news outlets as construed from hyperlinks. In an examination of Swedish news media, Nygren et al. (2018) examine the extent to which hyperlocal outlets overtake legacy media. These efforts more fully consider the range of actors at play in constituting a media system, but they still often do so from the perspective of news organizations.

In contrast is work explicitly concerned with the structural underpinnings of news. Webster (2011) presents a model of audience attention shaped by individual actions within a structured environment. Consumers interact with institutional structures—the news media, large technology firms—and make rational decisions among a limited set of options. In aggregate, these straight-

forward choices shape the distributions of attention we observe across the media landscape. Similarly, Benkler et al. (2018) attempts to conceptualize the spread of partisan misinformation via a networked model of publishers, platforms, and audiences. Their study conceptualizes the media environment as a system of pathways, along which information travels. Information pathways and collective decision making are also prominent in the examination of agenda setting in Russell Neuman et al. (2014), which further focuses on the *dynamics* at play in media processes.

Of the theoretical perspectives considered, these structural studies hew closest to the paradigm seen in C. Anderson (2010). They consider heterogeneous actors with consequential connections, and they observe how those connections play out in context. Taking this perspective a step further, we can group these approaches together under a common theoretical framework, one that digital media researchers could leverage to motivate this type of work. To do so, we can turn to a parallel theoretical framework in another field, that of complex systems studies.

### 1.2.3 The complexity paradigm

Complexity is an amorphous topic (Ladyman & Wiesner, 2020), but it broadly encapsulates key characteristics of a system's function and structure: "Common to all studies on complexity are systems with multiple elements adapting or reacting to the pattern these elements create" (Arthur, 1999, p. 107). That definition encapsulates two tenets, common in descriptions provided by theorists of complexity, that are key to applying this lens to digital media systems.

First, complex systems are networked. They are characterized by interactions among their constituent parts. When aggregated, those interactions allow us to trace structural patterns of behavior: who interacts with whom in a social group, for example, and how those interactions form the basis of cliques (Simon, 1991). This network paradigm is flexible in a couple ways. It is flexible to any context in which interactions take place, as applicable to biological systems like the brain as it is to

human communication (Castellano et al., 2009; Ladyman & Wiesner, 2020). That flexibility allows researchers to transfer network models across fields of research, enabling sophisticated computational analysis of systems' structural and dynamic properties (M. E. J. Newman, 2003). Studies of human socialization, for example, leverage network topology to map out the broad structure of association within groups (Castellano et al., 2009). Network models are also flexible in the configuration of connections among components. In contrast to the rigid unidirectional association of, for example, a linear model, networks allow conceptual arrangements in which many actors might influence each other simultaneously (Sherry, 2015). Corresponding methodological tools allow researchers to analyze network structures (M. E. J. Newman, 2010), providing formal approaches through which flexible configurations of actors can be encoded.

Second, complex systems exhibit emergence:

> The starting point of complexity science is the fact that some of the behaviour of large collections can be novel, in the sense that the parts on their own, or in small numbers with small numbers of interactions, do not display it. Emergence is surprising because what will happen cannot be anticipated by thinking about the behaviour of isolated individuals or collections involving only small numbers of interactions among individuals. (Ladyman & Wiesner, 2020, p. 21)

These systems are self-organized, with regular, distinguishable patterns appearing without any top-down direction governing collective behavior (Sawyer, 2005). This characteristic often results in an overarching order within complex systems, born out of collective regularity despite seemingly erratic individual behavior (Castellano et al., 2009). As a consequence, the aggregate behaviors of a system are difficult to attribute to any singular cause, as they "bubble up" out of many disparate interactions (Waldherr et al., 2021). At the same time, though, the importance of small-scale

interaction to emergent behavior allows us to distill group behaviors down to simple rule sets, applied over a large population (Sherry, 2015).

Complex systems, then, can be broadly conceptualized as networks of interactions that exhibit distinct characteristics or behaviors at the collective level. They are often modeled via formal network methods or modeling of individual behaviors at scale. These approaches look radically different from many empirical studies of digital news phenomena, wherein the main concern is drawing a linear explanatory relationship between x and y (Sherry, 2015). They allow an expanded view of a system's dynamics, incorporating facets like non-linearity and feedback loops among components (Benkler et al., 2018; Miller & Page, 2007). They also shift the research focus away from particular instances of empirical observation and toward the processes that govern systems and the mechanisms by which they unfold (Sawyer, 2005).

Already, the complex systems approach has appeared in some theorizations of digital media. Qvortrup (2006) conceptualizes digital media as simultaneously increasing social complexity (by increasing interconnectedness) and managing it (by building systems that mitigate its impact). This view primarily focuses on media's relationship to *social* complexity and its real-world consequences. Sherry (2015) goes a step further, exploring the potential utility of complexity science for the study of communication systems. Communication, he argues, is inherently complex, and has several parallels with complex systems: Communication processes happen via interaction in groups, they operate via nonlinear feedback loops, and they are dynamic. These parallels focus on the process of human interpersonal communication in the strictest sense, but they provide on-ramps for introducing complex systems methodologies into communication studies (Sherry, 2015). Broadening this focus, Waldherr et al. (2021) argues for a widespread adaption of complexity science in computational communication study. While, they claim, empirical communication research has seen little uptake of complexity theory, many communication phenomena can be well

understood via this lens. Looking at communication systems in terms of complexity also "has profound consequences for the way scholars formulate hypotheses. Simple schemes of dependent and independent variables are not adequate, because complex systems are driven by adaptive behavior and resulting feedback loops, often leading to nonlinear outcomes." (Waldherr et al., 2021, p. 159). Waldherr et al. (2021) therefore argues for a view that reevaluates both theoretical and methodological approaches to system-level studies.

These arguments represent a growing interest in the complex systems perspective from researchers within communication studies. They take varying facets of complexity research—the network focus, the emphasis on non-linear dynamics, the emergent structures—and explore their applicability to common phenomena within communication studies.

We can extend this perspective further, more fully integrating the complexity perspective into communication theory and applying it specifically to a system-level view of digital news media. In addition to the benefits described above, doing so addresses several challenges unique to the study of digital media. First, we lack the explanatory power to account for the workings of many news processes, and we require a theoretical framework that explicitly addresses unpredictability in its conceptualization. This is evident in news consumption, where the drivers for a person's decision to click on an article are often unclear (Arapakis et al., 2017; Kormelink & Meijer, 2018; Taneja & Yaeger, 2019). By conceptualizing these decisions as one part of a probabilistic system with many possible outcomes, we may be able to more fully grasp the influences at play (Ladyman & Wiesner, 2020). Second, it gives us the flexibility to incorporate disparate actors under a unified framework, and to do so dynamically. It is this flexibility that allows Kaiser et al. (2019) to demonstrate the dynamics of the far-right media ecosystem over time as it developed around the 2016 presidential election. Finally, complexity gives us the ability to examine relationships among actors within the system as *generative*. This is a key characteristic of emergence—parts of a system are not

just connected via interaction, they are connected in a such a way that they produce novel higher-order dynamics (Castellano et al., 2009). Rather than just examining the effect of x on y, we can characterize the relationship between the two, model its characteristics and outcomes, and measure how those outcomes affect z. This allows us to avoid placing disproportionate weight on the most visible aspects of news media (e.g., attributing news coverage decisions to journalists).

## 1.3 Defining a news system

In order to apply a system-level lens to the study of news, we must first define the system of interest. In this work, we are broadly interested in understanding the *process* of digital news: how news coverage gets produced, where it gets distributed on the internet and through what channels, and how and why audiences decide to consume it. This overarching abstraction covers a wide range of phenomena, which involve a wide range of actors. To bridge the gap between the specific empirical contexts of news phenomena and our abstract formulation, we can first identify at a high level what *kinds* of actors are generally most prominent.

This work focuses solely on the actors most directly involved in that digital news process. Journalists and news organizations generate coverage. Platforms play a substantial role in distributing it around the web, especially via algorithmically-curated channels. And audiences both consume coverage and provide feedback—explicit or implicit—on it. Other actors are influential to the digital news system. Regulators play an important role in defining media markets, for example, and a wide array of informal news producers also contribute coverage (C. Anderson, 2010; McChesney, 1996) Further reflection on this broader range of actors appears in Section 6.5. However, in the literature motivating the empirical studies in this work, such actors are not treated as directly consequential to the news process of interest. In the following, we draw from prior work on news production, distribution, and consumption to motivate the actors included in our conceptualization

of digital news media.

### 1.3.1 Journalists

As explored above, journalists are a central focus of newsroom-centric theories of news. They exercise professional judgment in selecting sources, constructing news articles, and framing coverage, making them consequential individual actors in news production (Weaver & Wilhoit, 1986).

Those same characteristics now apply to an expanded set of individuals. Citizen journalists perform primary source news gathering, outside of the bounds of a traditional newsroom (Wall, 2015). And even within the profession of journalism, the traditional model of a full-time news gatherer situated in a newsroom has started giving way to alternative modes of working. Freelance and contract arrangements are increasingly common, and they often displace full-time roles (Marín-Sanchiz et al., 2021).

With that expanded definition comes an expansion of the magnitude and range of news decision making individual journalists can exercise. Outside of the newsroom process, in which organizational priorities and professional editors at least partially determine a journalist's focus, the individual journalist has much more editorial responsibility (Rosenkranz, 2016). They take on the role of assignment editor, deciding on their own stories and seeing through their execution (Rosenkranz, 2018; Storey et al., 2005). They also make the range of decisions described above about how those stories are constructed without external input. In some cases, this makes their published work the sole reflection of their personal editorial judgment, rather than an institutional one (Rosenkranz, 2016).

Similarly, these types of individual journalists have greater editorial latitude. They can carry out a range of journalistic activities beyond the traditional news gathering process, including offering commentary and exploring alternative modes of presenting reporting (Davidson & Meyers, 2016).

Some are, in other words, outside the norm, offering a novel view of how digital journalism should look.

Because individual journalists play such a pivotal role in shaping and presenting news coverage, it is worthwhile to consider them as actors independent of the news organizations that might employ them. This also allows us to consider how various kinds of news workers interact with their employers, and how that relationship impacts the coverage they produce.

### 1.3.2 News organizations

Even as individual news producers take on a larger role, news organizations still occupy a key position as sources of credible journalism. From an audience-facing perspective, publishers remain the primary source of news: According to Pew data, 53% of U.S. adults report getting news directly from a news website or application at least sometimes, a higher share than search or social media (Forman-Katz & Matsa, 2022). Even if they are no longer unilaterally in control of the news agenda, news organizations still play a pivotal role in shaping audiences' awareness and perception of news events (Jungherr et al., 2019; Singer, 1997). At a micro level, news consumers form strong associations with their preferred news sources, ascribing trust to them as institutions.

From the perspective of news production, organizations also wield substantial influence on the process of shaping coverage. News organizations are sites of strategic decision making, wherein priorities get set around the types of news that receive reporting resources (C. Anderson, 2011). That decision making also extends to how news gets presented, in the form of framing, tone, and style (Hagar & Diakopoulos, 2019). Organizations bring abundant institutional resources to these kinds of decisions, allowing them greater reach than individual actors via massive audiences, access to technical resources and newsroom talent, and wide-ranging partnerships with other institutional actors (Hindman, 2018).

News organizations, then, exercise unique influence on both news audiences and other producers. They command a large share of attention to digital news coverage, allowing them to play at least a partial agenda setting role. Simultaneously, they set a strategic agenda for news coverage, dictating much of the day-to-day activity of journalism.

### 1.3.3 Platforms and algorithms

In this work, we consider platforms and algorithms along several dimensions by which they intersect with news.

First, platforms act as sociopolitical entities (Gillespie, 2010), with their own priorities regarding news. For news organizations, they must be reckoned with as intermediaries for large audiences (R. Nielsen & Ganter, 2018). For audiences, they make opaque judgements about the extent to which news is important to show to users, both in terms of discriminating high-value from low-value news consumers (Thorson, 2020) and in terms of the overall perceived value of news to the platform (Thorson et al., 2020). It is this power to determine value, to weigh news against the myriad other types of content on social media, and to convey that value judgement in the form of design decisions that makes platforms consequential actors in the spread of news.

In those design decisions, we see the mechanics of platforms influencing how and where news appears. The UI of a platform works to determine how news content looks on a platform—the extent to which credible reporting is distinguished from other content, for example, or the prominence a story's publisher is given. Features of the user experience, such as collective ranking (e.g., Reddit voting) or interactions with social networks, play an important role in governing news exposure (Lee & Kim, 2017). Similarly, algorithmic recommendations can surface news as a way to ensure users get exposure to important current events, or they can drastically reduce news consumption in favor of recommending other kinds of content (Damstra et al., 2023; Fletcher & Nielsen, 2018).

Via these mechanisms, platforms touch every aspect of the news process outlined above. They can help shape news consumption habits, by modulating exposure. By building large audiences of potential news consumers, they make themselves vital distribution channels for many producers. And in doing so, they gain the ability to shape producers' output according to their organizational priorities (Caplan & boyd, 2018).

This influence stems from a unique interplay between the sociopolitical aspects of platforms and the mechanisms by which they operate. Algorithmic recommenders, for example, are somewhat autonomous actors, in the sense that they can rank-order content for users without human intervention. However, the priorities they embody in those decisions also stem from the organization's strategic goals, and from their codified implementation (Gillespie, 2014). In the case of large digital platforms, then, platforms and their algorithms are distinct but inseparable entities. In considering the organizational stance of platforms, we can view them as akin to news organizations: They cultivate relationships with audiences, and they exercise strategic decision making that impacts news coverage. In examining the design of those platforms, and the various algorithms that interact with digital news, a different type of agency takes form. These are technical systems that work directly as news actors.

As with the full set of actors, the specific platforms chosen for examination in this dissertation—Reddit and TikTok—do not represent the full range of platform news actors. While specific motivations for selecting these platforms are provided in subsequent chapters, they broadly represents contexts in which system design, collective behavior, and algorithmic decision making interact; and in which large amounts of news consumption happen. As such, we view these platforms as productive sites for analyzing behaviors and interactions that may in future work apply to other, similar platforms.

Outside large platforms, some algorithmic systems directly interact with news production to

shape coverage. Many newsrooms leverage optimization systems, which attempt to tweak the form or substance of news coverage to boost, e.g., web traffic (Petre, 2015). Such systems are an important, distinct source of algorithmic influence, because they tend to carry weight inside newsrooms (Hagar & Diakopoulos, 2019). The technical systems provide seemingly objective guidance to editors by operating on internal newsroom data and presenting reader behavior in a semi-experimental context (Hagar & Diakopoulos, 2019). But for many of these systems, the same intermingling of organizational priority and technical mechanism persists. Newsrooms' optimization systems are often licensed from third party organizations, which bring their own considerations to bear in their products. Digital metrics company Chartbeat, for example, organizes traffic data on its dashboard according to a strategic desire to earn journalists' trust (Petre, 2015).

In the cases examined here, algorithms therefore operate via a fundamental entanglement—with the organizations that develop and deploy them, with the platforms through which they encounter users, and with the interfaces that communicate their decisions. To represent that entanglement, we emphasize the particularly connected nature of these actors in Fig. 1.1.

### 1.3.4 News content

Up to this point, we have considered actors who shape news content in some way. But it is also possible for news content itself to exercise influence on these actors. While unlike the technical systems explored above, news coverage lacks a decision making apparatus, its form and contents still shape how other actors interact with it. In the case of news producers, the relative coverage that stories receive at a given point in time often inform subsequent decisions about how to deploy reporting resources (Vasterman, 2005; Waldherr, 2014). In the case of platforms, news coverage can shape product prioritization and design, to the extent that major technology firms (e.g., Google and Apple) have designed products around optimally curating and presenting news

(Bandy & Diakopoulos, 2020). And in the case of the audiences who consume it, news coverage has the potential to shape reading habits—audiences might be enticed to read more on a subject by an article's presentation, or they might be fatigued from the news' negative tone and disengage altogether (Vermeer et al., 2020; Villi et al., 2022). In the background, these interactions might ultimately stem from the entities producing news coverage. However, the immediate, micro-level ways in which interactions occur with particular instances of news content are valuable to consider on their own.

### 1.3.5 Audiences

As explored in an array of news theories, news audiences are active. They make choices about which media to consume, choices that hinge on a complex blend of personal preferences, the relevance of news coverage at a particular moment in time, and deeply ingrained habits (Boczkowski et al., 2018; Webster, 2016). In doing so, they exercise agency over news media by determining the relative popularity of sources and stories.

Beyond those basic choices, audiences also indirectly interact with news producers. Through processes of datafication, their actions are codified, standardized, and aggregated for news publishers (Thorson, 2020). In this aggregate form, the choices audiences make about what to consume shape news coverage, positioning news consumers as consequential participants in the editorial process (C. Anderson, 2011).

Similarly, audiences occupy overlapping consumption and distribution roles in their interactions with platforms. As news consumers, they might encounter news coverage via the platform mechanisms described above. Their decisions about whether or not to engage with that news become similarly datafied, providing feedback to both the platform and the news producer. As distributors, though, news audiences also gain the ability to drastically amplify the reach of news

coverage (Bright, 2016). By sharing news to their social networks, news consumers simultaneously play a role in furthering its exposure and provide an additional signal for which producers can optimize their output (Bright, 2016; Karnowski et al., 2020).

There is, in short, a more explicit duality to the role of audiences than many other news actors. At times, audiences receive influence, as editors and algorithms attempt to shape their consumption habits. In the same process, though, audiences return this influence, implicitly leveraging their behavior to shape future efforts by those actors. This ongoing process of feedback is at the heart of many news phenomena, as producers continually attempt to reach engaged consumers.

### 1.3.6 Building connections

Together, these actors form the basis of a system of digital news production, distribution, and consumption, one within which a range of constant interactions shapes key processes. In Figure 1.1, we lay out the actors described here, as well as the relationships among them that this work addresses.[1] The studies in this dissertation represent further empirical efforts toward a fully-formed complex systems approach for digital journalism research. In each case, they go through a process of identifying relationships among media actors, positing a consequence of those relationships, then analyzing the extent to which that consequence arises. Narrowly, the first three studies encompass familiar delineations within news—production, distribution, and consumption—and incorporate established theory in those realms alongside a complexity perspective. More broadly, all four studies work in conjunction to map out a larger system of media actors, uncovering a unified conceptual structure.

Study 1 was conducted with Johannes Wachs and Emőke-Ágnes Horvát, and was published

[1]Of course, the relationships represented here are only a subset of all possible connections among these actors, raising an important question around how to conceptualize these unconsidered connections. This question is explored in detail in Section 6.5

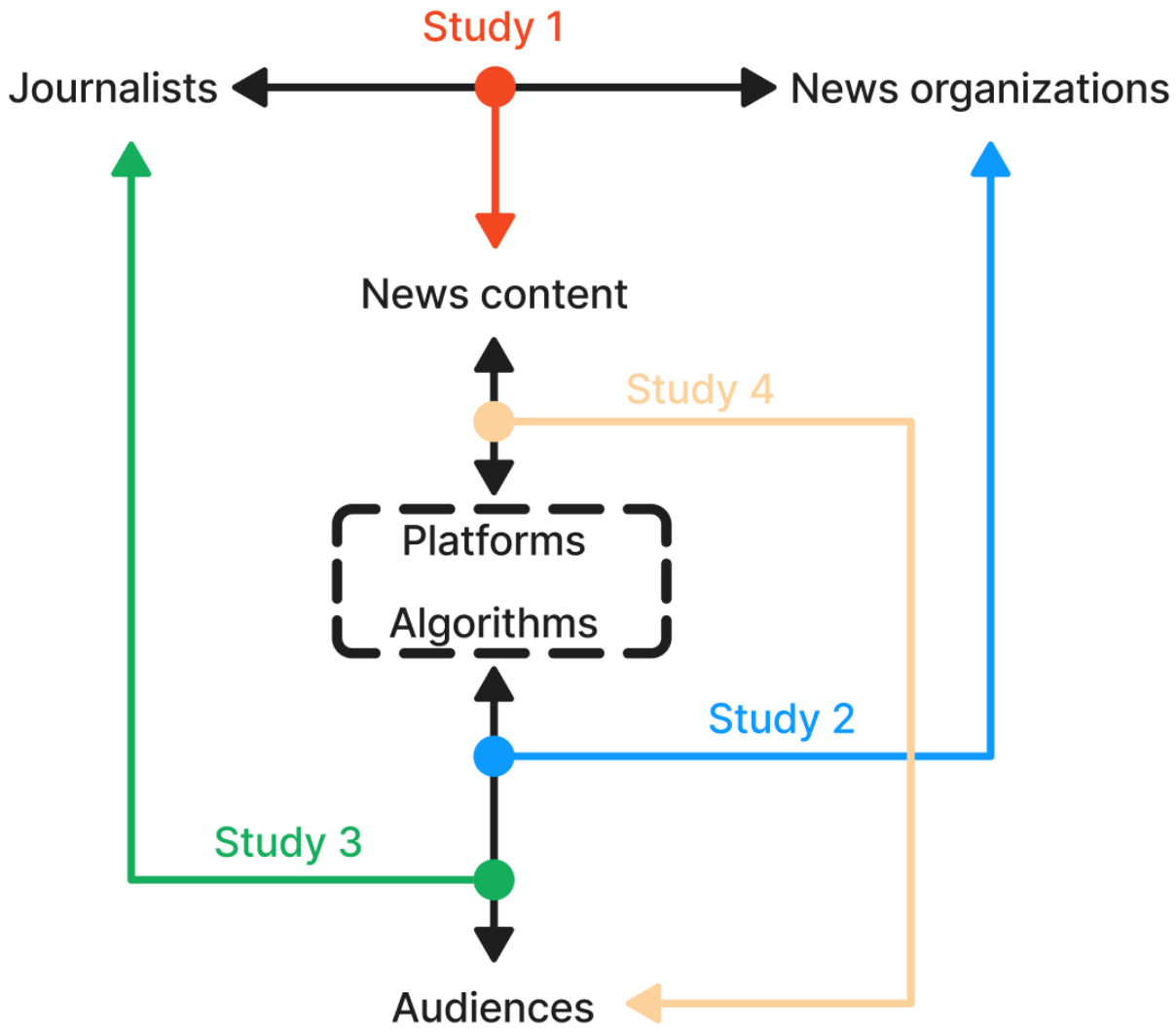



Figure 1.1: A conceptual framework of the digital news ecosystem, and the relationships among its components. From this system-level view, the connections among agents are generative, creating influence beyond that of individual components.

in *New Media and Society* (Hagar, Wachs, et al., 2021). In this work, we scrutinize the relationship between news organizations and the freelancers they commission, in order to highlight the consequences news production decision making has on audiences. We demonstrate the structure underpinning news freelancers' careers, one implicitly formed by the choices of writers, editors, and organizations. We find a network of writers divided along political lines, a division reflected in the topics writers cover.

Turning to news distribution, Study 2—conducted with Aaron Shaw and published in the *Journal of Communication* (Hagar & Shaw, 2022)—demonstrates a case in which structure is messier. In this study, we look at how individual preferences and social media platform design interact to influence the popularity of news outlets. We use data from Reddit's largest news communities to demonstrate that, contrary to prior work on news web traffic, outlet popularity is highly unstable over time. At the same time, we find that the high-level *dynamics* of attention allocation in this context do remain stable. We demonstrate modest predictive power when attempting to model which outlet will be the most popular on a given day, but we can predict how *relatively* popular said outlet will be. This finding, we argue, uncovers a key distinction between the *state* of a news distribution system at a given point in time and the underlying *process* that governs it.

Study 3 was conducted with Nicholas Diakopoulos and Burton DeWilde, and published in *Digital Journalism* (Hagar, Diakopoulos, & DeWilde, 2021). This work synthesizes Study 1's focus on the producer/audience connection with Study 2's reliance on the datafication of audience behavior through intermediaries. However, this study takes an approach much narrower in scope. We examine the data generated by interactions between news audiences and individual headlines, as captured by A/B testing software. We consider how these data might influence journalists' writing decisions, in the form of consistent datafied feedback. In doing so, we again find evidence of a turbulent system. In attempting to model the outcomes of headline tests, we find limited

predictive power from the features of headlines themselves, emphasizing the outsized role that unmeasured factors play in shaping decisions to consume news at a micro scale.

Finally, Study 4, conducted with Nicholas Diakopoulos, pulls together production, distribution, and consumption to examine how TikTok surfaces news. This study incorporates the full range of news actors examined in the work up to this point. We are concerned with the algorithm's response to individual behavior in the case of new users, the news audience's collective decision making in determining the popularity of news topics, and the ways in which these data might feed into news producers' behavior. We find a severe lack of news content on all fronts, which we characterize as a malfunctioning feedback loop among consumer, platform, and producer. In this environment, exposure to news appears to fall by the wayside.

In the following four chapters, we describe the details of each study in full. The final chapter synthesizes key theoretical themes from this work and presents a potential agenda for future research.

# CHAPTER 2
# WRITER MOVEMENTS BETWEEN NEWS OUTLETS REFLECT POLITICAL POLARIZATION IN MEDIA

## 2.1 Introduction

In the context of newsrooms, there is a well-defined relationship between news writers and organizations (Cohen, 2019). However, employment as a full-time journalist is on the wane. In its place, freelance labor makes up an increasing share of professional news production (Antunovic et al., 2019). This invites a more dynamic mode of news labor, one in which writers are transient, working arrangements are temporary, and labor is heavily mediated by personal contact between writers and editors (Gollmitzer, 2014; Hayes & Silke, 2018; Rosenkranz, 2018). This reconfiguration creates a new structure of news production, driven by novel tensions and motivations. In turn, it has the potential to impact key issues in news media in novel ways.

One such issue is polarization. Political polarization in digital news consumption is well established. While the existence of self-reinforcing isolation that is strong enough to *override* individual agency—so called "filter bubbles"—is hotly contested, there is substantial evidence of patterns of media consumption aligned with political identity in the U.S. (Bakshy et al., 2015; Macy et al., 2019; Shi et al., 2017). Whether by an explicit recognition of political identity or the implicit association of certain behaviors and traits with a politically-aligned community, in-group identity drives members of opposing political groups into silos (Coleman, 1988; Dvir-Gvirsman, 2017).

In news production, research often focuses on the individual priorities and professional norms countervailing to polarization. In particular, researchers pay close attention to the professional

norm of objectivity within journalism and how it gets enacted to minimize partisan influence (McChesney, 2003; Schudson & Anderson, 2009). This perspective focuses on the individual journalist as the counterpoint to polarization in news media, as they carefully construct coverage with an eye toward fairness, balance, and accuracy (Ryan, 2001). And while this view acknowledges the role of news producers in polarized environments, it also fails to address the structural forces shaping that production. At an individual level, writers outside of traditional newsroom positions must balance professional norms with strategies that ensure access to ongoing employment opportunities (Edstrom & Ladendorf, 2012). Freelancers must adapt to the demands of clients, and they must shape their work experience in such a way that it appears attractive to future employers (Leung, 2014). The mechanics of freelancing therefore complicate journalists' role as enforcers of objectivity, as they become more dependent on news organizations to set the parameters of journalistic work.

More broadly, research that does take a structural approach finds ample evidence of polarization in the distribution of digital news. Researchers have often sought to test the efficacy of cross-cutting content and conversation in reducing polarization, which exposes members of opposing groups to challenging viewpoints (Heatherly et al., 2017). But in a structure where exposure is only one piece of the puzzle, these efforts may fail to address underlying factors driving polarization. The earliest political blogs demonstrated an overwhelming preference for linking to and engaging in dialogue with like-minded outlets (Adamic & Glance, 2005). Similarly, work on the dissemination of digital election coverage found an isolated, self-reinforcing group of right-leaning outlets, one that largely disengaged itself from the rest of the news ecosystem (Benkler et al., 2018). Third-party social media platforms used for news distribution seem to largely amplify these producer preferences (Bakshy et al., 2015; Benkler et al., 2018; Wihbey et al., 2019). Therefore, even in a world where media consumers have the agency to cross political boundaries and discover opposing

views, the distribution of that media operates in such a way that discourages cross-cutting exposure. Polarization is an integrated problem.

This study attempts to extend that structural view into the production of digital news itself, by examining the movement patterns of digital news contributors from outlet to outlet. We treat these movements as a network, in which edges form between outlets who share contributors. Using thousands of online news stories from 13 digital outlets—with manually-validated information about the journalists, freelance writers, and political actors who wrote them—we demonstrate clear patterns in the trajectories contributors take when they write for multiple outlets. We then show that contributor trajectories across outlets align with the underlying political leanings of the outlets in our sample, more than any other characteristic of the outlets or their audiences. Furthermore, we show that those politically-aligned clusters differ in internal structure, with a highly interconnected cluster of right-leaning outlets. Finally, we link the political clustering present in this network to the content contributors produce. We find that, not only do both the topics writers cover in each cluster and the ways in which they present this content differ, but also that contributors that move between clusters tend to be less explicitly political. Taken together, these findings challenge the value of bridging content in reducing the polarization of digital media consumers. They also demonstrate the need for a more nuanced examination of media *producers* and their underlying motivations.

## 2.2 Background

In news' shift to digital production, newsrooms increasingly rely on freelancers and other outside contributors. This reliance on freelance work creates an open and competitive market for news production. Individual news contributors in this market must develop strategies for successfully navigating news outlets, especially in the midst of extreme political polarization. Those strategies can appear in the movement patterns of contributors from outlet to outlet, as well as the topics

they cover. We posit that such contributors who successfully cross political lines offer potential strategies for reducing polarization.

### 2.2.1 Digital Journalism's Shift to Outside Contributors

For digital news outlets, the shift to outside contributors is largely the result of resource constraints—layoffs push news producers into less secure work positions, such as freelancing, as news revenue continues to decline (Cohen et al., 2019). For professional journalists, freelancing has become the major form of news production work in some places (Hayes & Silke, 2018). Because freelancing relies heavily on individuals pitching news pieces to editors, this shift also represents an increased responsibility for contributors to decide what topics to cover, and for which news outlets (Rosenkranz, 2018).

At the same time, the type and focus of digital news coverage has moved away from traditional event-centered reporting. Rather than present an isolated account of a newsworthy occurrence, reporting often emphasizes analysis and interpretation of events in a broader context (Barnhurst & Mutz, 1997). That change has only been accelerated by blogging and news production on social media, opening up news analysis to a wide array of contributors outside the traditional journalism field (Lowrey, 2006; Singer, 2005). These concurrent shifts create a less formalized sphere of news production, one that extends beyond the traditional professional class of journalists. In turn, news contributors expand to include not only journalists, but also writers, academics, politicians, and other individuals who generate news analysis and interpretation for online audiences.

Such an open, digital environment for news production has two key implications for contributors. First, it creates rapid and constant competition, in which all news producers have equal access to technology and the ability to produce content (Munger, 2020). Second, external modes of labor like freelancing are most often temporary, meaning workers must constantly move from organi-

zation to organization (Berton et al., 2011). In these conditions, news producers must develop and carry out successful strategies for shifting from one position to another, in order to remain competitive.

For full-time news contributors, freelance work is often synonymous with precarious work. Income becomes far less stable, and freelancers must bear additional costs associated with reporting (Gollmitzer, 2014; Salamon, 2020). These challenges add additional pressure to finding successful strategies, since freelancers' livelihoods depend on a constant flow of new work. In particular, success in freelance work depends on the *patterns* of that work—what opportunities freelancers take, and which others arise in conjunction (Leung, 2014). In news production, those patterns appear most prominently via the outlets in which contributors publish articles. Given that emphasis, we ask:

**RQ1:** What patterns do news contributors, particularly external contributors, follow when moving among outlets?

### 2.2.2 Polarized Coverage for Fragmented Audiences

Digital news exists in a fragmented ecosystem. In contrast to mass media like broadcast or print, news consumers split their attention among a wide variety of sources (Webster, 2016). Audiences splinter into small groups as low-barrier, low-cost news sources tailor to their specific interests and identities (Taneja & Wu, 2018).

One prominent dimension of this fragmentation is political polarization. Past work has argued for the existence of "filter bubbles" or "echo chambers", completely isolated spheres of media consumption. The evidence for these kinds of all-encompassing divisions in news consumption is limited (Bruns, 2019). However, political beliefs and preferences are a clear *factor* for consumers in news source and story selection. People generally prefer to read news that aligns with their

political positions (Flaxman et al., 2016). Selective exposure along partisan lines also shapes readers' perceptions of news source credibility, thereby impacting what news they decide to engage with (Tsfati et al., 2014).

In turn, audience preferences feed into increasingly polarized news outlets. Digital news media are heterogenous, with varying agendas, topical focuses, and speeds of reporting (Harder et al., 2017). Those differences serve to distinguish outlets from each other, drawing in different audiences depending on their preferences. As a consequence, news outlets can become more polarized in response to competition, as a strategy for capturing market share (Mitchelstein & Boczkowski, 2009; Mullainathan & Shleifer, 2005). Even in markets with news consumers who do not explicitly seek out polarized viewpoints, models show that newspapers are served well by taking clear political stances (S. P. Anderson & McLaren, 2012).

At the same time, many news organizations and individual journalists go to great lengths to avoid the appearance of any partisan position or influence. As journalists developed a cohesive professional identity throughout the 20th century, industry-wide norms developed that shifted news' presentation (Schudson, 2001). In particular, journalists developed a strong link between professionalism and objectivity (Schudson & Anderson, 2009). This objectivity seeks to balance the perspectives and positions represented in news coverage, rather than advancing any particular narrative (Ryan, 2001). It is important to note that objectivity is often an ethical ideal, not a value reflected in published news. Even so, objectivity (or at least its appearance) is a valuable strategy for news organizations, as it enhances the credibility with readers that they rely on for sustainable growth (Singer, 1997).

Polarization and objectivity set up conflicting accounts of how news organizations position themselves in competitive markets, and they have distinct implications for the strategies outside contributors might pursue. Coverage polarization incentivizes politically-aligned niches of topics

and perspectives, while objectivity encourages broadly salient, more neutral information. These contrasting potential strategies lead to our second research question:

**RQ2:** Do the topics addressed by contributors vary, depending on an outlet's political leaning?

### 2.2.3 Structural Polarization and Crossing Political Boundaries

Structural factors are a major driver of polarization in the production, distribution, and consumption of digital news. As mentioned above, market competition pushes news outlets gradually further away from the center of the political spectrum, in an effort to better capture certain subsets of news readers (S. P. Anderson & McLaren, 2012; Munger, 2020). Algorithmic intermediaries also reinforce polarization via distribution, in that they reward more extreme offerings via recommendations (Blex & Yasseri, 2022; Ribeiro et al., 2020).

Even beyond these mechanisms, researchers have identified *network-based* characteristics of the digital news ecosystem that reinforce polarization. In particular, Benkler et al. (2018) examine the pathways along which news stories travel among outlets. They find that, across several modes of distribution, right-leaning outlets tend to amplify each others' stories most often. They are also isolated from the rest of the media ecosystem, lacking ties, in terms of hyperlinks and social media sharing, to center- or left-leaning outlets. Outlets also diverge in the topics they cover, depending on their political leaning. In this way, Benkler et al. (2018) paint a picture of a media ecosystem that is polarized, not because of any individual agent or behavior, but because of a broad structure of distribution. This account fundamentally addresses the *portability* of news stories and perspectives among audiences.

Outside news contributors add an important piece to this structural account, offering potential novel approaches for combating polarization. There has been much focus on "bursting" news readers' filter bubbles, by giving them indications of the partisan slant of their consumption habits

or intentionally exposing them to opposing viewpoints (Flaxman et al., 2016; Resnick et al., 2013). This approach attempts to tackle partisan isolation in the digital news ecosystem by focusing on cross-cutting exposure in *consumption*. News *producers* can provide a productive complement to that approach in a couple ways. First, from a networks perspective, movement across groups is key for valuable information flow (Granovetter, 1973). Cross-pollination increases exposure to novel ideas and productivity (Granovetter, 1973; Montgomery & Nyhan, 2017; Vedres & Stark, 2010). From an individual's perspective, bridging the gap among isolated groups presents an opportunity for potentially lucrative professional connections (Burt, 1992). By crossing political lines, news contributors may be rewarded with new connections, perspectives, and editorial approaches. They therefore have incentives to cross political lines that consumers do not.

Second, cross-cutting perspectives sometimes create unintended consequences, causing the recipient to become less receptive to opposing views (Bail et al., 2018). To some extent, that negative outcome may stem from the producer's choice of topic or perspective. In environmental news, some frames (e.g., focusing on public health) are more effective than others at engaging climate change skeptics (Bolsen & Shapiro, 2018). More broadly, news consumers' choice of topics often stems from social divisions like political affiliation (Knobloch-Westerwick & Meng, 2009; Tewksbury & Riles, 2015). These findings suggest that cross-cutting news must be tailored to its recipients, such as by more strategically selecting topics. By examining contributors who have successfully produced news across polarized audiences, we may get a better sense of which news topics appeal across political lines. To examine this possibility, we ask:

**RQ3:** Do contributors who cross political boundaries cover different topics from those who do not?

## 2.3 Data

To identify news contributors who moved from outlet to outlet over time, we relied on a manual review of cleaned byline data from a broad sample of news articles. We started with a publicly-available data set of news articles scraped from the homepages of major publishers (Thompson, 2017). The full data set contains 131,860 articles, published between June 2014 and July 2017, across 14 outlets: the New York Times, Breitbart, CNN, Business Insider, the Atlantic, Fox News, Talking Points Memo, Buzzfeed News, National Review, New York Post, the Guardian, NPR, Vox, and the Washington Post.[1] These outlets provide wide variation along a number of key characteristics: not just political valence, but also audience size, institutional prominence, geography, and the non-digital publishing channels they utilize. While these outlets represent only a subset of all digital news media, relying on a pre-made set of news outlets precludes any bias that might arise from our own selection criteria. The results we show below are also not dependent on this particular set of outlets, as removing outlets at random does not change the outcome of our analysis (see Appendix A for details of this robustness check).

We filtered this full data set in three ways. First, we removed any records with author fields that do not correspond to a person's name. To accomplish this, we removed any stories with bylines that matched generic phrases (e.g., "the editors", "anonymous"), outlets (e.g., "The Associated Press", "NPR staff"), or that contained bylines with multiple names. Second, because we are interested in movement between outlets, we limited our sample to articles by contributors who appeared in at least two outlets in the data set. Because Business Insider does not share any contributors with other outlets, we removed it from our analysis. Finally, we manually verified the remaining bylines, disambiguating cases of multiple people sharing the same name. This process identifies

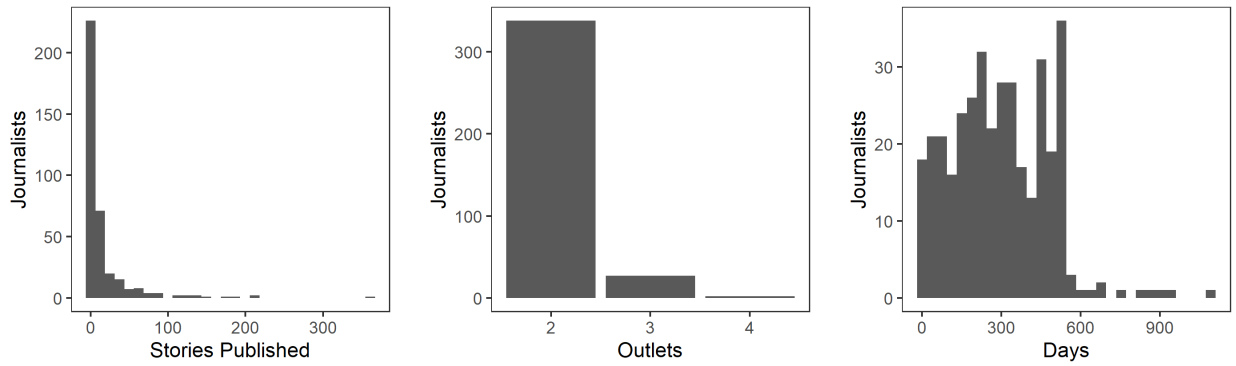
[1] Because Reuters is a wire service from which stories often appear across many outlets, we could not distinguish true contributor movement from syndication. For this reason, we excluded Reuters from our sample.

368 contributors, who wrote 6,032 articles and are the focus of our analyses. While these articles represent only 4.6% of our initial set, they are consequential in that they represent all inter-outlet movement within our sample. We are able to capture thousands of cases in which contributors move between outlets, laying the groundwork for a robust network model.

We next manually coded each contributor depending on their professional background. First, there are three types of professional journalists in our sample: Freelancers who write for multiple news outlets (97 names), journalists who are full-time employees at an outlet but write for at least one other (67 names), and journalists who move from one full-time job to another over the course of our sample (43 names). Overall, then, 207 of the contributors in our sample are professional journalists of some kind. We also find a couple of groups who write pieces in support of particular issues or causes. This includes activists and think tank members (62 names) as well as political commentators and politicians (27 names). Finally, there are smaller groups of academics (33 names) and authors (33 names) who write pieces primarily to promote their work. An additional 6 contributors do not fit into these categories.

Within our sample of 368 identified contributors, we analyzed publishing patterns for each individual (Figure 2.2a). Figure 2.1 presents the volume of stories, number of outlets published in, and number of days from first to last story for each contributor. We find that many contributors write across a couple outlets, and the majority publish fewer than 100 stories. However, there is widespread variation in how long contributors are actively publishing stories throughout our sample, indicating differences in their tempo of activity.

At the outlet level, we use a variety of characteristics to evaluate our eventual network's structure (Table 2.1). These characteristics fall under three broad types of measures, and they broadly capture information about the outlet and the size and composition of its audience. First, we used two measures of outlet size—total unique users and total time users spent reading each outlet—as




(a) (b) (c)

Figure 2.1: Distributions of a) story counts per contributor, b) unique news outlets that published each contributor, and c) the number of days from each contributor's first to last story in our data.

these are often utilized in studies attempting to classify news outlets (e.g., Hindman, 2018; McCombs & Winter, 1981). Second, we used a variety of audience characteristics—household size, whether or not a household has children, the race and gender composition of the audience, where the audience is located, and the median reader age and income. All measures except income report audience size, in terms of unique users, for certain groups. For example, household size reports total users for households with 1-2 and 3-5 members. All measures except region, age, and income report two groups, so we calculated the ratio of one group to the other to reduce the number of classifications. For audience region, we retained the largest region by number of users. Audiences often identify with the news they read, particularly in alternative media, which could create observable clusters of outlets within shared communities (Benkler et al., 2018; Chiricos et al., 1997; Couldry & Curran, 2003; Kaiser et al., 2019). Third we utilized a couple outlet characteristics: traditional publishing medium and political leaning. Publishing medium is a common classification approach for comparative work (e.g., Boczkowski & de Santos, 2007; S. Maier, 2010), and while the articles we analyze are all published online, outlets still maintain other dominant distribution media that may inform their overall publishing strategy. We utilized the outlet political leanings

generated by (Bakshy et al., 2015) in examining polarization on Facebook.

ComScore provides all our audience measures. For these, we obtained the mean monthly value of each measure from January 2016 to December 2017, when most articles were collected. We then divided each continuous measure into three groups, each of the size $(max - min)/3$. For political leaning, we use a measure devised by Bakshy et al. (2015). They analyze the self-reported political affiliation of Facebook users, assigning news articles an average alignment score based on who shared each article. They then average those article-level scores at the website level, producing a site-level alignment score ranging from -1 (left-leaning sharers) to 1 (right-leaning sharers). Bakshy et al. (2015) coarsen this measure by quintiles. We follow a similar approach, treating sites that fall outside of their central quintile as left- or right-leaning.

| **Measure** | **Source** |
|---|---|
| Total unique users | ComScore |
| Total time read | ComScore |
| Median reader income | ComScore |
| Median reader age | ComScore |
| Audience household size | ComScore |
| Audience children (yes/no ratio) | ComScore |
| Audience race | ComScore |
| Audience gender (M/F ratio) | ComScore |
| Audience region | ComScore |
| Outlet medium | Manual coding |
| Political leanings | See: (Bakshy et al., 2015) |

Table 2.1: Measures used for classifying outlets, along with their sources.

## 2.4 Methods

Here we describe the analytic approach we used to examine the network between outlets that maps news contributor movement, as well as to investigate the topic and tone of stories published.

### 2.4.1 Contributor Network Structure

To evaluate the network structure of contributor movement, we constructed a bipartite network of writers and outlets, utilizing only our filtered set of writers. A writer and a outlet shared an edge if at least one of the writer's stories appeared in the outlet within our sample. From this network, we constructed a one-mode projection for outlets, in which outlets share an edge if at least one writer published an article in both during the considered time frame (Figure 2.2b). Edges were weighted according to the number of writers outlets share. The number of articles per contributor across outlets also varies; however, the article count is much less relevant to our research question than the number of contributors shared. 31 pairs of outlets did not share any contributors.

We then compared this weighted projection to randomized bipartite networks of news outlets and contributors, to check for significant connections between outlets. We used a standard Markov Chain Monte Carlo (MCMC) algorithm to build random networks from the observed bipartite network (Cobb & Chen, 2003; M. E. J. Newman, 2001; Rao et al., 1996; Tumminello et al., 2011; Zweig & Kaufmann, 2011). Because our network is bipartite and has a skewed degree distribution, the MCMC approach provides the most exact null model comparison (Schlauch et al., 2015). We attempted $m*log_e(m)$ edge switches, where $m$ is the number of edges in the bipartite network (i.e., 784) to obtain a single instance of a random network that has a substantially different structure than the observed network (Gionis et al., 2007). We generated a set of 10,000 such random networks. Table A.1 shows the significance of edges in the projection; we retain edges where $Z > 1.96$. This procedure generates distinct and disconnected clusters of outlets.

### 2.4.2 Classification Evaluation

To evaluate the uncovered network clustering, we compared it to classifications from the mixture of audience and outlet characteristics described in the Data section (Table 2.1). We treated each

of these classifications as though they were partitions within our observed network, then used that partition to calculate the modularity of the classification. Ranging from -1 to 1, modularity describes the extent to which a network is split into distinct clusters (M. E. J. Newman, 2006, 2010). By mapping classifications onto our observed network as though they were clusters, essentially we calculated the extent to which each one aligned with the observed structure.

### 2.4.3 Topic Modeling

To analyze how the network structure of the journalism marketplace relates to content, we employed topic models (Mohr & Bogdanov, 2013). Specifically, we fit a topic model to the articles in our data set with Latent Dirichlet Allocation (LDA) (Blei et al., 2002) using the Mallet program (and its default parameter settings) (McCallum, 2002), accessed via the gensim library (Řehůřek & Sojka, 2010) of the Python programming language. The result assigned each document a vector of topics, allowing that a document consists of a combination of topics and that words appear in multiple topics. The number of topics is a parameter of the LDA algorithm—in the paper we presented results from a model set to find twenty topics. We found similar results, namely significant differences in topic prevalence across the clusters of outlets, when we fit the model to 15, 25, and 30 topics.

Before we fit the topic model, we processed the text of the articles. We extracted each word or token, lower-casing all characters and removing apostrophes. We removed English-language stop words and stemmed the remaining words using the Porter stemmer. We removed words that occurred in more than one-third of articles and those that occurred in less than 10 articles in our corpus. The general goal of these steps was to reduce the noise of the data while keeping its main signals relevant to our research intact (Hopkins & King, 2010). The order of our processing pipeline is in line with recent recommendations of best practices in topic modeling for communi-

cation research (D. Maier et al., 2018).

Our network analysis finds a clear two-cluster structure, one that maps onto outlets' political leanings. We checked whether there are significant differences in content between our derived outlet clusters by comparing the topic vectors of their articles. We identified which topics are over-represented among articles from individual clusters by calculating the average topic vector of articles in the cluster and then comparing this vector to average topic vectors calculated after randomizing the political lean of each article's outlet. This null model, which we generate 1,000 times, presents a random assignment of political label. For each topic, we calculated a Z-score comparing the prevalence of the topic in the observed articles with the average prevalence of the topic after randomization and scale by the standard deviation. We conducted this process separately for articles in each cluster.

### 2.4.4 LIWC Features

Besides the significant differences in content between the derived clusters, we also examine important stylistic differences. This inquiry is driven by the observed effect of news story presentation on audience reception, especially along partisan lines. In particular, past work has demonstrated how stories that are subjective and convey emotion reinforce partisan isolation in news consumption and increase virality on social media (Berger & Milkman, 2012; Flaxman et al., 2016; Xu et al., 2020). Here we investigate how these semantic aspects of news stories vary between our outlet clusters. We apply the Linguistic Inquiry and Word Count (LIWC) dictionary (Tausczik & Pennebaker, 2010), a widely used tool for investigating stylistic properties of text, to each article in our corpus. LIWC assigns text scores in various linguistic (e.g., the use of pronouns, prepositions, punctuation) and psychological (e.g., the use of words with significant positive or negative emotional valence) dimensions curated by human experts. LIWC features have been used to an-

alyze the effectiveness of various kinds of persuasive writing from crowdfunding pitches (Horvát et al., 2018) to public advocacy appeals (Bail et al., 2017). A study of democrats and republicans on Twitter using LIWC found significant differences between the characteristic linguistic style of the two groups (Sylwester & Purver, 2015), for instance that republicans were more likely to use words expressing negative emotion.

Using the LIWC software, we scored each article in our corpus along 83 dimensions of linguistic features. We also considered two additional sentiment-based features from the VADER library (Hutto & Gilbert, 2014). We compared the distribution of each feature between the politically-aligned clusters using a Mann-Whitney U test. We find that 50 out of 85 features have significant differences (Bonferroni corrected $p < .01$) in their distributions between the two groups of articles. The number of significant differences suggests substantial stylistic differences in the writing presented in the two clusters. The full table of features and differences are presented in Appendix A.

## 2.5 Results

To address RQ1, we first investigated the patterns common across contributors' movements. Since movement occurs at the level of the news outlet, we analyzed what contributors' next steps tended to be from any given outlet (e.g., do contributors who write a story for the *New York Times* tend to also write for the *Washington Post*?). Fig. 2.2c demonstrates these trends, by showing our outlet-to-outlet network with significant edges only. This network of statistically significant movements between news outlets reveals striking patterns in contributors' publishing histories. In particular, two clusters of news outlets emerge: One low-density cluster comprised of nine outlets, and a dense cluster of four outlets. In the low-density cluster, only three outlets have connections with more than two outlets, creating a chain-like structure. From a contributor perspective, this indicates that

jumps around the cluster are unlikely, but that travel to each outlet from every other is possible. In the high-density cluster, however, 5 of the 6 possible edges are present, indicating high mobility throughout the cluster. Between clusters, though, movement is highly unlikely. Outlets across clusters do not just lack significant positive connections; they have statistically significant negative edges. See Table A.1 for further details.

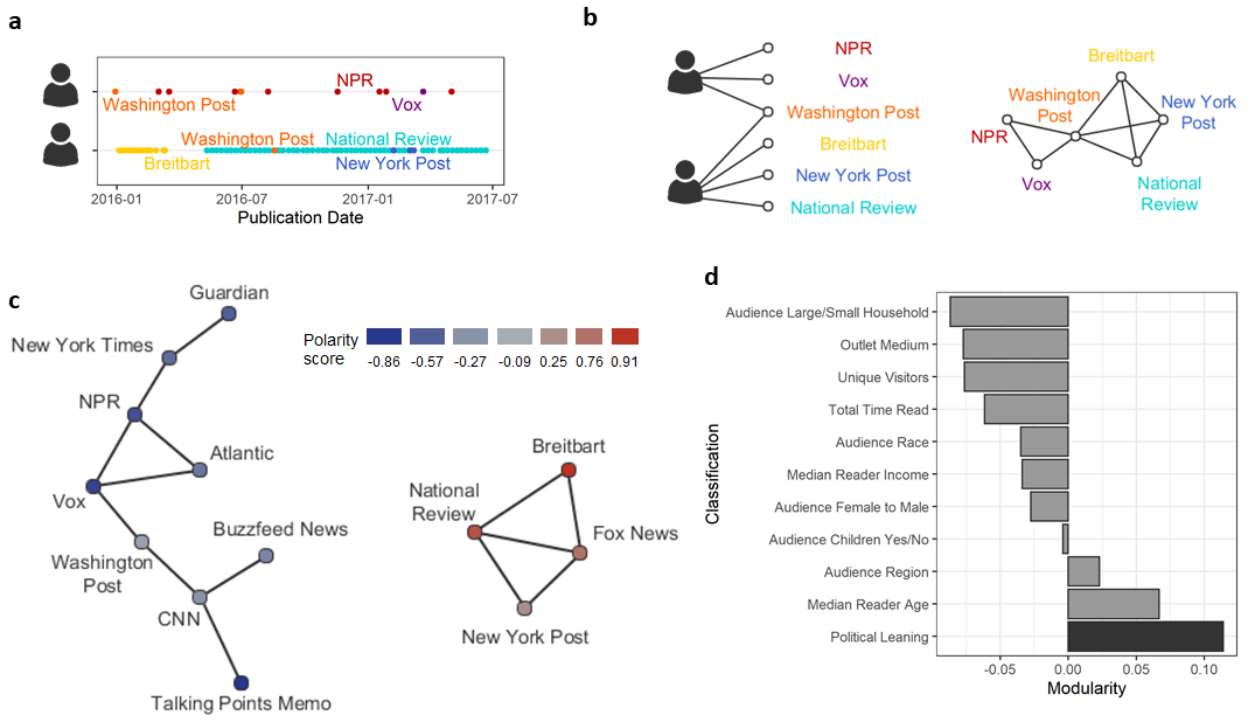



Figure 2.2: Mapping contributors' trajectories between news outlets. a) Example trajectories indicating the outlets for which the selected two contributors wrote articles. b) Illustration of how we build the network of outlets based on which pairs of outlets shared contributors. c) Significant connections between news outlets ($Z > 1.96$), colored according to political leaning scores assigned by Bakshy et al. (2015). d) Modularity for each outlet classification applied to the outlet network.

This clustering structure suggests that there are some constraints or barriers to the movement of contributors between outlets, in particular as contributors seem to lack the flexibility to move between clusters. However, those constraints apply unevenly. Looking across contributor types, the

majority of non-journalist authors, and academics cross between clusters at least once. However, less than a third of freelancers and other types of journalists move between clusters. For them, association with one cluster may represent a choice between two mutually-exclusive publishing paths, which have difficult-to-reverse consequences for future opportunities (Leung, 2014). In contrast, politicians make up 36% of all cross-cluster contributors, but only 5% of single-cluster contributors. Independent of the topics they focus on, then, politicians appear to move more freely between clusters.

These patterns are observable without taking into account any contributor or outlet attributes. Rather, they are structural characteristics that emerge from a combination of the editorial process and labor market. If the market structure of contributor-outlet relationships is built out of implicit patterns in activity, what heuristics—audience size or composition, for example—align with this structure? Are new classification mechanisms needed to understand the outlet landscape?

### 2.5.1 Structure of Contributors' Movements between News Outlets

Traditionally, work that examines multiple outlets has divided them by their publishing medium—comparing print newspapers to news websites, for example, or radio to television (S. Maier, 2010). In this network, though, the outlet's traditional medium does not seem to align with the clusters we observe: Magazine contributors also write pieces for radio and digital-native sites, for example journalists shared by *The Atlantic*, *NPR*, and *Vox*. To validate this observation, we categorized each outlet using a variety of metadata, described in the Data section. Then, we measured how well each of these classifications captures the clustering dynamic in our observed data using modularity (see details in Methods section). Our goal is to determine if traditionally-accepted classification strategies can capture similar information about contributor movement patterns as our network structure, and if not, whether any other characteristics might provide a useful proxy.

In Figure 2.2d, we show the modularity of each classification. Accordingly, outlet political leanings assigned by (Bakshy et al., 2015) correspond best to the observed clustering. If we map the leanings of each outlet to our observed network (Figure 2.2c), we see a dense cluster of right-leaning outlets and a loose grouping of left- and center-leaning ones. Given this dichotomy, we subsequently refer to these clusters as left/center-leaning and right-leaning. This finding suggests that the political leaning of a news outlet is a key structural factor in where contributors will publish, pointing, as in prior work, to a polarized online news ecosystem (Adamic & Glance, 2005; Bakshy et al., 2015; Benkler et al., 2018; Kaiser et al., 2019; Wihbey et al., 2019). While this structure impacts movement, though, does it affect the content of news articles across the network?

### 2.5.2 News Coverage Effects

The Z-scores derived from our topic model comparison across clusters (see Methods) indicate the most relevant words of topics statistically over-represented among left/center- and right-leaning outlets in the first columns of Table 2.2. A Z-score with absolute value greater than 1.96, corresponding to a p-value of 0.05, is considered a statistically significant deviation from the null hypothesis that cluster lean labels are unrelated to the distribution of topics. Articles appearing in center/left-leaning outlets are significantly more likely to be about science and research (Topic 8), the media and press (Topic 3), and healthcare (Topic 16), among others. In right-leaning outlets, articles are more likely to be about conservatism and liberalism (Topic 10), prominent Democratic Party politicians (Topic 13), and Republican politics (Topic 4).

17 out of 20 topics are significantly over-represented among writings appearing in either the right or left-leaning clusters of outlets. In relation to RQ2, this indicates that the partition of outlets according to transitions by contributors aligns with differences in content. Our findings prompt a

question: Do those contributors that do transition between the left and right-leaning clusters adapt their writing topics to the venue? Or do these transitioning contributors fill in particular niches within the partisan spheres they visit?

To address this question, consider three groups of contributors: those who write only for center/left-leaning outlets, those who write only for right-leaning outlets, and those who write for both. Prior literature suggests that these groups will focus on different topics, for two reasons. First, if outlets gravitate toward certain political positions, they will often emphasize pieces of information relevant and favorable to that political group, and omit those that are not (Bernhardt et al., 2008). Therefore, if contributors write stories that are in line with the editorial approach of the outlet as a whole, we would also expect the information they produce to shift depending on their group of focus.

Second, it is challenging for either individual journalists or outlets as institutions to credibly maintain multiple political positions in their work (S. P. Anderson & McLaren, 2012). A contributor who writes a piece about the dangers of climate change, for example, cannot also credibly write a piece denying its existence. It is then in contributors' best interest to establish and develop a consistent perspective in their writing throughout their careers. The contributors who remain in only one cluster will generate that cluster's representative content (e.g., freelancers publishing in left-leaning outlets will focus on the media), while contributors who move between clusters will write about topics that aren't particular to either. In other words, in response to RQ3, we expect that contributors moving between the left and right-leaning outlets tend to write about politically more neutral topics.

| **Topic: Keywords** | **CntrLeft Z** | **Right Z** | **Trans-CntrLeft Z** | **Trans-Right Z** |
|---|---|---|---|---|
| 1: attack, isi*, islam, muslim, war | 0.42 | 0.8 | *-3.69* | *-5.29* |
| 2: vote, voter, percent, parti*, poll | *-2.63* | **4.07** | 0.52 | 0.58 |
| 3: news, media, post, press, twitter | **9.99** | *-9.97* | *-5.17* | -1.16 |
| 4: senat*, cruz, candid*, gop, parti* | *-5.62* | **7.24** | **2.20** | **4.70** |
| 5: realli*,talk, didn*, lot, someth* | **4.75** | -0.86 | *-5.55* | 0.38 |
| 6: game, play, team, season, serv* | **6.31** | *-5.65* | 1.86 | **6.85** |
| 7: women, school, children, famili* | -1.04 | **3.16** | *-4.33* | **5.62** |
| 8: research, studi*, human, found, univers* | **13.07** | *-12.37* | *-7.51* | **2.18** |
| 9: film, music, play, movi*, book | **5.23** | *-3.47* | *-2.66* | **3.37** |
| 10: conserv*, power, left, liber* | *-23.59* | **27.43** | **11.93** | *-4.57* |
| 11: compani*, million, bank, busi* | *-6.64* | *-8.70* | *-3.24* | **13.07** |
| 12: court, law, immigr*, case, rule | -0.36 | -0.76 | **18.86** | -1.25 |
| 13: clinton, obama, hillari*, bill, email | *-15.61* | **14.45** | **2.30** | *-6.63* |
| 14: white, black, christian, commun*, religi* | *-3.30* | **5.38** | -1.29 | *-2.86* |
| 15: tax, percent, job, econom*, rate | *-3.92* | **3.16** | -0.28 | -0.91 |
| 16: health, care, plan, bill, insur* | **7.70** | *-7.14* | *-3.64* | -0.02 |
| 17: china, obama, unit, iran, foreign | **3.11** | *-2.89* | **6.92** | 1.35 |
| 18: offici*, investig*, secur*, inform, depart | **6.20** | *-6.21* | *-2.68* | *-5.58* |
| 19: citi*, build, water, climat*,area | **4.69** | *-5.10* | -1.54 | -0.65 |
| 20: polic*, offic*, crime, gun, case | *-2.06* | **2.03** | -0.50 | *-4.51* |

Table 2.2: Topic keywords and their statistical over-representation (bold) and under-representation (italic) within news outlet clusters. The first two columns of Z-scores present statistical over/under-representation of topics within the center/left- and right-leaning clusters. The last two columns record the statistical over/under-representations of topics in articles by contributors who move between clusters, compared to those by contributors who stay within the center/left- and right-leaning clusters, respectively.

We tested this idea by considering within-cluster differences. For both the left and right-leaning cluster of outlets we compared two kinds of contributors: those who stayed within the cluster, which we call purists, and those who transitioned. For example, among all articles in right-leaning outlets we compare which topics tend to be covered by purists and those which tend to be covered by transitioning contributors. We repeated a similar experiment to the one used to determine topics over-represented among left and right-leaning outlets. Within a cluster we calculated the average topic vector of articles written by purists. We compared the average topic vector to those calculated from 1,000 randomizations realized by shuffling the purist/transitioning contributor labels. This randomization allowed us to test the statistical over-representation of topics by purists or transitioning contributors within the partisan clusters.

The results of this investigation are shown in the last two columns of Table 2.2 for articles in left-leaning outlets and right-leaning outlets, respectively. Here we report the relative prevalence of topics in articles written by transitioning authors within the two clusters, compared to their purist counterparts. In both clusters there are many topics which are significantly over and under-represented in the writings of transitioning authors. This indicates that switching contributors in both partisan clusters occupy specific niches with their writing.

Among the articles in right-leaning outlets, transitioning contributors are more likely to write about finance (Topic 11), sports (Topic 6), and families (Topic 7), among other topics. They are significantly less likely to write about Democratic Party politicians (Topic 13), investigations (Topic 18), and conflict in the Middle East (Topic 1). We note that of the six topics significantly under-represented in the writings of transitioning contributors relative to purist, three topics are globally over-represented in right-leaning outlets (Topics 10, 13, 14). We interpret this as suggesting that transitioning contributors are less likely to write about extremely partisan topics.

The same analysis of left-leaning outlets suggests a similar pattern. Transitioning contributors

writing for left-leaning outlets seem to avoid topics over-represented in those outlets on a global level (Topics 3, 8, 18). Two of the five topics in which switching contributors are over-represented are globally over-represented among right-leaning outlets. Future work should investigate whether this observed relationship is because transitioning contributors avoid partisan topics in order to remain widely employable, or perhaps because the kind of contributor who is able to switch between the partisan clusters tends to specialize in these topics.

### 2.5.3 Stylistic Differences in Language

Our LIWC and sentiment analysis also generate clear differences across outlet clusters. We report the top six most distinguishing features by effect size (measured by the AUC) in Table 2.3; the full table of features can be found in Appendix A. Three of these features are significantly over-represented in articles in center/left-leaning outlets: *hear*, capturing the use of words describing the act of listening; *percept*, capturing the use of observational words relating to perception; and *focuspast*, capturing the use of past-tense verbs. Three are over-represented in right-leaning outlets: *affect*, counting the use of words with significant psychological content; *negemo*, counting the frequency of words with negative emotional connotation; and *certain*, counting the use absolute words such as "always" or "never".

| Feature | M-W U | AUC | CntrLeft Avg. | Right Avg. | Diff. | Bonf. P |
|---|---|---|---|---|---|---|
| **hear** | 3144691.0 | 0.65 | 1.08 | 0.71 | 0.37 | $< 10^{-89}$ |
| *affect* | 3242798.0 | 0.64 | 4.33 | 5.10 | -0.77 | $< 10^{-77}$ |
| **percept** | 3260857.5 | 0.64 | 2.38 | 1.87 | 0.51 | $< 10^{-75}$ |
| *negemo* | 3438952.5 | 0.62 | 1.88 | 2.36 | -0.48 | $< 10^{-55}$ |
| *certain* | 3476109.5 | 0.62 | 1.06 | 1.30 | -0.24 | $< 10^{-51}$ |
| **focuspast** | 3616885.0 | 0.60 | 4.09 | 3.45 | 0.64 | $< 10^{-38}$ |

Table 2.3: The top six LIWC features distinguishing texts from center/left- (bold) and right-leaning (italic) outlets, by their AUC scores.

At a high level, then, we observe widespread differences in the relative usage of certain linguistic and semantic markers across right- and center/left-leaning outlets, giving us an affirmative answer to RQ2. Not only are audiences across these clusters exposed to different *content*; that content is also subject to differing *presentations*. In particular, we see that the usage of affect and emotion varies across these clusters, suggesting that the subjective frames with which writers present news stories depend on the preferences of the outlet and its audience.

## 2.6 Discussion

This work addresses an understudied area of news media polarization: structural production forces driving partisan leanings. By tracing the individual news articles published by writers across a sample of diverse outlets, we trace an emergent structure of freelancer behavior. That structure aligns with a clear partisan divide in the digital news ecosystem, and it appears without any explicit consideration of audience behavior or preferences. Nor is it driven by any top-down organization on the part of newsrooms.

Our results demonstrate the ways in which structural factors can work against individual ones, such as a professional commitment to objectivity. Some contributors in our sample are not journalists, and some are explicitly political figures. However, those that do fall under the umbrella of professional journalism often stay within partisan bounds. Zooming in, that dynamic emerges from many individual interactions between freelancers and the editors who hire them. Freelancers generate story ideas, shape them for specific outlets, and pitch editors at those outlets for assignments (Rosenkranz, 2018). Editors evaluate pitches in the context of the overall coverage mix they want to produce, and from the perspective of freelancers' reputation (Christin & Petre, 2020). From that simple set of motivations and interactions, a higher-order emergent structure takes shape. The driving mechanisms in specific cases may vary, and those possible mechanisms warrant further

study. But at a high level, somewhere within the editorial process of pitching, selecting contributors, assigning stories, and producing news coverage, a dynamic arises that structurally prefers contributors whose publishing histories ideologically align with a publication's own. Given the ways in which objectivity is so forcefully conveyed as a key standard at many outlets, it is also worth examining where outlet interactions with structural forces cause this norm to break down.

Because of the integrated nature of polarization within this system of production, assigning it any one cause belies the complex interactions at play. Moving beyond the interpersonal interactions, polarization also shows up in the topic and style of news coverage that writers produce. Distinct areas of interest and modes of presentation arise for each partisan outlet cluster, affecting news coverage at the level of individual articles. Prior work has demonstrated differences in news coverage approaches across the political spectrum (Schiffer, 2006; Xu et al., 2020), but not in the context of a broader structure of news production. By situating polarization in this way, we invite deeper questions of how it is *motivated*. At the outlet level, selective exposure and profit maximization are promising candidates for mechanisms of polarization (S. P. Anderson & McLaren, 2012; Stroud, 2010). However, honing in on these factors as the *primary* drivers runs the risk of ignoring broader connections, leading to ineffectual pinpoint interventions. Producer-targeted interventions ignore audiences' desires for similar perspectives to their own. Audience-targeted interventions ignore producers' incentives to publish traffic-driving partisan coverage. Neither fully grapples with the feedback mechanism—running through the production, distribution, and consumption of digital news—tied to revenue incentives outlets must confront to remain sustainable (S. P. Anderson & McLaren, 2012; McChesney, 2012). Future work should aim for more holistic interventions that address this integrated polarization.

We also find that topics less related to politics are more common for contributors who move between clusters. Relatively few contributors produce cross-cutting political news, reducing au-

diences' exposure to opposing viewpoints. Future work should examine whether more politically neutral topics are an effective starting point for cross-cutting exposure. In particular, it would be worthwhile to investigate whether the commonality of some topics also holds from an audience perspective–does sports coverage cut across political preferences, or does the partisan affiliation of a particular outlet affect all the coverage it produces? These results further highlight an interesting distinction between cross-cutting *topics* and *contributors*. Political topics are less common for cross-cluster contributors, but politicians make up a large share of those contributors. While politicians are still in the minority of all cross-cluster contributors (36%), their prominence raises questions about differing expectations and perceptions across types of contributors. These distinctions should be explored in future work.

Finally, this work also finds parallels in other structural examinations of polarized news media. Benkler et al. (2018) identify a very similar network structure in news media distribution to what we see here: a loosely connected group of left/center-leaning outlets, a dense core of right-leaning outlets, and very little activity between. Using a totally separate sample, and examining a different aspect of digital news media, we find an ecosystem with identical characteristics. This suggests the potential for a broader-reaching networked model of media polarization, one that might apply across discrete news processes. At the same time, Benkler et al. (2018) also demonstrate how the periphery of right-leaning outlets amplify conspiracy theories and misinformation into the core of mainstream media. In contrast, we see very little movement from the periphery inward. This network structure is potentially good for stemming the flow of misinformation, but potentially bad for amplifying the content-based partisan leanings we see in both clusters. Guess and Coppock (2020) find that readers often update their views in response to information that contradicts their preconceptions. However, this updating process can only happen if readers are exposed to cross-cutting content in the first place, an unlikely prospect in a politically isolated media environment.

### 2.6.1 Limitations

This study has several important limitations. First, we cannot control for all the potential biases that may have arisen in data collection via scraping. We demonstrate the robustness of our results to differences in sample and significance level with additional checks (Appendix A). First, we check the robustness of the network to varying Z-score thresholds. We also examine the impact of removing individual outlets from the network on its structure. However, it is still important to acknowledge potential biases introduced by this dataset.

Similarly, our sample represents a subset of all news outlets. While the outlets used here are prominent sources of digital media from across the political spectrum, they are only one part of the news ecosystem. On a larger scale, other factors may interact with polarization to provide a more complex picture of contributor movement. For example, prior work stresses the importance of considering the organizational context of news outlets when evaluating their coverage (Becker & Vlad, 2009). Factors such as the extent to which an outlet depends on freelancers, its reputation as a publishing venue, or its geographic focus may impact contributors' venue choices, and they may do so unevenly. This in turn might create an *unevenly* polarized network on a larger scale, affecting some outlets far more than others. To better understand how other factors might interact with, or even mitigate, the polarization we observe, future work should attempt to capture a broader subset of the news production ecosystem.

Finally, outlet political leaning is a difficult quality to measure. It originates from the perspective of the audience, rather than any objective measure of the news coverage an outlet produces. We use a rigorous, peer-reviewed method to ascertain our scores (Bakshy et al., 2015), but other approaches may categorize outlets differently.

### 2.6.2 Conclusions

This study demonstrates the polarized structure of digital news contributor movement. It ties that structure to differences in news coverage content and perspective across the political spectrum, and it highlights the topic differential across inter- and intra-cluster movement. Together, these results show the interconnected nature of polarization and its consequences across various aspects of news production. These findings fill an important gap in the literature on news media, highlighting structural factors on the production side which may contribute to polarization.

# CHAPTER 3
# CONCENTRATION WITHOUT CUMULATIVE ADVANTAGE: THE DISTRIBUTION OF NEWS SOURCE ATTENTION IN ONLINE COMMUNITIES

## 3.1 Introduction

Chapter 2 explored a structure of news production implicitly generated by many one-to-one, transactional interactions between journalists and news organizations. These interactions, at scale, produced well-worn pathways among outlets, observable as a network over an extended period of time.

Similarly, this study examines an ongoing collective outcome—relative news outlet popularity—as it results from repeated, individual action. However, this study also presents an alternative mode of building structure. We examine news behavior in the presence of a system intermediary, where product design interacts with user intent to shape attention flows. We do so with an eye toward a formalized underlying process of collective attention to news, one that operates within a probabilistic system of news distribution. To guide this inquiry, we ask: What mechanisms drive the distribution of attention to news sources in the digital age?

The rise of digital media, social networked sites, and online communities seems to exacerbate a tendency toward concentration (Webster, 2016). Web-based patterns of traffic, attention, and participation follow extremely skewed, unequal distributions, often reproducing existing hierarchies of power and influence. A small set of media outlets (themselves increasingly concentrated in terms of ownership), posts, and stories gain massive, viral audiences through large-scale online collectives engaged in the production, aggregation, and dissemination of news.

Often, this concentration is tied to cumulative advantage as a driving mechanism. Findings of concentration have identified cumulative advantage as the mechanism by which "the rich get richer" across a wide variety of areas, including news source attention and web traffic (DiPrete & Eirich, 2006; Hindman & Rogers, 2018). In the case of news attention, popular offerings benefit from broad access to casual consumers, and from a relatively loyal audience compared to more obscure outlets (Ehrenberg et al., 1990; Taneja et al., 2012). These factors set up a compounding mechanism of divergent popularity, the process by which concentration forms.

The driving factors of cumulative advantage, in the context of news, appear to center around disparate access to resources and community dynamics. Hindman (2018) identifies several advantages enjoyed by the most prominent news outlets, including better technical infrastructure and access to advantageous partnerships. Online communities also seem to reinforce concentration dynamics in news exposure, falling into what Webster (2016) has called "obstrusive" and "unobstrusive" structures. Such dynamics reinforce the status quo of relative popularity, ensuring popular news outlets remain so.

At the same time, audience research tells a story that suggests individuals attend to a wide array of voices and sources in their online information consumption practices, even if that attention tends to follow channels that reproduce the "persistence of the popular" (Taneja & Webster, 2016). In addition, some empirical work challenges the tight association between cumulative advantage and concentration. The reception of digital content by audiences is often difficult to predict, or cannot adequately be modeled by a cumulative advantage process (Arapakis et al., 2017; Gleeson, Cellai, et al., 2014). In aggregate, news consumption is driven by interactions between agents and structures, creating the possibility for unexpected emergent behaviors at the system level (Waldherr et al., 2021; Webster, 2011). Indeed, news consumption is often incidental, driven largely by structure and context (Taneja & Yaeger, 2019). These accounts suggest a view of aggregate atten-

tion that, while still potentially concentrated, is far more volatile than the well-ordered cumulative advantage narrative would suggest.

These factors lead to a potential distinction between the *state* of news attention distributions and the *process* governing them. At any given point in time, news attention might be concentrated toward a particular outlet over others, as it benefits from a combination of institutional resources and platform favoritism. At the same time, the process underpinning news attention over time may still be fundamentally driven by uncertainty, fueling the volatility observed in news consumption more broadly.

We investigate the distribution of news source attention in online communities using computational methods that combine simulations with observational data analysis. Specifically, we consider two paradigmatic explanations for the patterns of attention in digital media that correspond to concentration and volatility respectively.

To compare these mechanisms directly, we adapt them into two formal models of source attention: one where source attention is driven by cumulative advantage and the other by stochastic fluctuation. We then use the models to conduct simulations and compare the results to data from the largest news sharing communities within Reddit, an online platform with approximately 52 million active users (Patel, 2020). To further examine the specific factors at play in audience attention, we use a predictive model to assess how the mechanisms of community behavior (Cheng, Adamic, et al., 2014), recent activity (Gleeson, Cellai, et al., 2014), and content text (Berger & Milkman, 2012) might contribute to concentration or fragmentation dynamics.

We find that attention to news sources shared on Reddit adhere to a stochastic attention model, rather than a cumulative advantage model. These results contradict prior studies of cumulative advantage in source attention online that did not consider online communities (Hindman & Rogers, 2018). Online information attention may fragment, even in the presence of countervailing tenden-

cies towards concentration. Our predictive models similarly highlight the stochastic qualities of this attention process, and they reflect an attention market weakly driven by community behavior and cyclical popularity. Multiple factors, including the design and use of features of Reddit, may shape these results. Our findings support a vision of the networked public sphere that may be at once concentrated in its distribution of attention and fragmented in where that attention focuses over time. We use these findings to propose theoretical and public agendas for understanding the role of social media in news attention allocation and sustaining balance and diversity in news attention.

## 3.2 Background

Debates over audience behavior and the allocation of news source attention have a long history in communication research that predates the internet and digital media by several decades (Rubin, 1993). Rather than attempt an exhaustive review of these discussions, we focus on a prevailing account of online news source attention dynamics (concentration around popular offerings) and the corresponding mechanism to which many prior studies attribute it (cumulative advantage). Into this framework we introduce countervailing empirical observations of attention instability—cases in which audience behavior is largely unpredictable and inconsistent. This tension, we argue, can be resolved by considering the context in which digital news attention decisions occur. A number of studies have considered how online communities and social media present distinct environments that drive substantial attention to digital news through the collective behavior of users, design affordances, and algorithmic systems. However, few (if any) studies interrogate directly whether news source attention in online communities might adhere to either of the mechanisms described above or contribute to the patterns of concentration or fragmentation.

### 3.2.1 Concentration in online news attention

The narrative that attention to news online tends to become more concentrated follows from patterns observed in many digital, networked attention systems, of varying sizes and scales. Across the internet, the largest sites consistently dominate all others in terms of traffic (Hindman & Rogers, 2018). Wikipedia traffic skews disproportionately toward the most popular pages (Ratkiewicz et al., 2010). On social media platforms, only a small fraction of posts ever "go viral" (Gleeson, Ward, et al., 2014).

The concentration of attention extends to digital news media. Large news sites maintain popularity to the exclusion of smaller sites (Hindman & Rogers, 2018). That inequality affects the viability of smaller and local outlets, as audience size and engagement drive news sites' economic prospects via advertising. Because of this commercial link, concentrated attention impacts ownership and coverage. Consolidated ownership removes space for dissenting voices and small outlets, centering news coverage around official narratives (McChesney, 2003). The imperative to attract attention and sustain traffic can further exert isomorphic pressures on content, creating a more homogeneous news ecosystem (Boczkowski & de Santos, 2007).

This concentration holds despite the abundance of media options, and proliferation of access channels, afforded by digital contexts. When presented with a wide array of available news providers, audiences stick to a narrow subset of common, popular offerings (Fletcher & Nielsen, 2017). Despite high-profile arguments to the contrary, empirical research finds that political polarization does not seal people into ideological "filter bubbles." On the contrary, many readers are regularly exposed to cross-cutting news media (Guess et al., 2018). Even in the case of misinformation, readers who fully disregard established news outlets are a slim minority (Nelson & Taneja, 2018). Overwhelmingly, empirical research demonstrates evidence for a news audience that is more connected than it is disparate, more concentrated than it is fragmented.

### 3.2.2 Mechanisms driving audience engagement with digital news

Attention dynamics are often observed in aggregate—an audience level characteristic driven by individual behaviors. Those behaviors and their interactions give rise to the driving mechanisms of concentration. Chief among them is cumulative advantage, a process of compounding concentration over time. In the case of news attention, cumulative advantage is aided by variations in audience availability, and by the disloyal audiences of smaller offerings.

Cumulative advantage describes a general mechanism by which initial resources compound, increasing resource concentrations over time (DiPrete & Eirich, 2006). In digital media markets defined by immense choice, cumulative advantage characterizes patterns of audience concentration, what Webster and Ksiazek (2012) call "the persistence of popularity."

Past empirical studies have found patterns of digital news attention consistent with cumulative advantage. Among political news sources, the bulk of traffic accrues to a small number of highly popular sites (Nelson & Webster, 2017). A simulation by Hindman and Rogers (2018) suggested that a cumulative advantage model of audience behavior reproduced empirically observed attention allocation across a wide sample of news sites. Hindman and Rogers (2018) posit several potential mechanisms driving cumulative advantage for large news sites, including better technical infrastructure, more resources for news production, and a larger traffic base from which to grow. This work provides the strongest evidence that attention to news online is both highly concentrated and characterized by cumulative advantage.

Driving this pattern of cumulative advantage are two key features of news audience behavior. First is availability. While a small subset of the news audience are heavy readers, most dedicate a limited amount of time to consuming media (Webster, 2009). As a mechanism to cope with the wide array of available media offerings, that group of casual consumers develops repertoires—subsets of (usually popular) offerings that they routinely turn to, to the exclusion of others

(Taneja et al., 2012). This dynamic creates a common, popular set of media, along with a smaller, less popular set accessed by heavy media consumers. The browsing and reading behaviors of audiences may drive them to consume both fringe and mainstream sources to varying degrees, rather than isolating in well-defined "filter bubbles" (Webster & Ksiazek, 2012). Those behaviors also align with corresponding structures of media distribution and use. Web browsing activity follows a recursive "curatorial architecture"—users start at a major hub (e.g., Google or Yahoo!), then divide out into different browsing pathways (Taneja & Wu, 2018; A. X. Wu et al., 2021).

Audience availability helps explain the shape of concentrated attention distributions, skewed toward popular offerings. Another important component is the stability of those distributions—the extent to which offerings' relative popularity remains the same over time. In examining digital news attention, Hindman and Rogers (2018) demonstrate clear variation in stability. The most widely read outlets remained firm in their position, while less popular ones struggle to secure footing. This differential treatment is partly driven by a phenomenon called double jeopardy. Because of a lack of familiarity and fewer resources, smaller offerings in many markets also experience the most disloyal audiences (Ehrenberg et al., 1990). The same holds true for digital news, where, for example, the audience for misinformation is both disloyal and dwarfed by that of mainstream media (Nelson & Taneja, 2018).

Together, variations in audience availability and double jeopardy allow concentration to take hold in digital news attention. That concentration surfaces in audience behavior, in the resources available to news organizations, and in the architecture of news distribution (Benkler et al., 2018; Hindman & Rogers, 2018). Despite the well-ordered dynamics suggested by concentration, news consumption is also defined by uncertainty. Predicting reader behavior at the individual level is a challenge, as is demonstrating a link between any feature of a news article and its ultimate popularity (Arapakis et al., 2017). To some extent, the uncertainties around audience preferences

and exposure dynamics arise from the mechanisms by which users discover news. Social media and search engines shape the discovery process by driving increasing proportions of traffic and engagement to online news (Trilling et al., 2017). On community-oriented platforms, structure and feedback govern who and what receives attention (Cheng, Adamic, et al., 2014). For news content, audiences on social media respond to certain textual and semantic cues (Berger & Milkman, 2012).

These factors interact with the broader framework of concentration, necessitating an understanding of platform-specific characteristics along with high-level governing processes. In addition, they require an understanding of how these factors, along with audience-level behavior, interact to produce system-level attention outcomes. Similar to the Webster and Ksiazek (2012) structurational theory of attention, we argue for a view of news attention that scrutinizes complex characteristics and aggregate attributes. Examinations of communication systems from the complexity perspective stress the need for fully considering the impact of context at varying levels, from user interactions to platform design (Waldherr et al., 2021). They also advocate for new models of understanding aggregate attention, often moving away from the cumulative advantage concentration narrative (Gleeson, Cellai, et al., 2014).

The question of exposure dynamics is therefore one of news consumers' interactions with third-party platforms (rather than news sources per se)—What impact do these systems have on aggregate patterns of attention? Our study addresses this question with a specific focus on social media recommender systems.

### 3.2.3 How social media recommender systems shape attention to news

Social media and recommender systems are a major source of news exposure and engagement—about 18% of U.S. adults primarily rely on social media sites to get news (Mitchell et al., 2020). Sites like Facebook, Reddit, and Google incorporate tacit as well as explicit recommendation systems

that leverage behaviors and preferences to sort, rank, and filter news content. Users of these sites interact with news content and each other, generating additional feedback into both social and algorithmic systems that drive attention around the web.

Prior research offers few direct insights into how social media sites (individually or as a whole) contribute to aggregate patterns of news attention. Most work emphasizes attention concentration (with some exceptions—see Bandy & Diakopoulos, 2021). As discussed above, audiences tend to follow well-worn, established pathways through the web, resulting in relatively predictable concentrations of viewership and consumption (Fletcher & Nielsen, 2017; Taneja & Webster, 2016; Taneja & Wu, 2018; Webster & Ksiazek, 2012). Social media and recommender systems generate their own strongly unequal, concentrated distributions of attention and popularity (F. Wu et al., 2009). These distributions recur across many types of sites and measures, including links, likes, "up-votes", views, and more (Cheng, Danescu-Niculescu-Mizil, et al., 2014; Lampe & Resnick, 2004; Salganik et al., 2006; F. Wu et al., 2009). A rise to prominence within any given site, while rare, produces a cascade of attention. Social hierarchies and status orders that emerge within social media and recommender systems further reinforce these patterns through cumulative advantage in terms of whose posts or perspectives tend to attract attention (Gilbert, 2013; Shaw, 2012).

However, just because attention follows a consistent distribution does not mean the position of any single entity within that system is pre-ordained. Indeed, the rankings produced by attention cascades can be quite difficult to predict for any given entity, even if the dynamics of the system are precisely understood (Cheng, Adamic, et al., 2014; Salganik et al., 2006). This volatility, combined with site-specific features like personalization and sorting algorithms, could contribute to audience fragmentation, even if news consumption overall remains concentrated. In this case, the state of the distribution becomes decoupled from its driving process, potentially challenging the established mechanisms of concentration.

Do the attention dynamics in social media and recommender systems reinforce or unsettle the tendency toward audience concentration among online news sources? The countervailing findings and contradictory predictions of prior research motivate the empirical inquiry that follows. As with past work reporting mixed patterns of attention dynamics overall, we do not anticipate a single, comprehensive answer that applies to all sites or audiences. Rather, we expect that specific sociotechnical affordances and configurations of social media systems may drive attention patterns in distinct ways. Empirical analysis of news source attention dynamics in these environments can inform both critical understanding as well as future design interventions to shape the networked public sphere.

## 3.3 Methods

This paper pursues a combination of observational analysis and computational simulation to understand how Reddit, one of the most prominent social media and recommender systems, contributes to the allocation of attention to news. Posts on Reddit often focus on current news items and consist of a link to an original story written and published by a different media organization (e.g., The New York Times, CNN, etc.).[1] Reddit users ("Redditors") may comment and vote on posts, producing data that the site uses to rank and sort the submissions. News stories that achieve high rankings attract millions of pageviews via Reddit users clicking through to the source (Barthel et al., 2016). These clicks and pageviews feed into the kinds of engagement metrics and outcomes that motivate our study as well as media organizations' own understanding of audience attention (C. Anderson, 2011). The rest of this section provides additional background on Reddit as the setting for our work; the sample, data, and measures we use; and a detailed explanation of the analyses.

[1] Advance Publications, itself a major media organization, owns a majority stake in Reddit but does not intervene in the system by which stories get ranked within Reddit.

### 3.3.1 Reddit as a Research Setting

Reddit is a popular social media platform with about 52 million daily active users (Patel, 2020). The site consists of volunteer-moderated groups called "subreddits," each of which focuses on a topic or theme (e.g., politics, video games, pet videos). Redditors may either submit content or respond to submissions within subreddits. Submissions ("posts") can take the form of text, multimedia elements, or hyperlinks. Every post starts out with one point. Users can comment on the post and "upvote" or "downvote" it to add or subtract a point from its total score. The site ranks and orders posts roughly according to this score in its default view, with a weighting toward newer and fast-rising content (Figure 3.1).[2] Users most often click through to the highest items in rank-ordered lists, making submission votes key to the traffic a news site ultimately receives from a subreddit (C. Barry & Lardner, 2011).

Reddit has particular relevance as a site of study for news attention. While only a fraction of news consumers uses Reddit, 70% of U.S. adults who do use it get news there (Barthel et al., 2016). This situates Reddit as both a news destination and a channel for news attention, directing readers out to news outlets via hyperlinks. In addition, Reddit's surface area for news exposure goes beyond the dedicated following who contribute to its communities. News stories can appear in a dedicated "News" tab on the mobile application's homepage, as well as in a generic feed of popular items from across the site. These features enable news exposure for casual users as well as those disproportionately invested in news-centric communities.

[2]The Reddit ranking algorithms are not public, but decay function weighting has been a part of the site for many years (Munroe, 2009). Redditors can select from a few different sorting algorithms (e.g., "What's hot" vs. "Controversial" emphasize upvote count and upvote + downvote count respectively). Additional details about Reddit appear in Chandrasekharan et al. (2018) and Gilbert (2013).

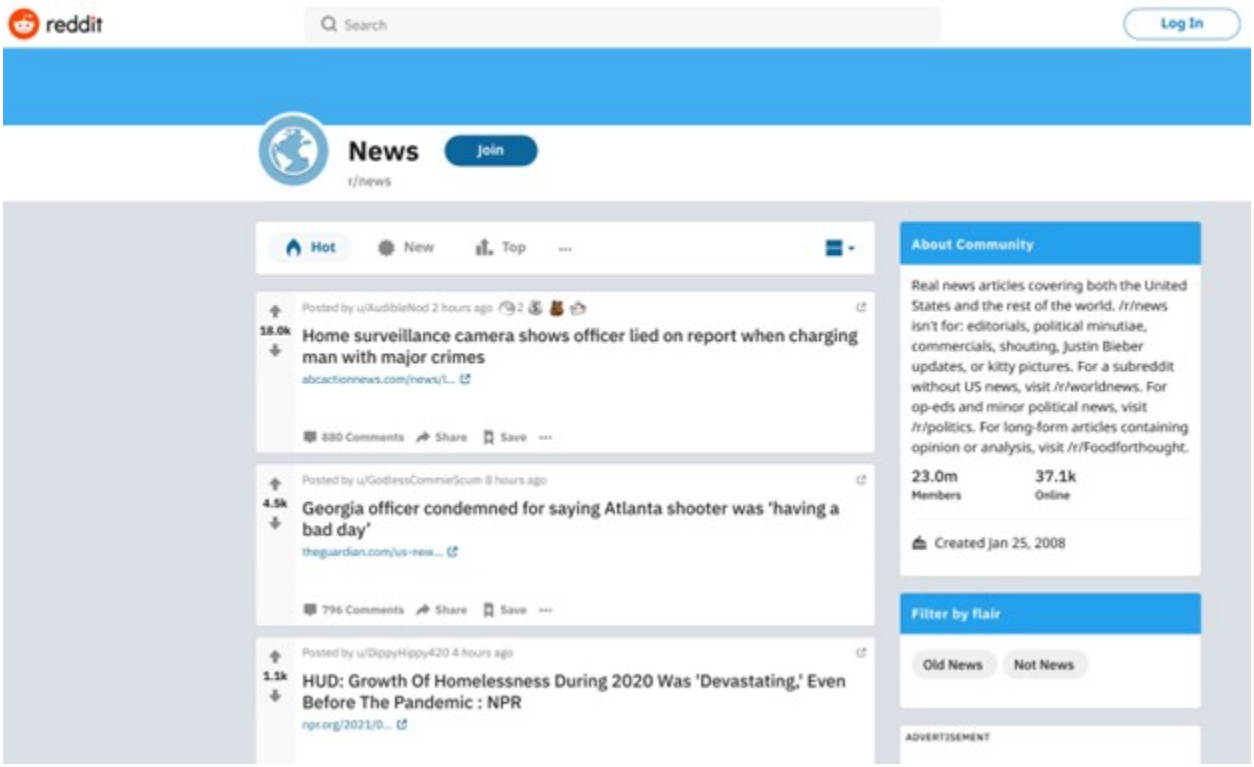



Figure 3.1: The default view of the r/news subreddit.

### 3.3.2 Data

The empirical sample is three large, news-focused subreddits: r/news, r/worldnews, and r/politics. While other popular subreddits include links to news, the three we consider consist exclusively of news links, are among the very largest subreddits (with over 60 million members combined), account for an overwhelming number of the most prominent (highly ranked) news-focused posts, and drive millions of visitors to news sources around the web (Barthel et al., 2016). As mentioned above, these subreddits are also responsible for surfacing prominent news stories throughout the rest of Reddit, extending their reach beyond the community itself. Additional information on our selection process and characteristics of each subreddit appear in the "Subreddit Details" section of Appendix B.

The data we collect and analyze for these subreddits come from the Application Programming Interfaces (APIs) of Reddit.com, PushShift[3], and the Alexa Web Information Service. We used custom scripts to perform data collection and included only public data licensed for access and reuse (although access to the Alexa data is subscription-based).

### 3.3.3 Measures and Analysis

We combine data from the sources above to construct measures at the levels of subreddits, posts, votes, and individual Redditor accounts. Additional details for these measures appear below in the corresponding descriptions of our analysis, which proceeds through two complementary studies: Study I builds theoretically-motivated simulations of news source attention patterns in subreddits and compares the simulated results against empirically observed data. Study II expands the scope of the inquiry and uses a machine learning classifier to predict source attention rankings based on attributes of the Reddit posts, news sources, and news content.

[3]https://pushshift.io, A third-party site managed by Jason Baumgartner that publishes a large corpus of Reddit data

### *3.3.3 Study I: Simulating Source Attention*

We use computational simulations to approximate the impact of cumulative advantage and stochasticity respectively on source attention allocation in news subreddits. Formal models of each mechanism generate hypothetical daily submission scores across news domains within each subreddit as though the process were governed by either cumulative advantage or stochasticity alone. We then compare the two sets of simulation results against the data in our sample. In designing the simulations, we build on the work of Hindman and Rogers (2018), who adopt a similar approach to analyze web traffic data.

The simulations and the empirical benchmarks against which we evaluate them draw on submissions data for r/news, r/politics, and r/worldnews from PushShift. Our sample ranges from January 1, 2018, to September 30, 2018, and contains 630,248 total submissions. Submissions' scores reflect PushShift's most recent record at the time of collection. To identify the news source, we collect the top-level domain (TLD) to which each submission links. We removed any TLD that averaged one point per submission or less, given that voting on Reddit is "underprovisioned" (Gilbert, 2013). Table 3.1 contains full descriptive statistics for the sample.

| **Subreddit** | **Posts** | **Sites** | **Minimum score** | **Median score per site** | **Maximum score** |
|---|---|---|---|---|---|
| r/news | 167,082 | 4,423 | 2 | 53 | 2,048,504 |
| r/politics | 293,824 | 2,204 | 2 | 408 | 20,586,977 |
| r/worldnews | 169,342 | 3,966 | 2 | 34 | 4,942,395 |

Table 3.1: Descriptive subreddit statistics, detailing the volume of posts, unique domains posted, and score range for each.

To simulate cumulative advantage as a mechanism of source attention, we adapt the model underpinning the web traffic simulation in Hindman and Rogers (2018). Formally, our model is:

$$s_{x_j(t+1)} = \gamma(s_{x_j(t)} + g_j s x_j(t) + 1) \tag{3.1}$$

Equation 3.1 expresses the mean submission score $s$ of a website (top-level domain) $x$ in rank $j$ one day in the future ($t + 1$), as a function of the current day's mean submission performance ($x_j(t)$) and the percentage daily growth rate ($g_j$). We derive the daily growth rate $g_j$ empirically, by drawing from all observed daily growth rates in our sample. We also include an indicator term ($\gamma$) to capture whether a site has any submissions in the subreddit on a given day. This value was drawn from a weighted binomial distribution, modeled on the empirical probability distribution—$P(present)_{x_j(t)}$—that a site had any submissions in a subreddit on a given day. Since submissions start with one point, we also added one to the mean performance on each day.

To simulate fragmentation as a mechanism of source attention, we use the stochastic model expressed in Equation 3.2:

$$s_{x_j(t)} = \gamma(\theta + 1) \tag{3.2}$$

In this model, $s$, $x$, $j$, and $t$ have the same meaning as before. We introduce $\theta$, a daily performance score, which we randomly draw from the distribution of observed domain performance per day. This ensures that simulated outcomes reflect the characteristics of our sample.

To mirror our sample, we simulated each model for 273 days, initialized with the same number of sites as in the observed data (Table 1). We used mean score per submission[4] as our domain-level daily performance measure. We initialized each site with a total score randomly sampled from across each subreddit's entire distribution.

We compared each simulation to our empirical data in terms of how well it reproduced the

[4]Results do not change using median score per submission.

observed inequality and stability of attention. To measure inequality, we calculated the Gini coefficient of mean points per submission per news domain.[5] To measure stability, we first ranked news domains by their mean submission score per day. Rankings allow for scale-free, robust comparisons across days and subreddits. For each domain, we constructed a sequence of the ranks that domain inhabited. We treated those sequences as Markov chains, using them to calculate the probability of moving between any two ranks (Makhortykh et al., 2021). We then transformed these probabilities to normalized Z-scores to facilitate robust comparisons over subreddits and time. We encoded days on which a site did not appear in the subreddit as 0-ranks. Because the number of domains present varies by day and subreddit, we focus on transition probabilities among the top 50 ranks, the maximum number present for 95% of days across subreddits.[6]

Finally, we compared the two distributions of simulated, normalized transition probabilities to those calculated from the empirical data. We used visualizations and calculated the Jensen-Shannon Divergence (JSD) between each simulated and observed distribution. JSD provides an information-theoretic, scale-free measure of "distance" between two pairs of probability distributions (see Klingenstein et al., 2014, for a related application), allowing us to evaluate which simulated model more closely approximates the observed data.

### *3.3.3 Study II: Predicting Source Attention*

To explore specific factors that might drive either cumulative advantage or fragmentation in source attention dynamics, we use machine learning in Study II. Prior work on attention dynamics in online communities and social media suggests several important predictors of attention allocation, including community behavior, recent activity, and features of content presentation. Study I cannot

[5]Calculating the Gini coefficient for total points produces higher values but does not impact our conclusions.

[6]Since we are using pairwise probabilities between ranks, the results presented here are not impacted by the ranks filtered out by our threshold. We conducted two additional analyses with identical results—one which included the top 100 ranks, and one which had no cutoff.

differentiate between any of these, so we designed a machine-learned classification model to do so. This analysis extends the inquiry into cumulative advantage and fragmentation as attention-driving mechanisms: Attention predictability is high in the former, but low in the latter. The machine learning model allows us to directly evaluate predictability and the types of features that affect it.

We generated four types of features (Table 3.2): cumulative advantage, community behavior, recent activity, and content text. The motivation and operationalization of each category follow.

| **Feature** | **Category** |
|---|---|
| Domain site speed (seconds) | Cumulative advantage |
| Domain site speed (percentile) | Cumulative advantage |
| Inbound links to domain | Cumulative advantage |
| Domain web traffic rank | Cumulative advantage |
| Total users submitted | User |
| Total submissions | User |
| Submitted by active, high-status contributor | User |
| Mean submission score, previous 7 days | Recent activity |
| Time between first and most recent submission | Recent activity |
| Percent of past 7 days present | Recent activity |
| Mean submission sentiment | Content text |
| Minimum submission sentiment | Content text |
| Maximum submission sentiment | Content text |
| Mean submission word count | Content text |
| Minimum submission word count | Content text |
| Maximum submission word count | Content text |
| Domain | Metadata |

Table 3.2: Features used to train the random forest classifier on mechanisms driving attention allocation in news subreddits.

Cumulative advantage features incorporate additional off-platform differences in resources and prominence that may account for shifts in news source popularity as characterized by Hindman and Rogers (2018). Using the Alexa Web Information Service API[7], we gathered how fast each site

[7]https://awis.alexa.com/developer-guide

loads, its total inbound links, and its overall traffic ranking across the web.

Community behavior features represent more specific interactions between Reddit users and a domain, to capture the influence of community structure and feedback on source performance (Cheng, Danescu-Niculescu-Mizil, et al., 2014). For each day, we counted the total number of users who shared at least one submission from a domain, and the number of submissions shared from that domain in total. Since prior work associates user status with post performance, we identified Reddit users in our sample whose submission count and score per submission were above average (Gilbert, 2013). We used this information to generate a binary indicator of whether a domain was shared by a more active, higher status contributor.

Recent activity features reflect the idea, as shown in Gleeson, Cellai, et al. (2014), that recent activity is more predictive of performance than cumulative activity. For each domain, we calculated a seven-day moving average submission score, the time difference between a domain's first and most recent submission within our sample, and the percent of the past seven days for which a domain had at least one submission. Content text features are motivated by prior work demonstrating how headline writing style affects a news story's spread on social media (Berger & Milkman, 2012). We calculated the sentiment (F. A. Nielsen, 2011) and length (Kuiken et al., 2017) of each submission's text. In both cases, we used the daily minimum, maximum, and mean values. We included one additional metadata feature—the domain associated with each entry—to allow the model to learn continuity across days for a single domain.

Using these features, we trained a random forest classifier to predict daily domain performance within each subreddit. We use the Ranger R package's random forest implementation, with 100 decision trees and otherwise default tuning parameters (Wright & Ziegler, 2015). Rather than forecast exact rankings, this model predicted whether a domain would perform above a percentile threshold. We evaluated models with 50%, 25%, 10%, and 1% thresholds. This approach follows

| | r/news | | r/politics | | r/worldnews | |
|---|---|---|---|---|---|---|
| | Present at $t+1$ | Absent at $t+1$ | Present at $t+1$ | Absent at $t+1$ | Present at $t+1$ | Absent at $t+1$ |
| Present at $t$ | 0.46 | 0.54 | 0.59 | 0.41 | 0.5 | 0.5 |
| Absent at $t$ | 0.03 | 0.97 | 0.05 | 0.95 | 0.03 | 0.97 |

Table 3.3: Probabilities that a domain moves in or out of each subreddit from day to day, given that domain's starting state. Sites that start out in a subreddit have about a 50% chance of remaining. Sites that start out absent almost always remain so on the next day.

past work using thresholds to make tractable prediction tasks out of complex ranking problems (Arapakis et al., 2017; Cheng, Adamic, et al., 2014). We use balanced accuracy to summarize model performance because it corrects for imbalance between classes with different numbers of observations.

## 3.4 Results

### 3.4.1 Study I: Simulating Source Attention

We first summarize the empirically observed characteristics of news source attention dynamics in the data from the three subreddits. Attention concentration is high in all three subreddits, as evidenced by the Gini coefficients of their mean points per submissions across domains: $G(news) = 0.92$, $G(politics) = 0.78$, and $G(worldnews) = 0.93$.

Attention to specific news sources is far from stable in these subreddits, however. Table 3.3 shows the within-subreddit probability that any news source present (absent) on one day is present (absent) on the following day. Most sources do not appear on sequential days. Sources that do appear have about even odds ($\mu = 48\%$) of not appearing multiple days in a row.

When domains do appear on sequential days, there is little consistency in how they rank in terms of upvotes received. Row A in Figure 3.2 uses a heatmap to visualize normalized between-rank transition probabilities from one day to the next for each subreddit. The absence of any large

transition probabilities between any pairs of ranks (which would appear as lighter, "hotter" cells in the grid) indicates the absence of stability. A news source that receives an abundance of upvotes and attention one day is not likely to do so the following day.

In summary, these results indicate high concentration of news source attention within subreddits, but also high instability of attention to specific news sources. Further analyses underscore these findings and appear in Appendix B, under "Additional Descriptive Analyses". We use these empirical results to evaluate our simulations. In addition, further simulations at weekly and monthly resolutions produce similar results.

In contrast to the empirical data, the simulated cumulative advantage model produces extremely high concentration ($G = 1$ for all three subreddits) and highly stable attention. Row B in Figure 3.2 shows the normalized rank transition probabilities derived from this simulation. These results align well with a theoretical cumulative advantage system: Top ranks change hands infrequently (the joint transition probabilities are lighter and "hotter" in the top-right of each grid), while lower-ranked sites show higher volatility.

The simulation of a stochastic model of attention allocation replicates both the attention concentration and rank transition instability we find in the observed data. In terms of attention concentration, the Gini coefficients from the simulation resemble the observed values closely—$G(news_{sim}) = 0.88$; $G(politics_{sim}) = 0.6$; $G(worldnews_{sim}) = 0.88$. Row C of Figure 3.2 shows that the model also produces rank transition probabilities consistent with those we observe empirically—all ranks experience uniform volatility.

The Jensen-Shannon Divergence (JSD) between each simulated and empirical rank's probability distributions confirm the visual evidence and the proximity of the stochastic simulation to the observed data. Comparing the stochastic simulation to the empirical data produces a mean JSD value that is, depending on the subreddit, 1.6 to 3 times lower than in the cumulative advan-

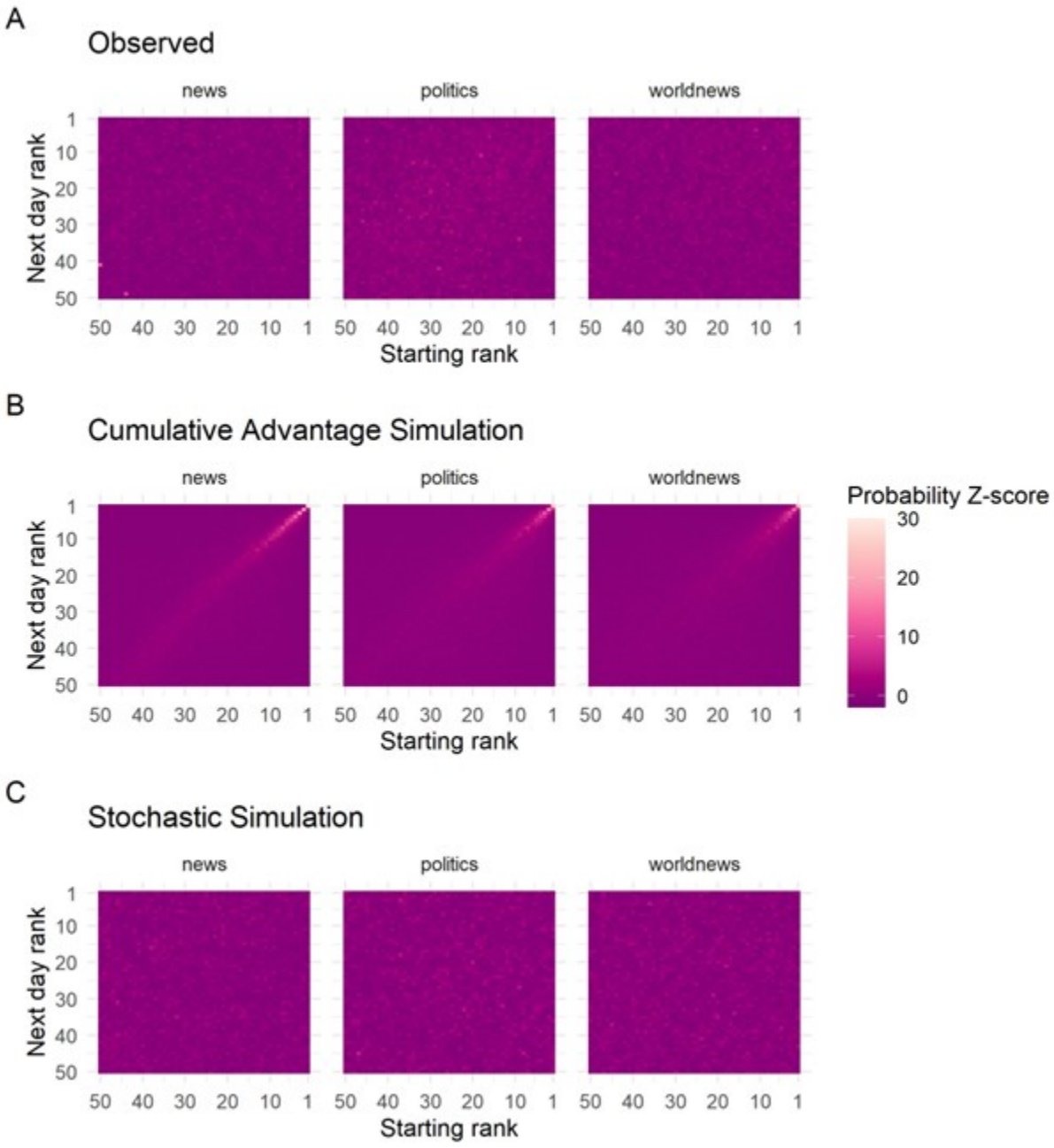



Figure 3.2: Study I: Normalized transition probabilities between all ranks across subreddits. Here we show results from our empirically observed data (A), the cumulative advantage model (B), and the stochastic model (C).

tage comparison. Across all ranks, the stochastic model better approximates day-to-day domain performance (Figure 3.3).

### 3.4.2 Study II: Predicting Source Attention

The simulation results indicate that news domain attention in subreddits resembles a stochastic process much more closely than a cumulative advantage process. But perhaps the models are just impoverished? Maybe other factors about the subreddits, the stories involved, or the news sources themselves might help explain the observed (stochastic) variations in attention and resolve the discrepancies between our observations and prior work documenting cumulative advantage? The results of our statistical prediction models suggest otherwise.

A summary of each prediction model's performance appears in Figure 3.4. The least restrictive models—those predicting whether a domain will wind up in the top or bottom 50% of the score distribution—perform best, peaking at 0.74 classification accuracy for the r/politics subreddit. The performance declines across the more restrictive models, with the models using 10% and 1% cutoffs performing no better than a random 50% classification accuracy baseline. The drop-off suggests that the models discriminate somewhat between the worst and best performing sources but fail to determine what features distinguish the highest performers.

To see which features the models identify as most predictive of domain attention, we examine the permutation feature importance from the most successful model—the 50% threshold model for the r/politics subreddit (Table 3.4). By far, the most predictive features come from the community behavior and recent activity categories. In particular, the 7-day average submission score (0.08) and shared by active, high-status user (0.03) features have the highest permutation importance overall. Cumulative advantage and content text features add very little predictive power to the model, with a maximum permutation importance of 0.01. Across all models, none of the features produce large

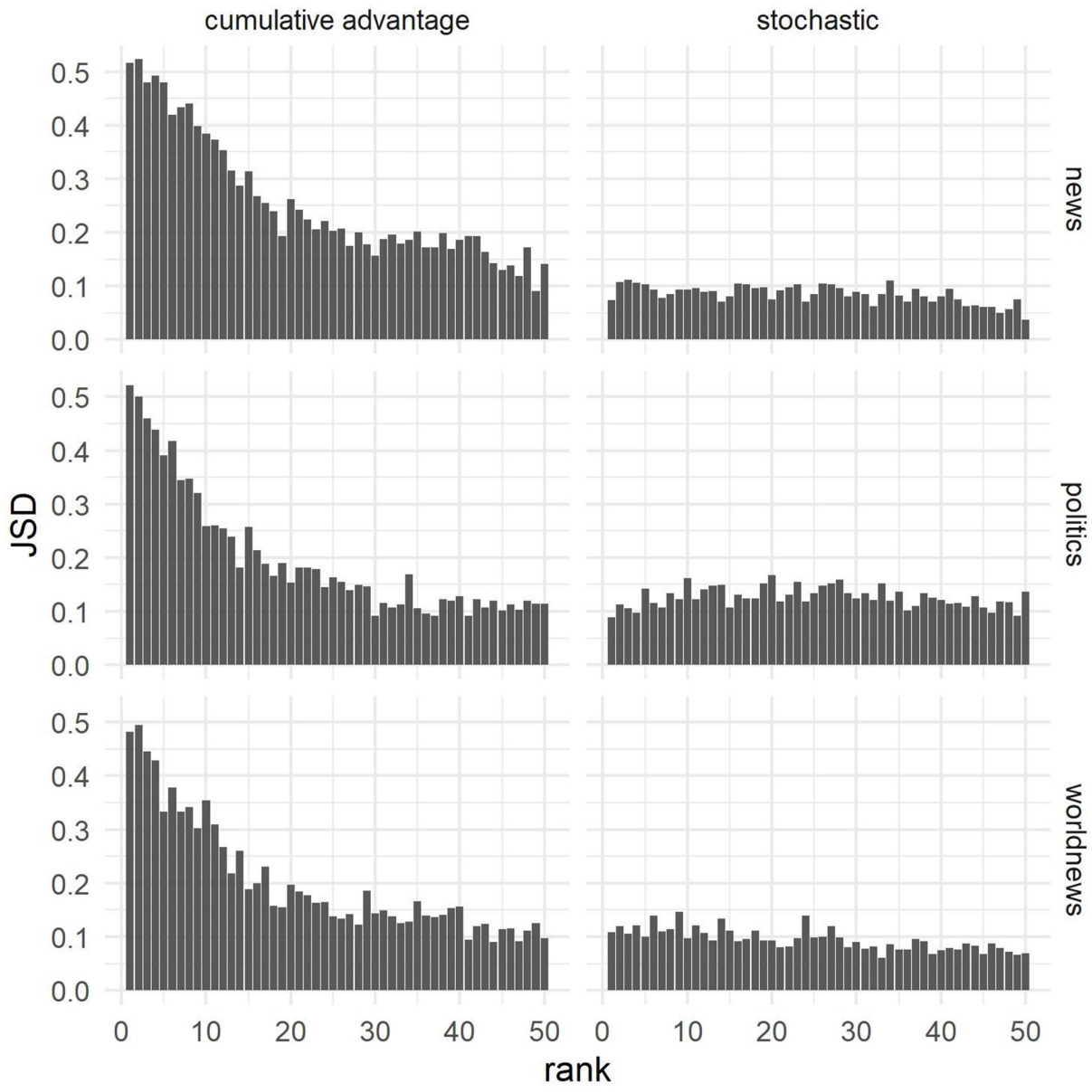



Figure 3.3: Study I: JSD measures for each rank, with each simulation compared to empirical data across subreddits. Overall lower JSD values in the stochastic simulation indicate a better fit to empirically observed behavior.

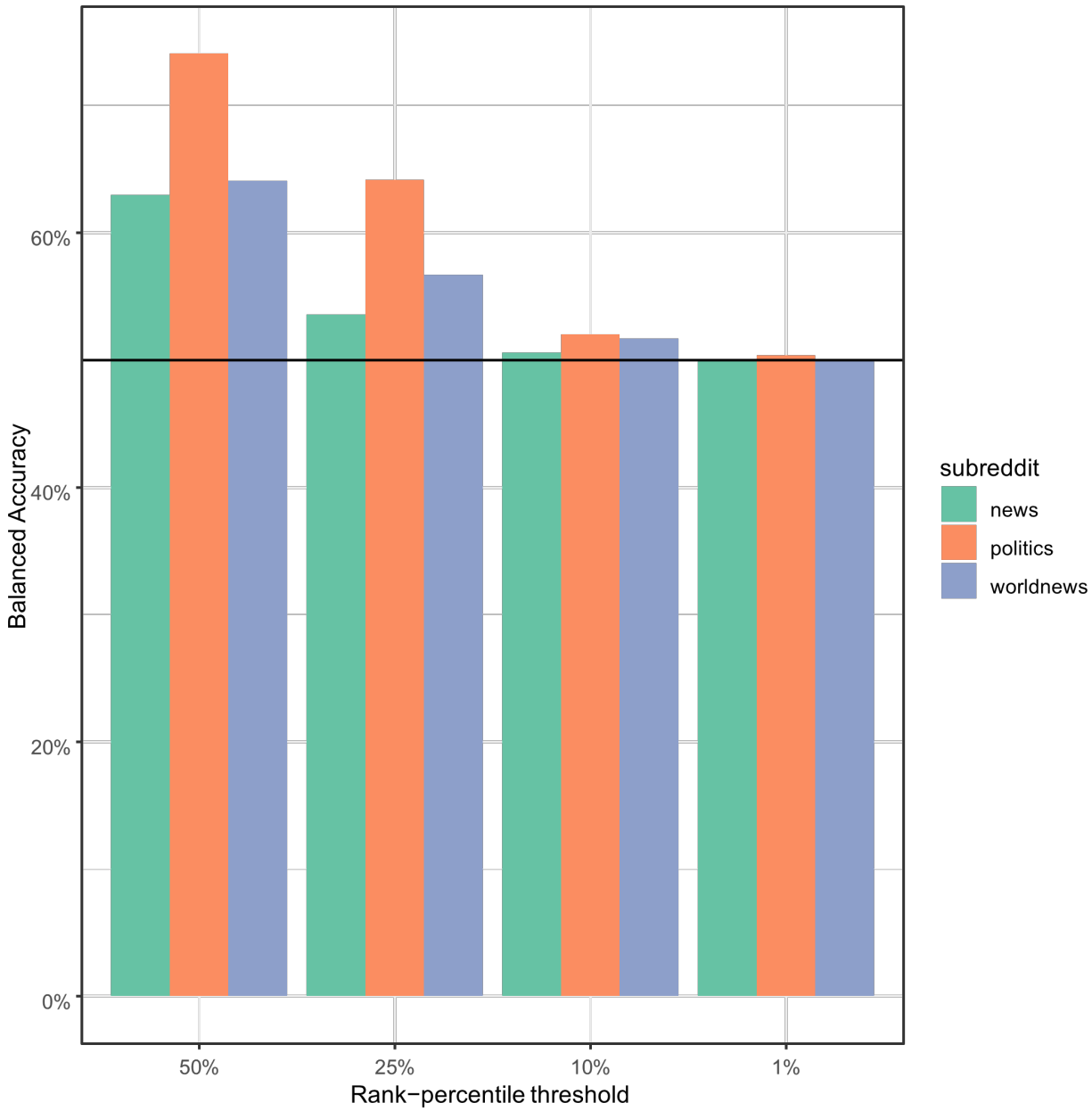



Figure 3.4: Study II: Prediction model classification accuracy by subreddit and rank-percentile threshold. The horizontal line indicates a random (50% classification balanced accuracy) baseline. The models distinguish between sites that perform in the top or bottom half of the attention distribution. Performance declines as the threshold gets more restrictive.

predictive improvements, but are rather almost equally unimportant to model performance. These results indicate that the models cannot predict daily news source attention and lend additional, indirect support to the stochastic model of attention allocation.

| Feature | Importance |
|---|---|
| Mean submission score, previous 7 days | 0.081 |
| Submitted by active, high-status contributor | 0.027 |
| Total users submitted | 0.021 |
| Percent of past 7 days present | 0.020 |
| Total submissions | 0.018 |
| Maximum submission word count | 0.011 |
| Inbound links to domain | 0.009 |
| Domain site speed (percentile) | 0.008 |
| Mean submission word count | 0.008 |
| Domain site speed (seconds) | 0.008 |
| Domain web traffic rank | 0.007 |
| Time between first and most recent submission | 0.006 |
| Minimum submission sentiment | 0.006 |
| Maximum submission sentiment | 0.005 |
| Mean submission sentiment | 0.004 |
| Domain | 0.004 |
| Minimum submission word count | 0.004 |

Table 3.4: Permutation importance values for each feature, in the r/politics classification model predicting whether domains appeared in the upper 50% on a given day.

## 3.5 Discussion

News source attention on Reddit follows a highly concentrated distribution, but in a pattern that resembles a stochastic process far more closely than a cumulative advantage process. This is inconsistent with the strongest forms of the concentration narrative developed in prior work (e.g., Hindman & Rogers, 2018) and suggests that online communities can contribute to the fragmentation of news source attention. At the subreddit level, high Gini coefficients indicate substantial

concentration in source attention. However, we find extreme instability in which sources accumulate attention from one day to the next—a result incompatible with an explanation based on cumulative advantage. A stochastic allocation model approximates the observed domain ranking patterns more closely. Furthermore, the predictability of domain performance is low.

These results nuance our understanding of audience-level news attention dynamics in a few ways. First, they extend the Webster (2011) structurational framework of attention. That framework emphasizes the importance of interactions—between agents and structures—in our understanding of audience-level dynamics. With this emphasis, Webster (2011) lays the groundwork for system-level studies of complex dynamics. Recent calls for a complexity perspective in communication research mirror that approach, and we take it up here (Waldherr et al., 2021). In doing so, we pay particular attention to separating the state of news attention from the process driving it (Simon, 1991). Our results reflect a state seen in prior work—that of extremely skewed attention in favor of popular offerings. However, they also demonstrate a clear detour from established process. Prior work argues that instability and concentration can co-exist, or are even causally linked to some extent (Hindman & Rogers, 2018). However, in this conceptualization of the mechanism underlying concentration, the volatility of individual decision making still results in stability over time. In contrast, here we find an apparent lack of any pattern in attention allocation. The process we observe does not fundamentally alter the *distributional* characteristics of attention, but it does reduce the predictability of individual performance.

Our findings also highlight the potential of applying alternative models of popularity in digital cultural markets more generally, outside of digital news. In evaluating popularity dynamics on social media and in app marketplaces, researchers have shown cumulative advantage to be a relatively poor predictor of popularity (Gleeson, Cellai, et al., 2014; Shulman et al., 2016). Rather, as we see here, popularity follows a cyclical "boom and bust" cycle (Gleeson, Cellai, et al., 2014).

Lorenz-Spreen et al. (2019) model similar collective behavior across several contexts, arguing that the cycle of popularity growth, saturation, and decline in attention markets is accelerating. Our examination of digital news may therefore represent one instance of a broader trend, away from treating popularity as cumulative, and toward adapting a cyclical framework of attention markets. We refer this possibility to future research.

For system designers, the behavior shown here highlights potential tensions between algorithmic recommenders and news and information delivery. As described above, Reddit's algorithmic sorting introduces decay into post ranking over time, by design. This characteristic works to continually inject novelty into subreddits. And while decay on its own doesn't inherently preclude authoritative news outlets from maintaining prominence, our findings suggest that the manner in which users interact with Reddit's systems does create volatile patterns of attention. From one perspective, these results suggest a need for more explicit guardrails to ensure stability. However, from another perspective, the volatility of source selection seen here is a step toward maximizing viewpoint diversity. The system may operate with minimal gatekeeping, favoring a participatory selection and ranking process driven by community values (Helberger, 2019). Several characteristics of the system—its mode of signaling information authority, how it incorporates user decision making, and the available opportunities for expressing communal agency—should be explicitly considered in order to facilitate the desired balance between stability and expression.

Returning to the complexity perspective, we situate the aggregate outcomes observed here as emergent properties of interactions. Rather than attempting to wholesale displace established mechanisms of audience attention, we argue that the contexts studied in this work modulate those mechanisms. Audiences, operating on digital platforms, put in contact with sorting and recommendation algorithms, and placed within explicit community contexts, behave in unexpected ways. At an individual level, the behaviors of news consumers may follow expected patterns: turning to

consistent sources of information (Makhortykh et al., 2021); responding to the framing, political orientation, or emotional valence of news coverage (Hasell & Weeks, 2016); or gravitating toward salient topics (Tewksbury & Riles, 2015). But at the same time, those individuals interact with simultaneous acts of curation—by news sources, by Reddit users, and by the platform itself (Thorson & Wells, 2016). Those interactions shape the context of each individual news ranking choice, which, when aggregated, seem to drastically affect how news audiences prioritize information. As a result, future studies of community-level governance and participant behavior may reveal more precise mechanisms by which news source attention gets allocated in online communities. In line with Gilbert (2013), our results suggest that one of the attributes of subreddit community participants—whether a domain was submitted by a high-status user—improved the accuracy of our prediction models more than other features. This suggests that community-level participation, institutions, and status might shape news attention allocation more than factors about news sources, but we recommend treating such a claim as exploratory based on the evidence presented here. Future work can investigate these findings more deeply across multiple communities and formally test this result against alternative explanations.

### 3.5.1 Limitations

This study examines three communities in one social media platform to broaden our understanding of attention allocation. However, there are many other digital attention markets that may behave differently, and that warrant further scrutiny. Reddit is both a channel for news attention and a destination itself, representing only a small slice of news consumption on the internet. That slice is not necessarily representative of all news consumers, as Barthel et al. (2016) note important demographic differences between Reddit users and the general population. In addition, our sample covers a limited timeframe, and news content may appear outside of these three major subreddits.

Voting is also likely a conservative measure of attention. Many more users are exposed to, and likely engage with, a news story than we capture here. The "Robustness Checks" section of Appendix B contains several robustness checks designed to ensure that subreddit-specific rules or submission behaviors do not invalidate our use of voting behavior as an attention measure. We also conducted robustness checks to ensure that unusual submission-level behavior did not influence the aggregate dynamics we observe.

Finally, our approach relies on using top-level web domains (TLDs) as proxies for individual news organizations. There are some cases where this may not be a 1-to-1 mapping. For example, a story might be syndicated across many different sites from one outlet. As such, our approach is only able to account for dynamics among distinct sources of news as they appear on Reddit, not the more complex relationships on the production side of digital news.

### 3.5.2 Conclusions

This study paints a picture of news attention driven not by cumulative advantage, but by boom-and-bust cycles that happen at random. However, even in the absence of any discernible pattern of attention, extreme inequality in vote allocation still emerges. As such, it is crucial to identify alternative modes of organizing collective information resources with this emergent inequality in mind. Public goods like digital news require a stable, equitable share of attention to effectively serve citizens' information needs. Platforms' current approaches to shaping news attention fall short.

Social media and recommender systems may play multiple and contradictory roles within larger attention markets. The results of this study indicate that news-focused subreddits contribute to the instability of news source attention on the web, while reinforcing its concentration. Future research should seek additional means of evaluating these conclusions and compare them across

multiple social media platforms and recommendation systems. The results we observe here may be driven, in part, by features of Reddit that contribute to the rapid turnover of rankings on the site (e.g., the use of an exponential decay function in the ranking algorithm that tends to prioritize newer content). By understanding how different sites and systems generate specific distributions of attention, future research can inform the design of interventions that correct persistent inequalities or attention market failures in the networked public sphere.

# CHAPTER 4
# ANTICIPATING ATTENTION: ON THE PREDICTABILITY OF NEWS HEADLINE TESTS

## 4.1 Introduction

Newsrooms increasingly optimize their output for audience attention in both coverage areas and presentation (C. Anderson, 2011; Fürst, 2020; Petre, 2015) hoping to ensure their viability in an attention-dependent digital media marketplace (Webster, 2016). But it's hard to judge how much a news producer's decisions actually matter in attracting audience attention. To be sure, news audiences are influenced by the articles, modes of presentation, and writing styles that news organizations employ (Jungherr et al., 2019; Kuiken et al., 2017). However, readers also rely on their personal preferences and backgrounds when composing their news diets (Garrett & Stroud, 2014; Lamberson & Soroka, 2018). The interactions between these influences are difficult to untangle (Kessler & Engelmann, 2019).

Underlying these influences is the process of datafication, by which audience behavior is quantified, encoded in standardized formats, and presented back to news producers (Van Dijck, 2014). In Chapter 3, datafication took the form of codified audience preferences presented as rankings. In that case, platform features enabled an ongoing expression of collective audience priorities. Those priorities were highly visible to the audience, could be directly altered in the form of casting votes, and operated on artifacts provided by members of the audience. In this study, we turn our focus to a different process of datafication, one that happens at a micro scale, largely invisible to news consumers, and provides direct feedback to journalists. In doing so, we examine the possibility

of encoded audience data to feed into a desire to optimize consumer behavior, in turn shaping newsroom strategy.

One focus of audience-facing optimization is the headline, the first (and often only) piece of text a reader might see from a news article. Headlines must summarize articles' contents and entice readers to click; a goal of headline optimization is to accomplish both with a concise and attractive writing style (Dor, 2003). Many studies have attempted to detail effective, attention-grabbing headline writing strategies (Kim et al., 2016; Kuiken et al., 2017; Rayson, 2017). A/B testing allows journalists and editors to write multiple variations of an article's headline, display those variations to different members of the audience, then select the best-performing version (Hagar & Diakopoulos, 2019).

Many of these efforts stop short of addressing broader questions about how news attention works online. By considering headline writing in isolation, they tend to overlook the role of audience preferences, attitudes, and habits (Kormelink & Meijer, 2018). Predicting the performance of any piece of digital content solely based on its composition is a known challenge (Arapakis et al., 2017; Martin et al., 2016). There is a tension, then, between the individual decisions being encoded by A/B testing and the aggregate information conveyed to journalists. The data captured by testing systems do not fully represent the intention or context news consumers bring to force when clicking on a headline. But journalists often attribute the outcomes of A/B tests to the way a headline is written, neglecting those hidden factors (Hagar & Diakopoulos, 2019).

This work interrogates the relationship between headline writing style, audience factors, and performance. We focus on headlines that appear on news publisher home and section pages, in the context of A/B tests, using a large-scale, real-world dataset. First, we examine the extent to which a headline's textual features can predict its performance. Second, within those bounds of predictability, we demonstrate the relative contribution of content-based features to headline

performance.

Our predictive model achieves modest success relative to our baselines while also demonstrating limits to content-based prediction. To further analyze the importance of writing style, we draw from a comprehensive range of theoretically and empirically motivated textual features. We show that, while several features' outcomes agree directionally with prior work, associations between feature usage and headline performance are weak overall. These findings suggest areas outside of a headline's composition, such as audience behavior and preferences, may warrant further study in understanding and predicting headline performance.

## 4.2 Background

In this section we consider related work on news audience attention and its relation to headlines. We also introduce background on the predictive approach we take in modeling that attention.

### 4.2.1 Audience Attention and Headlines

Perceptions, preferences, and beliefs shape what news sources a reader might seek out. Similarly, partisan preferences and attitudes toward news organizations impact what kinds of sources readers get exposed to (Flaxman et al., 2016; Kessler & Engelmann, 2019). A reader's familiarity with a piece of news, and their interest in that news, can affect the extent to which it draws their attention (Lamberson & Soroka, 2018). In addition, readers approach news with ingrained habits and routines of consumption (Makhortykh et al., 2021). Individual preconceptions—attitudes and behaviors that are difficult for a news organization to alter—play a significant role in news attention decisions.

Journalists and editors pay special attention to writing effective headlines. In print, an effective headline summarizes or highlights an article's most interesting points (Ifantidou, 2009). Digital

headlines tend to prioritize drawing in an audience from across distribution platforms, since news organizations often depend on reader clicks for ad revenue and subsequent subscription revenue (C. Anderson, 2011; Petre, 2015). That shift conflates what headline writers consider "good"—a value judgment traditionally based on professional standards of craft and ethics—and what gets clicked on the most, as seen in practices of algorithmically-optimized content distribution such as headline A/B testing (Diakopoulos, 2019; Hagar & Diakopoulos, 2019; Ross, 2017).

Rather than cultivating independent, editorial judgment, journalists may shift focus toward what the audience demands (Klinenberg, 2005; Ross, 2017). They in turn may place a greater emphasis on the kinds of coverage (e.g., soft news) and news values (e.g., proximity) that online audiences consider newsworthy (Trilling et al., 2017). In terms of headline optimization, this can result in clickbait, which attempts to generate curiosity by implicitly referencing material in the article without revealing its details (Blom & Hansen, 2015). These changes represent a shift in the specific qualities practitioners uphold as best practices in headline writing.

Using the direct measurement of audience response made possible by testing and other analytics tools, researchers and practitioners can more precisely evaluate the performance impact of writing strategies. We take this evaluation a step further by predicting headlines' performance from their contents, providing a view of how broadly writing approaches relate to the attraction and optimization of audience attention.

### 4.2.2 Headline Performance: From Explanation to Prediction

Using audience data, prior research has examined the effects of specific linguistic features on headline performance. Kuiken et al. (2017) measured headlines' click through rates in email newsletters and found a variety of textual features with a positive impact on headline performance, including average word length, lack of interrogatives, absence of quotes, use of personal or possessive pro-

nouns, and presence of sentimental words. Industry researchers studied the performance of 100 million headlines on Facebook, albeit not all from news publishers, to extract specific phrases and emotions that elicited strong engagement (Rayson, 2017). Kim et al. (2016) used news article click through rates from the Yahoo! homepage to evaluate the performance impact of various words and parts of speech. Much of this prior work operates through an explanatory lens: Using statistical approaches, studies attempt to demonstrate the extent to which proposed mechanisms are plausible drivers of headline success. In contrast, our research adopts a predictive approach, oriented toward developing models to predict unknown outcomes from previous observations.

Predictive modeling is complementary to more explanatory empirical work. First, predictive models help uncover new phenomena of interest. By identifying predictors that improve explanatory models, a predictive lens can enhance our empirical understanding of certain outcomes (Hindman, 2015; Shmueli, 2010). Second, prediction helps to establish the limits of what we might hope to understand about a phenomenon. As Tetlock and Gardner (2015) point out, knowing the limits of an outcome's predictability is itself valuable, in that it provides vital context to any accuracy measurements. Shmueli (2010) reinforces this notion, arguing that predictive models can help establish benchmarks for an outcome's potential explainability. Finally, predictive models help gauge the distance between theory and practice, testing how well proposed theoretical mechanisms apply in a given practical context (Shmueli, 2010). As we elaborate further below, the features that we operationalize to support our predictive model are theoretically motivated, and the usefulness of those features in the model help to establish the external validity of those theoretical ideas in the specific context of news headline performance (Margolin, 2019).

Predicting performance outcomes based on content alone is a known challenge. In many cases, predictive models require some early performance data from which to extrapolate (Szabo & Huberman, 2010). Other factors, such as social influence on digital platforms, have been shown to

affect performance outcomes more than content itself (Salganik et al., 2006). This prediction difficulty also holds true for news headlines (Arapakis et al., 2017). Applying a predictive lens to any content-based performance outcome should seek to establish bounds on predictability. Our work addresses the following questions:

RQ1: To what extent can the text of a headline predict its performance?

RQ2: What is the relative importance of various content-based features to headline performance?

## 4.3 Data

Our data come from Chartbeat, a company that provides analytics services to digital publishers. This includes their Engaged Headline Testing system, which experimentally compares multiple versions of an article's headline to determine which is most effective at attracting readers. In a headline test, the system presents different readers with different headline variants for the same article and measures how many people click on each variant. As differences in performance emerge across variants, the system shows higher-performing variants to a larger portion of the site's audience.[1] Once the system is confident of a statistically significant difference in performance across variants, it marks a test as "converged" (described in more detail below).[2]

For each test, these data contain the text of the headline variants, each variant's associated performance in terms of clicks and impressions, and metadata about when and on which page a test was run. All tests in our sample take place on the homepages or section pages of news sites. The dataset represents direct comparisons of headline constructions, with real readers, across many news sites of different sizes and types. Because each headline variant is only compared to other

[1] http://support.chartbeat.com/edu/headlinetesting/methodology.html
[2] http://support.chartbeat.com/edu/headlinetesting/orientationguide.html

variants within its test, and each test corresponds to one article, these data are well-suited for isolating the way a headline is written from the contents of its corresponding article.

### 4.3.1 Data Filtering

The complete Chartbeat dataset represents 1,023,996 A/B headline tests with 2,662,572 headline variants, run across 1,314 web domains between April 1, 2015, and April 30, 2020. To limit our analysis to tests with clear results and clean data, we first filtered out certain classes of tests. Since the statistical models used in our natural language processing pipeline were trained on English corpora, we filtered out any tests with a headline not written in English. We did this by first removing all headline tests run on domains that were manually determined to publish articles in a language other than English, then by removing tests with at least one variant tagged as consisting of more than 20% non-English words. Next, we excluded any anomalous tests with a headline that had zero clicks recorded. In addition, we removed any tests for which the system did not reach statistical confidence about the winning variant[3], including those that were prematurely canceled by a user. Since publishers can run A/B tests on non-headline text (e.g., section tags or sub-headlines) we manually analyzed a random sample of 300 variants in our dataset, ranging from one to 48 words long. We found that most headlines fell between three and 30 words and excluded any test with a variant outside that range. This filtering pipeline retained 140,918 A/B headline tests and 334,976 headline variants across 293 domains.

[3]Chartbeat's testing system distinguishes between hard convergence–in which the system is 95% confident that one headline is more successful–and soft convergence. In the latter case, the system selects the variant which it is confident no other headline beats by more than 25%. Because of this relaxed criterion for selecting a winner, soft-converged tests convey a less certain and clear-cut signal of performance for predictive modeling and are therefore excluded.

### 4.3.2 Deriving Performance Metrics

To measure headline performance, we use a normalized version of click through rate, which we call lift. To calculate lift, we first computed the raw click through rate (CTR) for each variant by dividing its total clicks by its impressions. Then, for each headline test, we took the mean CTR across variants and divided each variant's CTR by that average. The resulting metric represents a variant's lift relative to the test average. Normalization is necessary because headline tests occurred at different times, in different places on the homepage, and across different domains, making direct comparisons of raw CTRs across tests uninformative.

To illustrate, consider a headline test with three variants whose raw CTRs are 0.02, 0.06, and 0.04 clicks per impression. To calculate lifts, we divide each variant CTR by the test's mean CTR value (0.04), giving lifts of 0.5, 1.5, and 1.0, respectively.

### 4.3.3 Descriptive Analysis

The distribution of test counts across domains is heavily skewed. Most domains conduct a small number of tests (median=19), while a couple outliers conduct tens of thousands. This disparity results from variance in content volume and resources available for testing (Hagar & Diakopoulos, 2019), as well as differences in when domains first began testing.

Table 4.1 shows the distribution of the number of headline variants considered in each test. Most tests contain two variants, and very few have more than six.

Figure 4.1 shows the distribution of lift for winning variants. The median lift for winners was 1.23, indicating that the median winning variant garnered a 23% higher CTR in comparison to the average CTR of the test. Some variants perform far better, with 0.5% of winning headlines showing a lift greater than 2.

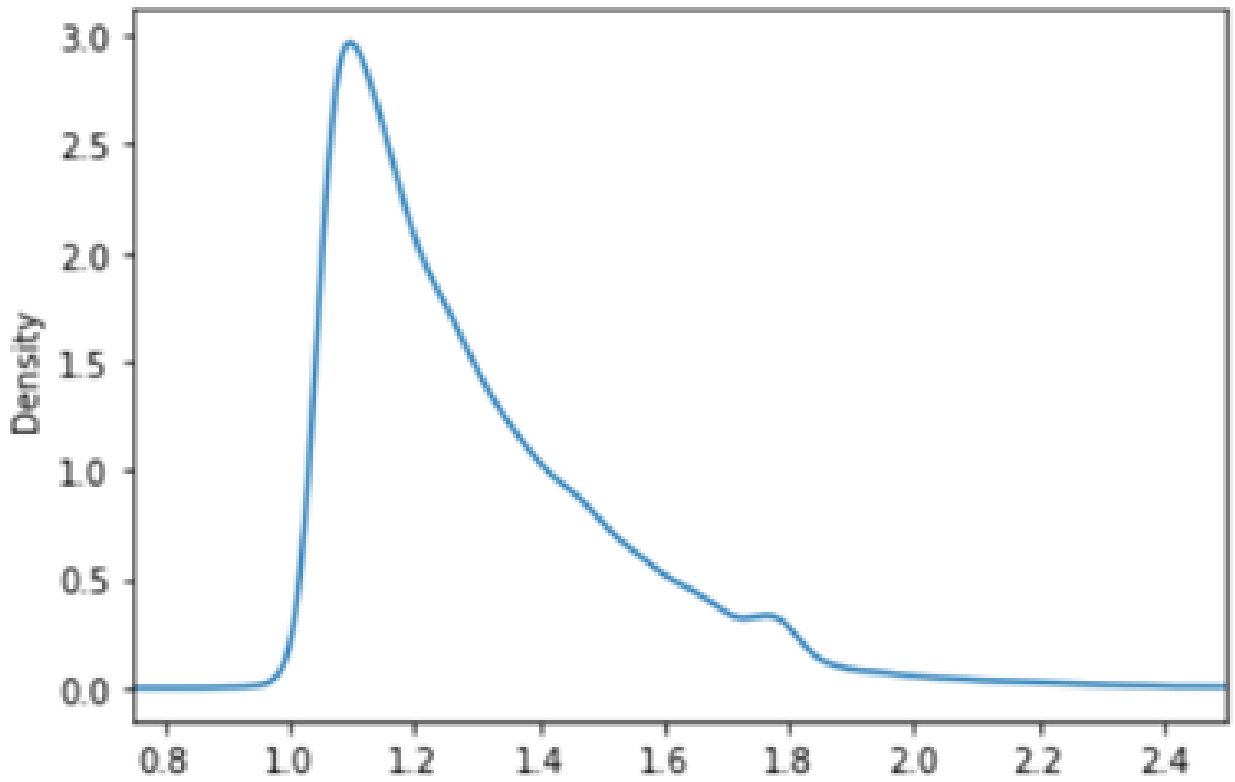



Figure 4.1: Distribution of lift for winning headlines indicating skew with concentration just above 1, a median lift of 1.23, and a long tail.

| Number of variants | Test count | % of total |
|---|---|---|
| 2 | 103,659 | 73.6% |
| 3 | 25,578 | 18.2% |
| 4 | 8,714 | 6.2% |
| 5 | 2,122 | 1.5% |
| 6+ | 845 | 0.6% |

Table 4.1: Count and percentage of tests by number of variants tested. 2-variant tests are most common, and almost all tests have fewer than six variants.

## 4.4 Methods

Our analysis uses predictive modeling to understand how headline writing impacts performance. Our processing pipeline operationalizes key features, trains models, and interprets their predictions. To better contextualize our model's performance, we also empirically estimated the upper limit of predictability within our sample.

### 4.4.1 Feature Engineering

To capture nuanced aspects of headline linguistics and semantics, we leveraged textual features motivated by prior research and theory across four categories: linguistic construction, news values, individual tokens, and semantic embeddings. We also incorporated contextual features about each headline test. Table 4.2 contains the full list of individual features and the source of each computed operationalization.

| Feature | Source | Category | # Dimensions |
|---|---|---|---|
| Part of speech | spaCy part of speech labels | Linguistics | 50 |
| Named entity type | spaCy entity labels | Linguistics | 18 |
| Word count | spaCy token count | Linguistics | 1 |
| Character count | String length | Linguistics | 1 |
| Mean word length | Character count / token count | Linguistics | 1 |
| Fraction of stop words | spaCy stop word labels | Linguistics | 1 |
| Reading level | Flesch reading ease score | Linguistics | 1 |
| Contains question mark | Pattern matching | Linguistics | 1 |
| Contains name | spaCy entity labels | News values | 1 |
| Contains A/V term | Pattern matching: custom dictionaries | News values | 1 |
| Contains location term | spaCy entity labels | News values | 1 |
| Contains magnitude term | spaCy entity labels | News values | 1 |
| Contains shareability term | Pattern matching: custom shareability dictionary | News values | 1 |
| Surprise | Empath dictionary cosine similarity | News values | 1 |
| Conflict | Empath dictionary cosine similarity | News values | 1 |
| Sentiment | VADER dictionary | News values | 1 |
| Lemmas | spaCy tokenization | Tokens | 825 |
| Semantic embeddings | spaCy large model pre-trained embeddings | Semantic embeddings | 300 |
| Datetime | Test metadata | Context | 5 |
| Domain | Test metadata | Context | 1 |

Table 4.2: Features and their sources, categories, and number of dimensions.

### 4.4.2 Linguistics

Linguistic features convey the syntactical components of headlines, including parts of speech and named entities. We also calculated the fraction of a headline made up by stop words (e.g., "the", "of", etc.), whether a headline contains a question mark, and each headline's reading level and length. These features are the basis for several prior empirical studies of headline writing styles (di Buono et al., 2017; Dor, 2003; Kuiken et al., 2017).

To identify part of speech types, named entity types (e.g., people and places), and stop words within headlines, we used spaCy, a state-of-the-art software package for Natural Language Processing (NLP) (Honnibal et al., 2020).[4] We indicate whether each part of speech or entity is present in a headline with a binary variable.

To determine a headline's reading level, we calculated its Flesch reading ease score (England et al., 1953). Our remaining features—word count, character count, mean word length, and the presence of a question mark—used simple tallies or character searches.

### 4.4.3 News Values

News values capture theoretical dimensions of how journalists determine and communicate an article's newsworthiness (Harcup & O'Neill, 2017). While there is not a universal set of news values across the literature, several qualities (e.g., surprise, conflict, proximity, sentiment) are common (Harcup & O'Neill, 2017; Karnowski et al., 2020; Kessler & Engelmann, 2019; Parks, 2019). Many of these concepts also appear in prior empirical evaluations of headline performance (Belyaeva et al., 2018; di Buono et al., 2017). Several of our news value operationalizations rely on dictionary approaches. To combat known issues with off-the-shelf language analysis dictionaries (Boukes et al., 2020), we select or create dictionaries specific to particular news values wherever

[4] https://spacy.io/api/annotation

possible.

Studies of news values often involve qualitative and contextual examination of headlines and articles (Harcup & O'Neill, 2017). Not all news values transfer well to a quantitative approach because they may require more nuanced examination (e.g., entertainment or topic familiarity—see Trilling et al., 2017). Some concern other article elements—such as body text or accompanying visuals—placing them outside the scope of our inquiry. In our analysis, we selected a subset of news values that could be operationalized from headline text: sentiment, reference to the power elite, magnitude, proximity, surprise, conflict, audio/visual signifiers, and shareability.

We calculated headline sentiment using a crowdsourced lexicon of sentiment intensity and valence to score texts with a continuous negative-to-positive measure (Hutto & Gilbert, 2014). The approach is designed for short texts and performs as well as or better than comparable lexicons on benchmark evaluations.

To measure references to power elites, we determined the presence of person entities, as labeled by spaCy. Some prior approaches assess the prominence of an identified individual, by measuring traffic to their Wikipedia page, for example (Arapakis et al., 2017). But because our sample contains a diverse range of outlets, we cannot rely on a single measure of prominence to assess a name's newsworthiness. Instead, we assert that any name included in a news headline carries weight for its intended audience based on the journalist's editorial judgement of importance and familiarity of the name to their audience. We also use spaCy to evaluate magnitude, which captures the scope and scale of a story. We identify this from the presence of comparative/superlative adjectives and adverbs, as well as numerical entities (e.g., percentages, ordinal numbers, and counts).

To get at the idea of proximity, we identified headlines with a location from the presence of spaCy's geographic entity labels. Other measures could record the distance between a place and a news organization, to quantify geographic proximity. However, many dimensions of prox-

imal salience (e.g., culture or topical interests) are not captured by distance (Hagar et al., 2020; McCombs & Winter, 1981). Even in the strictest geographic sense, physical proximity to an (inter)national news organization tells us little about a location's relevance to news audiences. As such, we eschew geographic proximity in favor of treating any named location as salient to a headline's intended audience. The presence of a location in a headline allows readers to make their own assessment of proximity, which may influence behaviors in a way that leads to patterns we can infer from the data.

Surprise and conflict were both calculated from dictionary expansion, a widely used approach to making lexicons more comprehensive (Gentile et al., 2019). We started with a list of synonyms for "surprise" and "conflict", drawn from Merriam-Webster. We then used Empath, a neural network-based lexical tool, to identify larger groups of related words based on these synonyms (Fast et al., 2016). Finally, we calculated headline-level scores for both surprise and conflict based on these expanded dictionaries. To do so, we relied on semantic embeddings, numerical vector representations of words described in more detail below. We use the embeddings from spaCy for each token in the headline and in the dictionary. We then measured the pairwise cosine similarity between every headline token embedding and dictionary token embedding, taking the maximum value of those similarities as the score. This value conveys the extent to which a headline aligns with terms that express surprise or conflict, and it helps to mitigate sparsity issues that might arise from attempting to directly match words in the dictionaries.

Audio/visual signifiers were drawn from the "perception" (273 words) and "see" (72 words) LIWC dictionary categories (Tausczik & Pennebaker, 2010). We also augmented these categories with a manually-curated list of A/V terms. For shareability, we created a binary indicator for whether the headline contained any matches to a series of phrases identified by industry research on social media shareability (Rayson, 2017). Full word lists for our surprise, conflict, A/V, and

shareability dictionaries can be found in Appendix C.

### 4.4.4 Tokens

Tokens refer to the individual words that appear within headlines. They allowed us to make finer-grained distinctions among categories—not just whether a headline has a name, for example, but which name. While this level of detail is often difficult to generalize, past research provides support for examining tokens when predicting headline performance (Kim et al., 2016).

We extracted the set of all lemmas from the headlines in our sample using spaCy. Whereas tokens may differ in conjugation or declension (e.g., "run" versus "running"), lemmatization normalizes tokens to their root form. We selected lemmas that were used 100 times or more, and that had a significant ($p < 0.05$) Pearson correlation to lift. For each of the remaining 825 lemmas, we created a binary variable indicating whether it appeared in each headline.

### 4.4.5 Semantic Embeddings

Word embeddings encode and make comparable the semantics of a text by representing word contexts as dense numerical vectors (Lau & Baldwin, 2016). As the product of deep learning models, the individual dimensions of these embedding vectors do not carry inherent conceptual meaning (Shin et al., 2018). However, word embeddings have proven valuable in headline performance prediction (Lamprinidis et al., 2018).

We used the built in pre-trained semantic embeddings from spaCy's large English model (version 2.2.5). These embeddings contain 300 dimensions and were trained on English language text from the OntoNotes 5 and GloVe Common Crawl corpora.[5] For each headline, we computed the average embedding vector across all tokens.

[5]https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-2.2.5

### 4.4.6 Context

Because headline tests occur on dynamic homepages across websites, we also represented broader contextual features that may be relevant to headline performance. We included the (ordinal-encoded) domain as well as the date and time (as year, month, day, day of week, and hour) as additional features.

### 4.4.7 Modeling

To assess our selected features' ability to predict headline performance, we trained a random forest regressor on a random sample of 75% of our filtered sample, then evaluated it on the remaining 25% using scikit-learn (Pedregosa et al., 2011). Random forests allow us to model potentially complex interactions between features, including non-linear relationships (Hastie et al., 2009). They also effectively incorporate the mix of continuous, binary, and categorical variables that we utilize, without a need for feature normalization. Finally, random forest models are highly interpretable, providing straightforward measures for the relative importance of each feature in prediction outcomes (Breiman, 2001). We compared both standard random forests and gradient boosted trees, and found slightly better performance with the former, which we report here.

We trained a regression model to predict lift at the headline level, then transformed those predictions into test-level ranks (i.e., ranking all variants in a test by predicted performance). We then measured how often the model correctly picked the test winner (precision@1). This approach is analogous to widely used learning-to-rank frameworks (Tatar et al., 2014).

To optimize our model, we ran a grid search over relevant parameters. We varied the number of estimators (50, 100, and 200), minimum samples required to split a node (2, 100, 200, and 400), minimum samples required for a leaf (100, 200, and 400), and the maximum features considered by the model when splitting (square root of total features versus log base 2 of total features). Our

optimal model contained 200 estimators, with split and leaf minimums of 100, and considered log base 2 number of features when splitting. We also ran 3-fold recursive feature elimination with cross-validation to refine the model's features. Recursive feature elimination removes features one at a time, then evaluates the model's performance on a held-out sample without each feature. It then evaluates each feature's utility to the model (See Table 4.3 for information on which features were retained in the final model).

### 4.4.8 Feature Interpretation

We calculated permutation importance to rank the relative contribution of each feature to overall model performance. This approach generates comparable importance metrics that do not depend on the scale or variance of features (Breiman, 2001). To examine the relationship between features and performance directly, we also calculated the Spearman correlation between each feature and lift.

### 4.4.9 Estimated Prediction Ceiling

Content-based performance prediction is a known challenge, because of inconsistencies in audience attention (Arapakis et al., 2017). This inconsistency is driven by factors that cannot be observed through content alone, such as social influence within groups or selective attention heuristics (Salganik et al., 2006; Zillmann et al., 2004). Because of these difficulties, it is unreasonable to assume that our model's performance would reflect a hard upper bound on predictability. To estimate a more reasonable prediction ceiling, we instead examined the consistency of 4,698 repeated tests within our sample. In these cases, multiple tests were run with the exact same set of headline variants, and at least one test reached hard convergence. Repeat tests may occur because of a user's desire to validate test results. In some cases, tests occurred across multiple domains

within the same organization (e.g., networks of local news sites). Because these tests varied along unobserved dimensions, we used them to gauge the impact of outside factors on our model's performance. We define the replication rate as the number of repeat test cases that always result in the same winner divided by the total number of repeat tests. The replication rate acts as a rough upper bound for the predictability of headline tests given variance in non-content-related factors. This straightforward calculation is not dependent on our feature engineering or modeling steps, allowing us to establish an estimated prediction ceiling independent of our model's performance.

## 4.5 Results

### 4.5.1 Content-Based Predictability

We addressed RQ1 (To what extent can a headline's written content predict its performance?) in two parts. First, we evaluated our model's ability to predict headline test outcomes. We measured our model performance using the precision@1 score, which indicates how often the model correctly identifies a test's winning variant. We focus on tests with 2–4 variants in our reporting, since they make up 98% of our sample. Our model's overall precision@1 score was 0.566. That score ranges from 0.4 (for 4-variant tests) to 0.47 (for 3-variant tests) to 0.61 (for 2-variant tests). In all cases, our model outperformed the test-level random baseline—calculated as 1 divided by the number of variants in a test—by at least 0.11. This performance suggests that it's possible to get some predictive power out of content-based features in this context. Headline writing style at some level does matter and can make the difference between a winning and losing headline. But predictability based on content alone also has clear limits.

We then examined these limits more closely by calculating a rough upper bound to predictability based on content alone. As described in our Methods section ("Estimated Prediction Ceiling"), we observed the outcomes of repeated tests and computed their replication rate. Values range

from 56.4% (for 4-variant tests) to 59.0% (for 3-variant tests) to 74.0% (for 2-variant tests). Figure 4.2 shows these replication rates (black dashes) against our model's performance. To further validate these estimates, we calculated our model's precision@1 score for only repeat tests. Our model performed within or slightly above the confidence intervals suggested by these replication measurements. The model can exceed this content ceiling because of the contextual information provided by the domain feature, allowing it to account for site-level differences.

These results help establish an estimate of the predictive power of a headline's textual features. While our model achieved modest performance, A/B test outcomes are clearly influenced by factors outside of how a headline is written. Even tests with identical variants change winners anywhere from 26% to 43.6% of the time, pointing to the important role of audience behaviors and preferences as well as other contextual factors in determining test outcomes.

### 4.5.2 The Impact of Textual Features

We next examined this predictive model in more detail, scrutinizing the impact of individual textual features in response to RQ2 (What is the relative importance of various content-based features to headline performance?). We found only marginal importance for any content features. Out of 1,212 features, recursive feature elimination retained 136 (11.2%). By the nature of our chosen model, these results reflect not just the effect of features in isolation, but also their impact relative to every other feature. Valuable features in this context provide information not already captured by other features, allowing us to compare utility to the model across feature categories (Table 4.3). The linguistic and lemma features fail to provide the model with much useful information beyond what other features offer. In aggregate, the linguistic, news value, and lemma features had comparable value to the model. By far, the most informative category seems to be semantic embeddings. Table 4.4 contains the Spearman correlation between each (non-embedding) feature

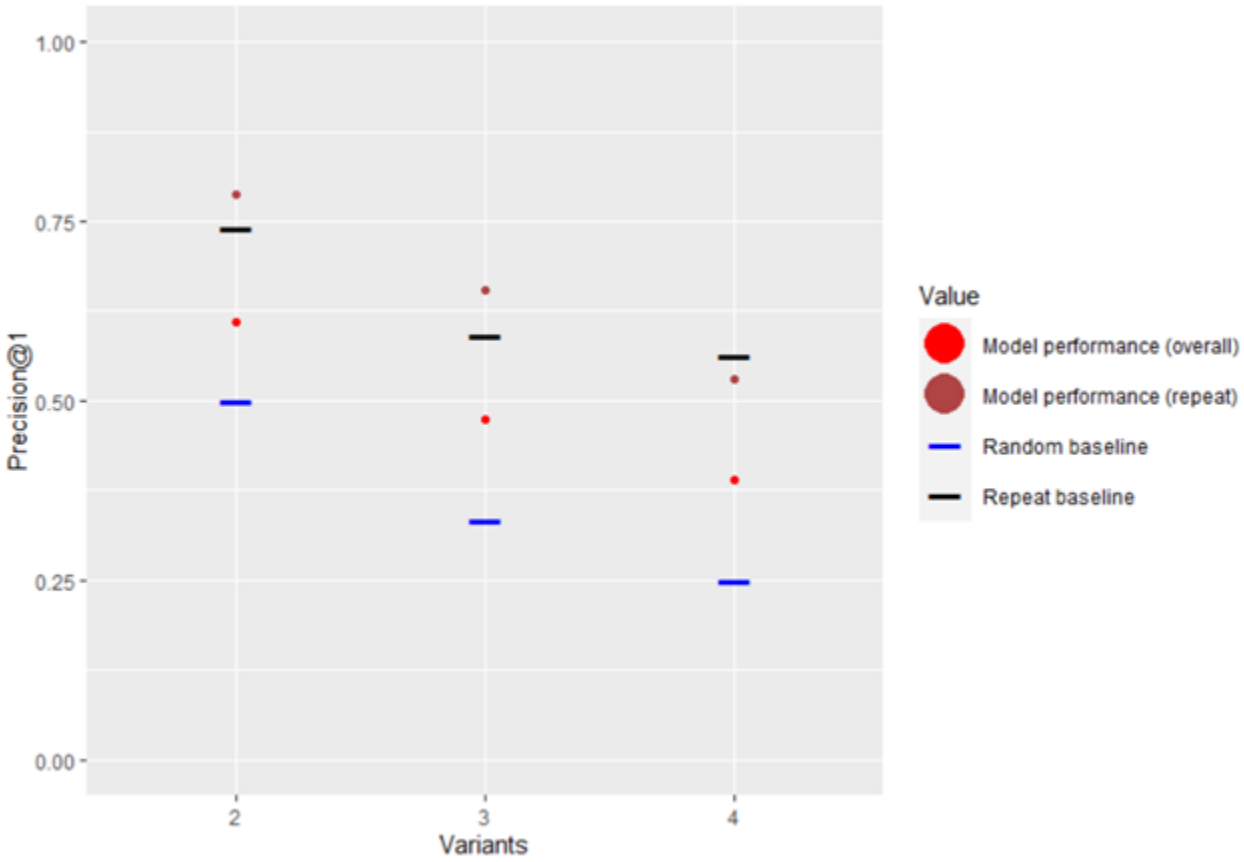



Figure 4.2: Model performance, compared to random baseline performance and our empirically estimated prediction ceiling. Our overall model performs about halfway between the baselines, while the repeat-only model meets or exceeds the content-only ceiling estimate.

and lift. All correlations are statistically significant ($p < 0.01$) except for that of the conflict score ($p = 0.13$). To further unpack the top-level statistics, we next detail the implications of the features chosen within each of our four categories.

| Feature type | Number retained (% of that type) | Total permutation importance | Features retained |
|---|---|---|---|
| Linguistic | 3 (4%) | 0.009 | Average word length, Number of characters, Fraction stop words |
| News values | 3 (38%) | 0.010 | Surprise, Conflict, Sentiment |
| Tokens | 2 (0.2%) | 0.011 | “Here”, “This” |
| Semantic embeddings | 127 (42%) | 0.338 | (see text) |
| Context | 1 (17%) | 0.001 | Domain |

Table 4.3: Number of features retained by category, as well as aggregate feature importance, and some of the specific features retained.

| Feature | Spearman correlation | Permutation importance |
|---|---|---|
| Average word length | -0.026* | 0.004 |
| Number of characters | -0.020* | 0.003 |
| Fraction of stop words | 0.034* | 0.003 |
| Surprise | 0.042* | 0.005 |
| Conflict | -0.003 | 0.004 |
| Sentiment | -0.028* | 0.001 |
| ”Here” (lemma) | 0.045* | 0.004 |
| ”This” (lemma) | 0.052* | 0.006 |
| Domain | - | 0.001 |

Table 4.4: Feature (Spearman) correlations with lift, as well as permutation importances. All correlations are statistically significant ($*p < 0.01$), except for the conflict score.

### 4.5.3 Linguistics

Three linguistics features appear in the final model. Contrary to past work, we find that parts of speech and named entities do not impact headline performance in a way that our model can distinguish (Kim et al., 2016; Kuiken et al., 2017). The remaining features' correlations suggest that simpler, less information-dense headlines perform better. Headlines with more stop words correlate with higher performance. As explored in Blom and Hansen (2015), clickbait acts as a forward reference. More stop words mean less substantive information, suggesting that the key content of the article may not be reflected in the headline. The importance of short headlines and shorter words also reflects the findings of Dor (2003).

### 4.5.4 News Values

The news values features selected by the model—surprise, conflict, and sentiment—focus on headlines' affect. We find a positive association between negative headlines and better performance. These results are in line with evidence of a tendency for people to react more strongly to negative

than positive news (Soroka et al., 2019). While other studies have found that news shareworthiness is predicted more by headline positivity (Berger & Milkman, 2012; Trilling et al., 2017), users may have a different motivation for reading than for sharing an article. Conflict scores have a non-significant negative correlation with performance, while surprise scores have a more substantial positive correlation. The lack of a performance boost from conflict-heavy headlines moderately aligns with past work. Trilling et al. (2017) find a statistically significant but minor increase in sharing for conflict-heavy headlines, while Valenzuela et al. (2017) find that a conflict framing reduces the probability that an article will be shared. The news value of "surprise" largely relates to an article's unexpectedness or contrast (Harcup & O'Neill, 2017; Kessler & Engelmann, 2019). These aspects of surprise are valuable in predicting test outcomes and are correlated with increased headline performance.

### 4.5.5 Lemmas

Only two words provide the model with meaningful predictive information: "here" and "this". Both lemmas are integral to common headline formulations (e.g., "here's why", "this is how") that are often identified as clickbait (Rayson, 2017). They also fulfill the forward-referencing role of clickbait by guiding the reader to a promised piece of information contained either later in the headline or in the text of the article (Blom & Hansen, 2015).

### 4.5.6 Context

As explored in Hagar and Diakopoulos (2019), organizations' testing strategies depend on their priorities, trust in the results of tests, and technical aptitude. Our domain variable implicitly captures these organizational differences. It also acts as a proxy for other variables, such as the size of the outlet, if the domain predominantly covers a particular topic, and facets of behavior that may be

specific to its audience. While the domain feature is less important than most others (permutation importance=0.001), its variation adds some site-level nuance to the model's predictions.

### 4.5.7 Semantic Embeddings

Facilitating clear interpretation and additional theorizing with semantic embeddings is an active area of research. Some extant approaches work to help understand entire embeddings in a relative sense (Kenter & de Rijke, 2015; Li et al., 2010), or transform pre-trained embeddings (Panigrahi et al., 2019). However, we are unaware of methods that elucidate the semantics captured by individual dimensions of pre-trained embedding vectors.

Instead of detailing the individual dimensions of our embedding features and their potential interpretations, we emphasize the utility that the embeddings as a whole provide in this prediction task. They comprise 93% of the features retained by the model (127 of 136), more than any other feature category. They also have the highest permutation importance of any category by far, at 0.338. Their prevalence relative to other features may be a result of the level of granularity they encode. Linguistic categories, for example, tend to encode a relatively broad level of information—labeling a token as a proper noun provides some information about its contents, but elides many details about linguistic context. Indeed, Dor (2003) makes the distinction between including names and concepts with high news value (which help headline performance) and those that have low news value (which hurt headline performance). In contrast, embeddings can capture nuances related to word context and semantics, which appear to carry the lion's share of value for predicting headline performance.

## 4.6 Discussion

This work explores the intersection of headline writing and audience attention. By examining the predictability of A/B headline tests, we develop ecologically valid insights into how writing strategies impact readers' propensity to click on articles.

We find that the predictability of headline performance based on content alone is limited (RQ1). Even when tests contain identical headlines, their outcomes often vary—up to 26% of the time for two-variant repeat tests. Our model draws some predictive power from textual features, but its performance is still hampered by our inability to measure a headline's context. Headline writing matters, but only to some extent. As noted in past work, content-based prediction proves challenging (Arapakis et al., 2017). There is a rich area of literature demonstrating the non-content factors that impact news reader attention, encompassing theoretical frameworks such as selective exposure, partisan preferences, and social influence (Fischer et al., 2005; Iyengar & Hahn, 2009; Lerman & Hogg, 2010; Messing & Westwood, 2014; Zillmann et al., 2004). Journalists' behaviors can also impact this process, by dictating the type, frequency, and quality of tests they run (Hagar & Diakopoulos, 2019).

Our results suggest these external factors play into news audience decision making, even at the micro level of evaluating headlines. They also demonstrate the dynamic nature of news audience engagement. Just as news stories shift in salience depending on the issues and events that are prominent at a given point in time, the writing strategies journalists employ to attract audience attention must depend on context (Waldherr, 2014). Applying static writing strategies to the fluid nature of online publishing (e.g. across time, duration, position) disregards that situational nature. At least some of a reader's decision to click on a news story occurs outside of the moment of exposure to a headline, requiring a contextual understanding of its presentation. Furthermore, a reader's

decision *in the moment* is shaped by a barrage of signals that interact in unclear ways, eluding efforts to make any forward-looking prediction about individual behavior (Gal & Simonson, 2021). To better understand and predict the relationship between news exposure and engagement, we need to incorporate a story's broader circumstances more explicitly into the study of its reception.

Most individual textual features do not substantially impact our model's predictive power (RQ2), though in aggregate the embedding features, which capture high-dimensional linguistic semantics and context, carry the most predictive power. Many features' correlations with headline performance are directionally consistent with past work, such as sentiment (Soroka et al., 2019), conflict (Trilling et al., 2017), and surprise (Kessler & Engelmann, 2019), but they are weak across the board. Our examination of these features stems from work which theorizes that linguistic and semantic formulations directly impact news engagement (e.g., Kim et al., 2016; Kuiken et al., 2017). However, given their limited impact, our results caution against over-emphasizing written headline composition when considering the complex factors influencing news attention decisions. While general trends might appear across a large sample for certain features, they do not provide hard and fast rules for improving performance in specific cases. Past work has already established that click behavior is a weak proxy for audience interest in a headline. Readers may click for reasons not related to an article's presentation, and not choosing to click does not equate to a lack of interest (Kormelink & Meijer, 2018). Our results reinforce this idea in an A/B testing framework, suggesting a need for more nuanced approaches to testing.

Many newsrooms treat testing as an objective source of optimization data, with a clear relationship between the writing strategies they test and readers' click behavior (Hagar & Diakopoulos, 2019). This perception creates an opportunity for ineffective headline writing approaches to gain prevalence, shaping news presentation through a misinterpretation of audience preferences. Similar to the process of writing to an imagined audience, journalists craft their headlines based on an

imperfect approximation of reader preferences when relying on A/B test results (Coddington et al., 2021). Test results (and behavioral analytics more broadly) record data at scale but fail to capture dimensions of audience engagement that are not easily quantifiable (Steensen et al., 2020). These divergent shortcomings complicate our picture of journalistic decision making. Prior work highlights the tension between journalists' professional priorities and the demands of audience metrics (C. Anderson, 2011). Using the framework of the imagined audience, future research should more broadly consider the constellation of audience feedback, as well as the potential shortcomings of its collection or presentation, when evaluating journalists' decisions.

For practitioners, our results stress the importance of adopting A/B testing in newsrooms. In our sample, the median test produces a 23% lift over the average variant click-through rate (even when incorporating less clear-cut soft-converged tests, the median test still generates a 19% lift). Without generalizable best practices for headline writing, continuous testing is the most effective way to achieve this lift because it optimizes the text in relation to specific (potentially unknown) audience and contextual factors. However, newsrooms and testing tool providers must also better communicate the statistical uncertainty of test outcomes and emphasize their context specificity when considering what generalizable lessons can be gleaned by running the tests.

In demonstrating the limitations of content-based prediction, this research suggests a few key areas for future work. First, advances in computational linguistics may allow for more sophisticated encoding of news values. Proximity and power elite could be measured using outside resources (e.g., Arapakis et al., 2017) or models trained via crowdsourced data. As additional signals of audiovisual elements, the images accompanying stories and their contents could be incorporated. Finally, analogous to the shareability measure used in Szymanski et al. (2016), fine-tuned language models could provide more sensitive substitutes for the dictionary approaches used to measure conflict and surprise.

Second, more advanced encoding and modeling approaches could improve predictive performance. Given the relative importance of semantic embeddings to our model, more sophisticated embedding approaches (e.g., sentence embeddings generated by a state-of-the-art model like BERT— Reimers & Gurevych, 2019a) might provide valuable information about a headline's composition. As new methods arise to interpret these semantic embeddings, we may be able to extract more actionable recommendations for practitioners from them (Panigrahi et al., 2019). Given neural networks' dominance in natural language processing tasks, they may prove more effective in predicting headline performance (Conneau et al., 2017).

A final area for future work is measuring audience characteristics and behaviors (at the individual level or perhaps clustered into groups—see Makhortykh et al., 2021) and studying how those characteristics interact with the outcomes of A/B headline tests. By examining longitudinal preferences of users through their reading histories and typical consumption patterns, future research might evaluate the consistency of their responses over time to textual elements.

### 4.6.1 Limitations

This work is also subject to several important limitations that may affect the applicability of our results. First, we only use one class of predictive models, on one subset of data. Other modeling approaches, such as alternative families of regression models or neural networks, may offer increased predictive performance. Our modeling task also only considers hard-converged headline tests. Since soft-converged tests convey a noisier performance signal, our model's performance would likely decline in real-world settings.

Second, we only consider one aspect of an article's presentation. Headlines on a news site homepage are a prominent driver of audience attention, but they are far from the only one. An article might have several distinct headlines—on the home page, the article itself, and social media,

for example. While some features we identify agree directionally with work on social media headlines, it is possible that the magnitude of their effectiveness varies depending on the source of an article's readership. In future work, cross-source comparison could help quantify the extent to which writing strategy effectiveness varies across audiences.

Finally, our feature engineering approaches introduce several limitations. Because of the scale of our sample, we may overlook words or phrases that are effective for particular audiences, but that our model would not register because of their global sparsity. In addition, several of our features are derived from dictionary-based approaches and may therefore suffer from sparsity. So-called "off-the-shelf" dictionary approaches have well-documented limitations (Boukes et al., 2020; Chan et al., 2021). While we attempt to address them with task-specific dictionary selection, data augmentation, and embeddings to reduce sparsity, our A/V operationalization may suffer from the limitations of the LIWC dictionary. Our measurements of power elites and proximity also rely on straightforward identification of named entities, potentially overlooking distinctions within the classes of people and places identified. Finally, because we focus on news values that can be measured quantitatively, we exclude a handful (e.g., entertainment/drama and relevance) that may influence headline performance (Harcup & O'Neill, 2017).

### 4.6.2 Conclusions

This study presents a large-scale, ecologically valid study of A/B headline tests, challenging the link between headline writing and performance. While practitioners benefit from ongoing A/B testing, our results suggest that they will struggle to obtain generalizable best practices from test results. News audiences are dynamic, and capturing their attention requires more than a staid approach to headline writing. Headline testing, and audience metrics more generally, are only one channel of reader feedback, one that needs proper contextualization and caveats. Equating tracking

audience behavior with knowing the audience encourages overreliance on incomplete data, driving flawed approaches to news story presentation.

# CHAPTER 5
# ALGORITHMIC INDIFFERENCE: THE DEARTH OF NEWS RECOMMENDATIONS ON TIKTOK

## 5.1 Introduction

Chapters 2-4 focus on a set of interactions that center on various components of the news system. In doing so, they demonstrate the extent to which interconnectedness fundamentally drives many key news processes, across production, distribution, and consumption. That interconnectedness is not limited to particular processes, though, and these studies gesture toward a larger underlying structure across digital news. In this study, we aim to integrate this earlier work into a system-level examination of news, on one platform: TikTok. We incorporate many of the modes of interaction explored in earlier studies—the feedback between audiences and platforms, the datafied feedback provided to newsrooms by algorithms, the alignment (or lack thereof) between journalists' strategic approach to presenting news coverage and audiences' preferences—into one set of interactions driving a particular context of news exposure.

News has long been a staple on social media platforms. Debates have formed around some aspects of news distribution on large platforms—whether they privilege certain outlets or perspectives over others, for example (Trielli & Diakopoulos, 2022), or whether credible sources get proper priority over misinformation and so-called "junk news" (Castaldo et al., 2022). However, the presence and importance of news as a component of social media's content mix has largely been treated as a given (Fletcher et al., 2021; Harder et al., 2017; Vermeer et al., 2020; Wojcieszak, Menchen-Trevino, et al., 2022).

However, news publishers may now be faced with a platform whose logic does not assume the institutional importance of news, in the form of TikTok. TikTok is one of the fastest-growing social platforms in the world, projected to reach over 840 million users in 2023 ("TikTok Users Worldwide (2020-2025)", 2023). Large news publishers, including Sky News and the Washington Post, have begun adopting the platform into their broader social strategies (N. Newman, 2022). But the experience of many publishers has been lackluster, leading some to abandon the platform entirely (Klug, 2020).

TikTok presents unique challenges for news, relative to other platforms. Its primary modality is short-form video, demanding news production outside the traditional workflow of written articles (N. Newman, 2022). It has an explicit focus on entertainment, particularly the music industry (Whateley, 2022). Its default user experience, the For You Page, is driven by an algorithmic recommendation system which generates a personalized feed of content based on user interactions (e.g. likes, shares, comments, follows), content-based features (e.g. captions, sounds, hashtags), and device or account details (e.g. language or country)[1].

The last point raises a key question for news on digital platforms: In a heavily algorithmic environment, driven primarily by recommendations designed to maximize user relevance, how will news fare? Abdollahpouri et al. (2021) argues that alternative recommendation schemes are required to ensure that credible news remains discoverable on algorithmic platforms. Thorson (2020) posits that news consumption will become uneven: Users who are interested in news will seek it out and receive positive reinforcement from the algorithm, while everyone else will get exposed only to other kinds of content. These arguments hinge on concerns about *exposure*—who sees what. Exposure to news is often incidental, as people encounter news coverage in the course of other activities (Taneja & Yaeger, 2019). Researchers have demonstrated benefits from incidental

[1]How TikTok recommends videos #ForYou: https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

exposure, in the form of increased awareness of current events (Lee & Kim, 2017). However, incidental exposure is not pure serendipity. Rather, it relies on the presence of functional pathways through which news can reach potential audiences (Thorson, 2020).

In the case of TikTok, those pathways largely flow through the recommendation algorithm. This research examines algorithmic news exposure in this context, attempting to understand the extent to which the platform might show users journalism or commentary around current events. Using web scraping and automated agent-based interaction we collect data from TikTok's account recommendations, For You Page, and trending hashtags. Our findings suggest a dearth of news exposure on TikTok, across three areas. First, TikTok presents new users with a mix of viral content (e.g., food videos, ASMR, makeup tutorials), omitting news almost entirely. Second, trending news topics on TikTok lean heavily toward entertainment (blockbuster films, K-pop groups, popular TV series). Finally, news producers are often not recommended to users by TikTok, and they receive relatively muted engagement. These findings highlight the need for explicit platform interventions to ensure broad access to journalism, such as labeling credible news outlets, elevating relevant coverage on trending topics, or giving journalism prominent placement outside the usual algorithmic flows.

## 5.2 Background

The process by which an individual encounters news on a social platform like TikTok relies on *interaction*. Users interact with the platform by engaging with its content. These interactions provide data to the platform, through which it shapes personalized recommendations. In turn, the platform interacts with the user, largely by means of an algorithm responsible for maximizing the relevance of the content users see. The interplay between user feedback and algorithmic recommendation shapes the user's platform experience.

However, for news consumption, this loop is insufficient. A platform must also contain a supply of credible journalism that it can recommend to users. This requires that news producers expend the resources necessary to publish their coverage to the platform. On many platforms, this is as simple as uploading a hyperlink. On a multimedia, entertainment-focused platform like TikTok, it represents a more substantial investment. That requirement creates another set of consequential interactions. News producers upload coverage to the platform. The platform signals the value of this activity, in the form of data around audience engagement.

This *relational* model idealizes a complex set of interactions that enable platform-mediated news consumption. Each interaction is essential, and influence flows from every actor to every other actor. As a consequence, a breakdown in any of the actions described above could interrupt the flow of news on a platform. The following sections explore user/platform and platform/news producer interactions in more detail, highlighting areas where the empirical reality of TikTok might diverge from such an idealized model.

### 5.2.1 User interest and algorithmic recommendations

News consumption spurs positive civic outcomes, and social media provide a key vector for exposure to news coverage (Shaker, 2014; Vermeer et al., 2020). However, this exposure is unevenly distributed (Thorson, 2020). While some users might actively seek out news, less proactive consumers might miss it entirely.

In some cases, this is a consequence of constrained resources. Social platforms are high-choice media environments (Prior, 2007). Users might like news, but they might have a stronger preference for entertainment content. A recommendation system, providing the most salient content for the user's interests, might fill their constrained time on the platform with only entertainment content (Damstra et al., 2023). Similarly, news producers deploy strategic resources to reach their

most promising audiences. In doing so, they exclude certain groups deemed as less likely news consumers (Thorson, 2020). There is no active decision on the part of the individual user in these cases of *unintentional* news avoidance.

*Intentional* news avoidance, on the other hand, stems from an active desire to not encounter news. News overload can play a role in avoidance, as consumers get overwhelmed by the volume of coverage they encounter (Villi et al., 2022). Similarly, fatigue around certain topics, or in the face of consistent negativity in the news, can drive people to reduce their consumption (Villi et al., 2022). Audiences may come to view the news available to them as unreliable or untrustworthy, or of limited utility. As news becomes less valuable along some dimension, consumers might decide to avoid it altogether (Toff & Kalogeropoulos, 2020).

These attitudes inform the broad types of interactions that a user might have with a platform around news content. They can send a positive signal by interacting with news content or following news accounts. They can seek out other types of content, thereby sending a neutral signal. Or they can send a negative signal, by blocking news accounts or leveraging platform tools to request less of a certain type of content.

To outline the platform's potential responses to these signals, we must also characterize its treatment of news. Many recommendation systems work by ranking items by their potential utility to a user, and then serving the top-ranking items (Barbieri & Manco, 2011). Depending on the system's definition of utility, news may never make an attractive recommendation for anyone but the most interested consumers.

This suggests the need for a balance between responding to user preferences and exposing users to diverse, credible information sources. By providing news recommendations independent of user factors, recommender systems can encourage incidental exposure (Abdollahpouri et al., 2021). This is the process by which users encounter news, even when not actively seeking it out (Lee &

Kim, 2017). When users who are less interested in news encounter it incidentally, they ultimately consume news from a broader range of news sources (Fletcher & Nielsen, 2018). There appears to be a network effect to this exposure, as users who consume news through social platforms have a more politically diverse news repertoire than those who go directly to news websites (Fletcher et al., 2021).

The degree of news exposure on a platform is therefore expected to be a complex interplay incorporating the intents of users and the intents of platforms. It is modulated by the responsiveness of platforms to users' intents, and it is contingent on the availability of news content on the platform. In this work we seek to shed light on this complex system, using observation to assess the precondition of news availability and simulation to assess varying user intents for news exposure together with how the platform responds. At a high level, the following question drives our inquiry:

**RQ1:** To what extent does TikTok expose users to news content?

### 5.2.2 Pathways to user engagement with news content

In many cases, the relationship between news publishers and the large platforms that distribute their coverage has been contentious, as platforms exercise an outsized amount of commercial power (C. Anderson, 2011; Caplan & boyd, 2018; Gillespie, 2010).

The negative impacts of platforms on journalism have, deservedly, received widespread attention from researchers (Latzer & Just, 2020). In the context of news *exposure*, though, this critical focus risks overshadowing the benefits that (some) outlets enjoy from established platforms. To varying degrees, large tech companies have explicitly encouraged the presence of news in their products.

News publishers benefit from affordances that heighten the reach of their output in several ways. First, technology firms often make design decisions on their platforms that give news content favorable presentation. Alphabet presents news coverage in privileged positions at the top of Google search results (Trielli & Diakopoulos, 2019b). Twitter surfaced news articles in its trending topics (Kwak et al., 2010).

Going a step further, some platforms provide dedicated products built around consuming the news. Google and Apple both operate news aggregation services, which provide avenues for publishers to accrue web traffic, convert new subscribers, or arrange paid partnerships. In Apple's case, this surface blends human and algorithmic curation, further highlighting news content that might not otherwise receive prominent placement (Bandy & Diakopoulos, 2020). As part of its paid subscription service, Twitter lets users leverage their social graphs as a news discovery engine, highlighting popular articles among on-platform connections.[2]

Even more direct still are the payments or partnerships that technology firms have offered to news publishers. These fall under three buckets. In some countries, platforms must pay legally-mandated licensing fees when news content appears in, e.g., search results (Kaye, 2022). In other cases, payments emerge from voluntary partnerships, as in the case of BuzzFeed's video contract with Facebook (Jones, 2023). Finally, some firms, most notably Google and Facebook, have operated grants to support local news and other forms of journalism.[3]

These facets all signal the value of news to platforms as a distinct and featured category of content. They all, everything else held constant, provide an advantage to news organizations over other on-platform content creators.

However, social platforms seem to be moving away from this position. According to Pew data, the most popular social media platforms among U.S. teens are YouTube, TikTok, Instagram,

[2]https://help.twitter.com/en/using-twitter/top-articles
[3]https://newsinitiative.withgoogle.com/programs/

and Snapchat, all multimedia-oriented platforms (Vogels et al., 2022). This has consequences for journalism, a form of content that often manifests on the web in text form (albeit often visually illustrated). Multimedia news is better suited to some types of coverage than others, and some types of news producer can adapt to its production process more smoothly (Boczkowski, 2004). Shifting social media modalities therefore also shift the kinds of news present on platforms.

Coupled with this shift, platforms like TikTok flatten news, presenting it alongside every other type of content creation. They deprioritize or completely remove the ability for organizations to link out to their work. Audiences are not given tools to evaluate the credibility of sources, or to differentiate an institutional news publisher from any other account (Bimber & Gil de Zúñiga, 2020). And gone are the financial incentives and preferential placement often given to news publishers. Without this aid, news organizations often struggle, or flat out refuse to engage with the platform (Klug, 2020).

With news exposure left at the mercy of the algorithm, active news engagement is also at risk. Some share of the users who are exposed to news content might go on to follow the account that produced it, or to provide commentary or opinion in a post of their own. In an environment where an algorithm dictates an increasing share of attention flow, these activities also become tied to that algorithm's mechanisms for news exposure. Therefore, it is valuable to understand how users engage with news in the context of a large-scale platform that appear almost wholly indifferent to news coverage as a feature of its offerings:

**RQ2:** In the absence of explicit platform differentiation and prioritization of news content, to what extent do users engage with news producers or news topics?

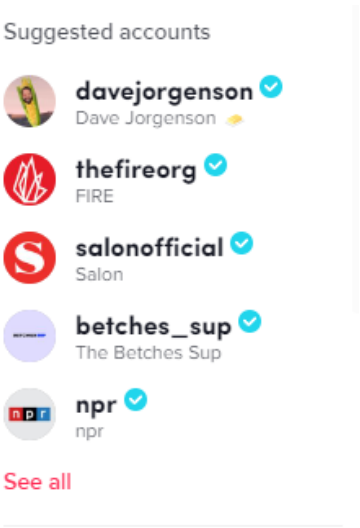
## 5.3 Data

For the purposes of data collection, TikTok differs from some other major social media platforms in key ways. At the time of this study, the platform offers no external-facing API. It is also less directly observable than platforms like Twitter or Reddit. On those platforms, there are retrievable metrics of global or community-wide popularity—posts appear in platform-wide trending lists, or are ranked highest among a set, indicating wide visibility to users. Tiktok, by contrast, primarily organizes its content through the algorithmically curated For You Page. This means that experiences of TikTok content vary widely from user to user, with little global commonality. And rather than emerging through user-defined taxonomy, sub-groups of users are algorithmically defined and dynamic (Gillespie, 2010). This occludes any attempts to directly collect popular content relative to a group of users, as those groups are fuzzy and constantly shifting.

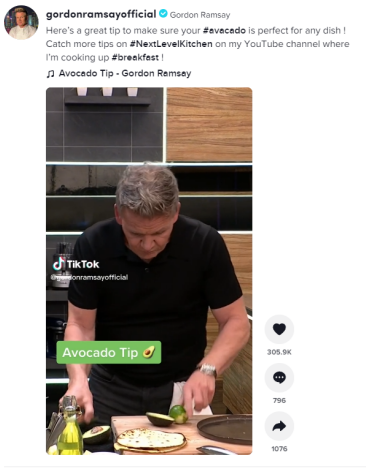
To address these challenges, we attempt to collect data that correspond to several facets of news content on TikTok. First, we simulate new users' exposure to news content (or lack thereof), through account recommendations and the For You Page. Second, we capture the kinds of news content that are generally popular across the platform, by scraping TikTok's list of most popular hashtags. Finally, we collect data about the popularity of established news organization accounts and their videos on TikTok, to examine news producers' experience of their audience on the platform. The following sections describe each of these data collection approaches in detail.

### 5.3.1 User-level Data

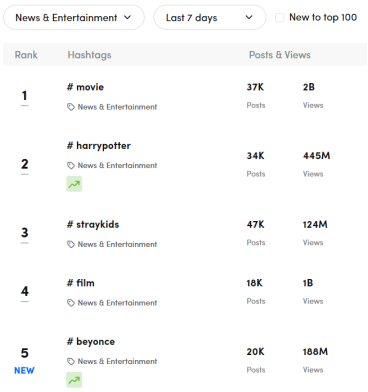
Our goal in collecting user-level data is to understand the extent to which a TikTok user might encounter news-oriented content or accounts. To do so, we collect data from two prominent algorithmic surfaces: recommendations of suggested accounts (Figure 5.1a), and the For You Page




(a) Suggested accounts for @washingtonpost




(b) Example For You Page video




(c) Trending news and entertainment hashtags

Figure 5.1: Examples of platform surfaces used for data collection. Each example is taken from TikTok's desktop website.

(Figure 5.1b).

### *5.3.1 Account Recommendations*

When a user visits the page of a TikTok account (on desktop) or follows that account (on mobile), TikTok lists additional accounts that it has algorithmically determined are related. This flow provides a potential avenue through which users might discover accounts producing news coverage—after following their primary news source, for example, they might get recommended additional news outlets, or accounts whose primary content consists of commentary on the news.

To explore this possibility, we set up an automated pipline to scrape recommended accounts from TikTok's website. On the desktop website, every account page contains a sidebar of "Suggested accounts", algorithmically curated for their similarity to the displayed user. The base probability of a news organization appearing in this sidebar for any given account is quite low—an

initial examination of recommendations from non-news accounts failed to find any. To account for this, we started the pipeline on four known, active news accounts—The Washington Post (@washingtonpost), NBC News (@nbcnews), National Public Radio (@npr), and PBS News (@pbsnews). This seed set provides the recommendation algorithm with a strong signal toward news publishers of various types, hopefully increasing our chances of getting recommended relevant accounts.

Our scraper collected the accounts recommended alongside each in our seed set, then fanned out from that recommended set to a further round of recommendations. We ran our scraper in a logged out, private browser window. We ran the fanout three times, then manually curated the list of resulting accounts to those that produced news. We used two criteria to determine inclusion in this set. Accounts had to be verified on TikTok, to ensure we didn't include imposters or account squatters. They then either had to a) correspond to an established news publisher brand elsewhere on the internet, b) belong to a professional journalist, as identified in their biography, or c) belong to an individual whose primary activity on TikTok was producing news aggregation or commentary, as identified in their biography and/or their three most recent videos. From that curated list, we ran an additional 3 rounds of data collection, repeating the process of account curation and subsequent fanout until saturation (i.e., TikTok's recommendations no longer surfaced new accounts that met our sampling criteria), a total of three times.

### *5.3.1 For You Page*

The main surface through which most users interact with TikTok is the For You Page, an algorithmic feed of short-form video. Users do not need to follow or engage with accounts for them to appear in the For You Page; rather, TikTok's recommendation algorithm attempts to ascertain user content preferences via watch time and other engagement metrics (R. Barry et al., 2021).

As the primary method through which users interact with content on TikTok, the For You

Page provides a key view into how TikTok responds to news (dis)interested users. To explore this interaction, we set up an automated data collection pipeline designed to mimic real-world viewing behavior on TikTok (Diakopoulos et al., 2021). Conceptually, our goal is to emulate the new user experience, one who has not interacted with or followed any accounts, and for whom the algorithm has no prior knowledge. This allows us to hew as closely as possible to realistic interaction between the user and the platform, without potential confounders like prior activity or changes to the recommendation algorithm over time.

To collect these data, we implemented an automated pipeline to interact with TikTok's website, evaluate videos relative to a probabilistic expression of "news interest", and determine whether to watch or skip videos based on that evaluation in order to provide that as a signal to the algorithm.

First, we created an automated agent to handle interactions with TikTok's website. We used Playwright, a software package for testing web apps, to control this agent.[4] This part of our pipeline opens a web browser (Chrome, in a private window with no browsing history or cookies), waits for an account to be logged in (TikTok login must be performed manually, because of CAPTCHA-like verification procedures), then scrolls through the For You Page one video at a time. The agent logs what videos it encounters, as well as metadata (caption, hashtags, likes, and reshares) about those videos where available.

This process sets up an automated agent that deterministically scrolls through every video it encounters. In order to allow agents to express interest in news, and vary their behavior accordingly in terms of watch time (one of the signals TikTok uses to assess user interest), we build a set of processing steps that ingest and analyze videos, then send a signal to watch or skip said videos based on their contents. First, at the start of the agent's session on TikTok, we query the New York Times' archive API[5] for the headlines of all articles published on the current day. This gives us

[4]https://playwright.dev/

[5]https://developer.nytimes.com/docs/archive-product/1/overview

a corpus of representative and timely news coverage text to compare videos against. Then, every time the agent swipes to a new post, we download its video and transcribe the corresponding audio using OpenAI's Whisper model (Radford et al., 2022). Using S-BERT[6], we generate embeddings for the video's transcript, as well as all the text in the New York Times corpus for that day (Reimers & Gurevych, 2019b). Finally, we compare the embedded video transcript against each New York Times headline embedding using cosine similarity. We take the maximum value from that distribution as a measure of how similar a video's content is to the news of the day, as published by an authoritative source. Intuitively, a high score would indicate that the text from the video was semantically similar to at least one of the headlines from the day. From that similarity score, we calculate the probability that an agent will watch the video. Each agent has a news preference value—ranging from 0 to 1—that is meant to express its level of interest in news-focused content. We multiply the video's cosine similarity to the news corpus by the agent's news preference value, producing a watch probability for the current video. This probability then signals the agent to either watch the whole video currently in frame, or to scroll to the next post.[7]

We collect data from agents with news preference values of 0.2, 0.4, 0.6., 0.8, and 1.0. Each agent's session lasts approximately 40 minutes, the length of a typical session on TikTok (Lebow, 2022). Within these parameters, we collected data in two phases. First, we collected data over 12 days (2022/12/6-2022/12/17), registering a new TikTok account for each agent, for a total sample of 60 accounts. Second, to gauge the effect of an active signaled interest in news on the algorithm's recommendation, we separately seeded five accounts by having them follow the five largest news organizations in our recommendation fanout sample by follower count—Complex, NBC News, E! News, Bleacher Report, and NowThis Politics. We assigned these agents the same range of news

[6]Implemented in the sentence-transformers Python package: https://www.sbert.net/

[7]Video processing takes up to 30 seconds, depending on the length of the video. This runs the risk that TikTok's algorithm picks up an erroneous signal of interest in longer videos. However, our pipeline processes video on a separate thread in batches, which introduces some random variation into delay times that is not linked to video length.

preference values and collected an additional day of data (2023/1/20).

### 5.3.2 Platform-level Data: Popular News Hashtags

To the extent possible, it is valuable to supplement our view of the individual user experience with an understanding of how news permeates TikTok as a whole. To capture this aspect, we leverage TikTok's Trend Discovery tool, an interface intended for advertisers but available publicly (Figure 5.1c). This tool displays the top 100 "trending" hashtags, ranked using a combination of post and view counts, over a 7, 30, and 120 day window. It breaks these hashtags down by category, including "News & Entertainment". To examine which news and entertainment hashtags are most popular, we scrape a daily sample of the webpage with their rankings over the course of a week (2023/1/20-2023/1/26). To mirror the constantly shifting nature of news coverage, we use the 7-day window of hashtag popularity.

### 5.3.3 Producer-level Data: News Account Popularity

Finally, we supplement the user and platform views with a direct examination of how well news organizations capture attention on TikTok. To do so, we scrape TikTok's website to collect the follower count, engagement, and viewership metrics of news accounts identified during the recommendation fanout.

These datasets provide a comprehensive view of the relationships that largely determine news exposure on TikTok. They allow us to examine interactions, at the individual level, between users and the recommendation algorithm. They give a view of news producers' experience of those recommendations, in the form of aggregate engagement data. And they provide an overarching view of trends in audience attention, in the form of popular news and entertainment hashtags across the platform.

## 5.4 Methods

Our approach aims to provide a *system-level view* of news on TikTok. Rather than focusing solely on the algorithm, the audience, or the news producers, we attempt to capture interactions among them. This lets us more precisely locate where in these relationships news dynamics might originate, and to emphasize the critical nature of their interactions in shaping the flow of news across the platform.

### 5.4.1 User-Level Data

To answer RQ1, we first examine the individual user experience of algorithmic recommendations on TikTok. We do so by analyzing the extent of news-related content in account recommendations and the For You Page.

For account recommendations, we follow the criteria laid out in the data section to identify news accounts. TikTok does not maintain a comprehensive list of news outlets on its platform (at least publicly), let alone the potential individual accounts that produce news commentary, so there is no way of knowing how comprehensive our approach is. However, our method posits that the recommendation algorithm will group together related accounts, and thus provide a reasonable snapshot of news media presence on the platform.

In the course of collecting For You Page data, our agents log the video files they encounter, the watch probability assigned to each video, and available metadata about the video creator and engagement. We leverage this information in several ways. First, we compare the accounts our agents encounter to those in our list of news accounts obtained through recommendations. This gives us an initial sense of the extent to which authoritative news sources appear in new users' feeds. Second, we compare each video transcript to the New York Times sample, using cosine

similarity. This is meant to capture the extent to which the algorithm picks up on and responds to strong news interest. Finally, to account for the possibility that our account list misses key news producers, we manually tag whether the videos each agent encounters are news related (i.e., whether they summarize, comment on, or repurpose coverage of current events). This gives us a fine-grained view of the extent to which the algorithm provides any opportunity for agents to signal their interest.

### 5.4.2 Platform-Level Data

In collecting the most popular news and entertainment hashtags on TikTok, we are able to measure the expressed topical interests of the platform's users in that domain. Further, we can leverage news text to compare those interests to that of the institutional news media, allowing for an examination of their relative news agendas. These efforts give a broader view of news consumption, providing additional perspective for RQ1.

To accomplish this, we again turn to New York Times headline data, this time over a 7-day rolling window to match our hashtag data. We conduct string searches against headlines in this sample, to identify cases where hashtag language appears in news coverage. We then manually examine those matches to rule out false positives.

As an additional step, we analyze the TikTok trending hashtags to capture the topics that they cover more broadly. We do so by manually tagging the set of unique hashtags, first as news relevant or not, then within a finer-grained topical hierarchy. This process gives us a view, independent of an institutional news comparison, of the types of news most relevant to TikTok users. Since TikTok combines news and entertainment in their hashtag rankings, it can also provide a view of the relative hard/soft news divide on the platform.

### 5.4.3 Producer-Level Data

To answer RQ2, we measure the distribution of account-level engagement (followers, likes, views) and video-level engagement (typical viewership) across the sample, to examine the extent to which attention might concentrate to the most widely-surfaced producers. These comparisons give us insight into two platform dynamics. First, they tell us about the supply of news on TikTok—on a given day, how many news videos are being produced, and from what accounts. This can help elucidate why news content may or may not appear in the algorithmic feed. Second, they help to uncover, from the producer perspective, an aggregate measure of relative success. News producers, institutional organizations in particular, must evaluate their output in terms of the audience it reaches. Profit-seeking organizations do so toward the goal of selling advertising or netting new subscribers. Nonprofit entities do so as a measure of impact and outreach. In both cases, the perceived response of audiences on a platform may factor into a producer's decision on whether or not to continue publishing on said platform.

## 5.5 Results

In all analyses, our results demonstrate an overwhelming disinterest in news coverage on TikTok. While some news organizations actively produce content for the platform, most accrue only a modest audience. It is unclear whether this apathy originates from user disinterest, or from a lack of platform efforts to boost news exposure. The For You Page algorithm surfaces virtually no news content, even when primed with active engagement signals. Very few "hard news" topics appear in the trending news hashtags, while entertainment and pop culture consistently dominate. The picture that emerges is one of a platform indifferent to journalism, in both technical features and engagement data. The following sections delve into each of our analyses in more detail.

### 5.5.1 Deficient news recommendations at the user level

#### *5.5.1 Account recommendations*

Our account recommendation fanout saturated after three iterations, at which point no new news accounts appeared. In total, our sample included 10,361 recommendations, of 2,005 unique accounts. Using the coding schema described above, we identified 18% of those recommendations as news-related accounts. However, after removing duplicated accounts across those recommendations, 6% of unique accounts (120 of 2,005) were news related. This proportion puts news producers, the starting point and expressed goal of our fanout, in a small minority of recommendations. It also represents a relatively impoverished attempt at producing relevant recommendations. To provide a small comparison, we ran a single iteration of a fanout from American football-focused accounts (detroitlions, buffalobills, nfl, bengals, chargers). 88% of those recommendations came from accounts that also focused on American football content.

These results suggest at least two possibilities. First, as indicated in past work (Klug, 2020), there may not be enough active news producers on TikTok to saturate account recommendations. Second, the recommendation algorithm may not recognize news producers as a coherent cluster of accounts, and thus fails to associate them for the purposes of suggesting accounts.

Of news accounts that do appear in recommendations, PBS News appears most often (Table 5.1 contains the most-recommended 20 accounts). Interestingly, the accounts represented cover a wide range of news producers. They encapsulate newspapers (latimes), individual professional journalists (itsrachelscott) and news commentators (briantylercohen), broadcast news (cbseveningnews), and others.

| Account | Recommendations |
|---|---|
| pbsnews | 232 |
| briantylercohen | 171 |
| zerlinashow | 158 |
| npr | 111 |
| itsrachelscott | 78 |
| latimes | 77 |
| verifythis | 66 |
| pbsdigitalstudios | 64 |
| theviewabc | 58 |
| theinfographicsshow | 38 |
| newsweek | 34 |
| theproblem | 34 |
| time | 34 |
| nbcnightlynews | 33 |
| cbseveningnews | 31 |
| ajplus | 29 |
| fullfrontalsamb | 25 |
| ken | 24 |
| cnn | 22 |
| nowthispolitics | 22 |

Table 5.1: The 20 news accounts that appear most often in news account recommendations.

### *5.5.1 For You Page recommendations*

Across the 60 automated agents in our first wave of feed data collection, we gathered a sample of 6,568 videos. Some are duplicated across sessions, as TikTok appears to pull from a common pool of popular videos for new accounts. To assess the news relevance of recommendations, we look for videos that either come from one of the accounts on our news producer list, or the contents of which address current events. None meet the first criterion and only a miniscule fraction—6 videos—meet the second criterion. Furthermore, these appearances are only comprised of two *unique* videos. One, which appears twice, is a post from short-form video news publisher Brut,

about an attempted mass shooting in Buffalo, New York. The other, which appears four times, is a video from Fox Soccer about a stray cat appearing at a press conference. The algorithm's recommendations do appear to increase slightly in relevance for news-interested agents (Spearman's $r(6873) = 0.07$, $p = 1.89 \times 10^{-8}$, between news interest threshold and video/news cosine similarity), but they do not strictly address current events. As far as our methodology can reveal, news does not appear to be a meaningful component of the new user experience on TikTok.

To give a sense of what *is* included in this video sample, we examine the hashtag data available from the video descriptions in our sample.[8] After deduplication and removing records without hashtags, we are left with 4,199 video descriptions. There are 20,911 total hashtags (4,127 unique; 5 on average per video description).

Table 5.2 ranks the most common hashtags across this video set. A couple key themes emerge. First, the most common hashtags overwhelmingly reference the platform itself. Approximately 3,300 instances reference the For You Page in some way. Another 600 reference going viral. These are both widely-acknowledged folk theories from TikTok creators about how the algorithm responds to tagging, which have not been confirmed or recommended by the platform (McLachlan, 2021). Also common are references to TikTok functionality—trends, stitches, answering questions, and duets are all commonly mentioned. These are instances of users classifying their videos not by their contents, but in the vocabulary of the platform's affordances. For those hashtags that do address contents, common categories of viral content dominate. In particular, sensory content appears dominant, as "food", "asmr", and the broader umbrella "satisfying" all make an appearance.

To corroborate these themes, we qualitatively examine a random sample of 100 of the videos. In doing so, we find a similar focus on generic viral content. Food, makeup, comedy, and ASMR

[8] In one case, TikTok placed an agent in what appeared to be a beta test for a new website interface, preventing our scraper from collecting anything but video files.

are common areas of focus. Lacking is any discussion or coverage of news events.

| **Hashtag** | **Frequency** |
|---|---|
| fyp | 1,798 |
| foryou | 727 |
| viral | 524 |
| foryoupage | 427 |
| asmr | 349 |
| fy | 180 |
| answer | 166 |
| funny | 154 |
| food | 143 |
| relax | 130 |
| xyzbca | 129 |
| satisfying | 126 |
| parati | 121 |
| asmrsounds | 119 |
| tiktok | 115 |
| duet | 111 |
| trending | 101 |
| trend | 100 |
| viralvideo | 100 |
| stitch | 94 |

Table 5.2: The most commonly occurring hashtags across recommended videos in our first wave of data collection. Most reference Tiktok affordances, or common categories of viral content.

In examining agents seeded by following news accounts, we do not see more news coverage in the For You Page.Only 3 of 465 recommended videos in this sample meet one of our news criteria. However, videos are generally more *relevant* across the board as measured by the similarity between video transcripts and the NYT headlines we collected. A t-test conducted across both distributions shows a significant increase in video similarity to news in the seeded sample ($t(6873) = 8.06$, $p = 4.52 \times 10^{-15}$, $\bar{sim}_{seeded} = 0.25$, $\bar{sim}_{nonseeded} = 0.23$). Figure 5.2 shows the distribution of videos' similarity to news text, by news relevance threshold, for the seeded and

unseeded agents.

In another qualitative exploration from this sample of videos, we find trends toward content that might *feel* or *appear* like news, but is not. One category encompasses content that draws from the U.S. justice system, but is framed as entertainment (e.g., a video captioned "Scary 911 calls..." and set to ominous music). Another includes wellness or health adjacent content (e.g., "Knowing this about drinking could save your life!"). This sample also includes general commentary from pundits, such as a clip of right-leaning political commentator Ben Shapiro speaking on a college campus. Finally, in one case, fringe misinformation crops up, in the form of a video about how giants were responsible for the construction of ancient monuments in Peru. While the algorithm increases its responsiveness to our priming in terms of our quantitative similarity metric, the actual contents it deems relevant are still at most tenuously connected to coverage of current events.

### 5.5.2 Soft news focus at the platform level

Our total sample over the 7-day period contains 700 observations of 163 hashtags in the News & Entertainment category. Of those hashtags, 48 appear across all 7 days of observations. Table 5.3 shows the top 10 of that subset by their average daily ranking (out of the top 100); the highest-ranked hashtag over this time period is "spirituality".

Already, a clear focus on entertainment across these hashtags emerges, as film (Avatar), television (The Walking Dead), and music (Spotify) dominate the most prevalent hashtags. References to prominent news events are glaringly absent, both in this truncated list and the full hashtag sample. The ongoing war in Ukraine, for example, does not appear in the sample of trending news and entertainment hashtags.

Unsurprisingly then, there is very little overlap with the New York Times data for this time period. This gap is evident in a couple ways. First, string searches of hashtags in New York Times

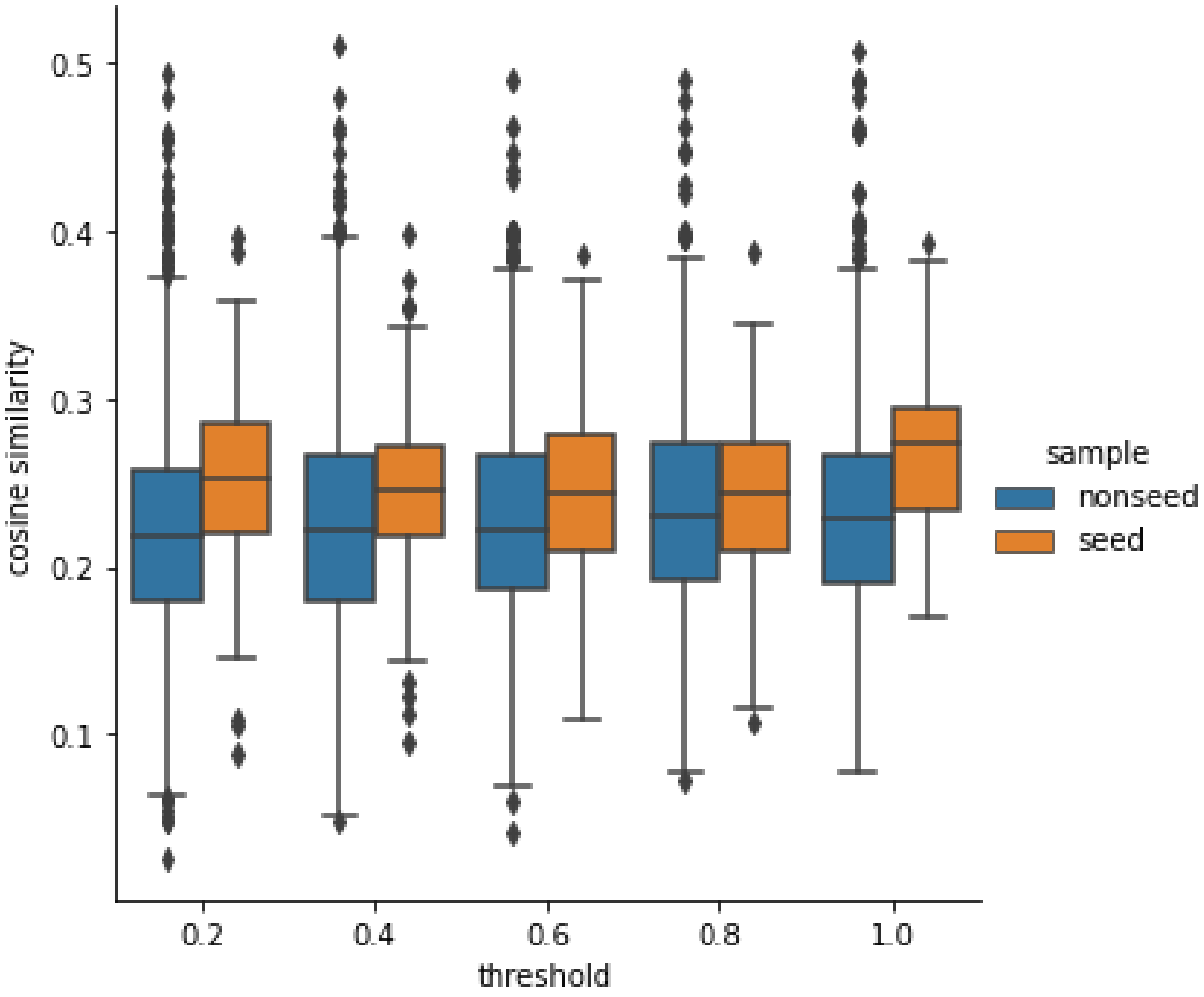



Figure 5.2: Distribution of cosine similarities, between video transcripts and news text, across news interest thresholds and samples. Values increase slightly for the sample seeded by following news accounts.

| Hashtag | Average ranking (out of 100) |
|---|---|
| spirituality | 4.3 |
| avatarthewayofwater | 5.1 |
| sigma | 5.3 |
| jennaortega | 7.7 |
| twd | 9 |
| lyrics | 11.9 |
| thewalkingdead | 13.6 |
| avatar2 | 16.9 |
| spotify | 19.3 |
| jakesully | 20.4 |

Table 5.3: The top-ranked news and entertainment hashtags over a 7-day collection window. Film, television, and music are heavily represented.

headlines find minimal overlap. Out of 1,360 headlines, only 110 contain terms from TikTok hashtags. What overlap does exist stems either from soft news—particularly entertainment—or from the ambiguity of hashtag text. Examples of the former include generic entertainment-related words (e.g., movie, actor, music, drama, songs) or specific references to entertainment properties (e.g., HBO, Wednesday Addams, Elvis). The latter category encompasses words like "show" or "drill". These are terms that, on TikTok, most often reference television or live performance. In New York Times headlines, they are deployed with alternative meanings ("George Santos's Mother Was Not in New York on 9/11, Records Show."; "South Africa to Hold Naval Drill With Russia and China Amid Ukraine War.").

Even in the sparse vocabulary the two samples hold in common, they are not evidently tracking a similar set of specific news events. Rather, there is some broad topical overlap, coupled with some incidental alignment.

These findings suggest that TikTok's popular news and entertainment hashtags are dominated by entertainment. To more comprehensively demonstrate this characteristic, we can categorize all

163 hashtags in the sample. Automated clustering approaches (using, e.g., BERT embeddings) will not work in this context, as the semantic meaning of hashtags is steeped in fast-moving pop culture and platform-specific trends. Instead, we can manually demonstrate the fraction of hashtags that fall under the categories of movies (references to films, actors, characters, or genres), television (references to series, actors, or characters), and music (references to artists and groups, albums and songs, or genres).

In total, 77% of hashtags in this sample—125 out of 163—fall under one of these categories. Movie hashtags are dominated by recent blockbuster release *Avatar: The Way of Water*. TV hashtags reference the show *The Walking Dead* and its characters most often. Music hashtags overwhelmingly reference K-Pop artists and groups.

Of the hashtags that do not fall under these categories, only one is explicitly news related. The hashtag "breakingnews" appears on two days of the sample, at an average rank of 93 out of 100. Among the most popular videos that use this hashtag are some news-related posts: A video from NBC News ranks highly, as does one from Peruvian news outlet La República. However, news coverage and commentary is also mixed in with misinformation (e.g., a CGI video purporting to show two supermassive black holes colliding) and comedy (e.g., the top video in the breaking news hashtag, which shows a talking potato delivering a satirical news broadcast).

These results demonstrate TikTok's overwhelming preference, as expressed by video production and consumption within popular hashtags, for entertainment and other forms of soft news. This suggests a largely separate agenda from that of institutional news, as represented by the New York Times.

### 5.5.3 Underwhelming response at the producer level

Finally, our examination of the 120 news producer accounts in our sample reveals that, for many, TikTok offers relatively sparse engagement. The median account in our news sample has 256,500 followers. 78% of accounts have fewer than 1 million followers, and only five accounts—Bleacher Report, NBC News, Complex, E! News, and NowThis Politics—have more than 3 million followers. TikTok shares little public engagement data, but third parties estimate it has approximately 135 million U.S. users (Howarth, 2022). This would suggest that the largest news accounts have followership from perhaps 2% of U.S. users, while the median's followers equate to 0.2% of those users.

The viewership this sample accrues paints a more conservative picture than its followership. Across their 5 most recent videos, accounts' median views ranged from 165 to 2,000,000. The median view count in this distribution, across 600 videos, was 4,703. To establish a loose on-platform comparison, we also collected a sample of 50 trending videos[9] on 2023/02/07, then calculated the same engagement metrics for the 5 most recent videos from the accounts that posted them. Median views in that sample ranged from 1,384 to 21,907,350, with a median value of 198,800. This suggests that the typical video from an account that is trending on TikTok is 42 times as popular as one from our sample of news producers. A Mann-Whitney U test conducted between the two view count distributions confirms this disparity ($U = 22982$, $p = 2.92 \times 10^{-50}$).

## 5.6 Discussion

In this work, we have demonstrated that across the most prominent surfaces a user might interact with on TikTok, news content is not readily available. The system responsible for recommending

[9] https://ads.tiktok.com/business/creativecenter/inspiration/popular/pc/en

accounts appears unable to consistently surface news producers. The system responsible for the algorithmically-generated For You Page appears reticent to include news content, at least as part of the new account experience. This remains true even when the algorithm is primed with agents that actively follow large news accounts.

At the platform level, these platform characteristics and user preferences work together to co-create a broad lack of trending news topics. Available data on the most popular hashtags on TikTok finds no mention of pressing current events. Rather, entertainment and pop culture dominate. From a design perspective, this suggests a potential need to separate out news and entertainment hashtags into separate categories, to prevent the former from getting drowned out. Similarly, while the news accounts in our sample are active in publishing videos to TikTok, they have not accrued particularly large audiences relative to the size of the active user base, or to accounts with trending content.

Our findings portray a mis-aligned feedback loop among producers, platform, and audience. In the case of users, news interest does appear prevalent despite the lack of news content. According to Pew data, one third of TikTok users regularly turn to the platform for news (Matsa, 2022). While future work should delve deeper into the habits of these users, we can speculate on a few potential explanations for the apparent mismatch between these figures and our results. First, users may be more aggressive in actively seeking out news content, leveraging features like the search bar (Navlakha, 2023). Second, TikTok users' definition of news might include some of the content that we identified in our sample. Whether or not a person perceives social media content as news depends on a range of factors, including its framing, the topics it addresses, and its source (Edgerly & Vraga, 2020). Calling something "news" is a highly subjective judgment, one that might apply to the entertainment news, influencer gossip, and pop science videos we see in our sample. Along similar lines, users may encounter citizen journalism or unverified primary source videos, types of content that often resonate with social media audiences (Leavitt & Clark, 2014). In these cases,

users might simply prefer content other than typical hard news fodder and shape their consumption habits accordingly.

Examining the platform, this work represents a continuation of the recognition that news is far from ubiquitous on social platforms (Thorson, 2020). Some users do encounter news frequently, but only to the extent that they actively seek it out, position themselves in relevant social networks, or are perceived of as salient targets for algorithmic recommendation (Lee & Kim, 2017; Thorson, 2020). This dynamic has the potential to create a self-reinforcing cold start problem: TikTok does not have enough information about new users to provide them high-quality recommendations, so it defaults to known viral content[10]. But it does not include news content in that default set, so it cannot gather engagement signals to inform subsequent news recommendations. This dynamic represents a subtle shift away from the widely disproven idea of filter bubbles (Bruns, 2019). The filter bubble argument positioned news as something that could be avoided in an active sealing off by an algorithm as it personalizes exposure to a user. In contrast, the experience of news avoidance on TikTok seems to happen almost by default as a matter of design. Explicit intervention on the part of the user (e.g. via search) or on the part of the platform designer (e.g. by deliberately including news in recommended content) could both counteract this.

In TikTok's specific case, it is also worthwhile to apply a sociopolitical lens to the company's approach at making recommendations. TikTok has a complex algorithmic decisionmaking system, one that is engineered to maximize user utility in near real time (Liu et al., 2022). Viewed through this lens, TikTok appears as an interface built around some sort of objective optimization function, one that dispassionately matches videos to viewers. However, this technical machinery does not capture the full extent of the company's decisionmaking. At a narrow scale, recommendations are subject to human intervention. Through a practice known as "heating", TikTok employees can

[10]https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

manually boost videos' exposure (Baker-White, 2023). This is often deployed as a strategy to woo advertisers or influencers to the platform. At a broader scale, the company has historically prioritized certain types of content over others. TikTok's explosive growth started in 2017, when it was merged with Musical.ly, an app that allowed users to upload lip sync and dance videos set to popular songs (Russell, 2018). TikTok still retains much of that focus on music, allowing users to select from an enormous library of songs to put in their videos and tracking which songs go viral on the platform. This focus has allowed the platform to garner massive influence in the music industry (Whateley, 2022).

Together, these examples do not paint a picture of a neutral platform focused solely on optimizing user recommendations. Rather, they demonstrate the company's priorities and the strategies it employs to achieve them. In this context, it becomes clearer why news struggles to find an audience on TikTok. The platform's priorities do not appear to favor journalism, strategically or technically.

News producers, then, require novel strategies for platform participation. TikTok, like many social platforms, operates with a logic distinct from that of traditional news publishers (Welbers & Opgenhaffen, 2018). By adapting their content to that logic—by leveraging a performative, illustrative presentation style, for example, or utilizing platform-provided editing tools—news producers may benefit from increased algorithmic amplification (Klug, 2020; Vázquez-Herrero et al., 2020). This strategy runs the risk of news producers shaping their coverage entirely toward characteristics rewarded by the algorithm (Caplan & boyd, 2018), but it provides a potential pathway toward short-term audience growth.

These results also suggest two urgent priorities for system designers. First, recommendation algorithms must be evaluated along dimensions outside of utility maximization. Assessing the democratic or public good function of a set of recommendations may help ameliorate some of the emphasis systems like TikTok's seem to place on entertainment content (Helberger, 2019). Second,

platforms as a whole should continually assess the non-algorithmic pathways they offer for news exposure. In the absence of mechanisms to signal the credibility and authority of journalism, it is becoming apparent that news organizations do not receive substantial attention (Bimber & Gil de Zúñiga, 2020). Explicitly designing mechanisms for regular news exposure may help to address some of the inequality that may otherwise develop in encountering credible news coverage (Thorson, 2020). This could, for instance, include deliberate feed diversification algorithms that increase incidental exposure and create "lower bounds" for societally desirable levels of news exposure.

### 5.6.1 Limitations

In evaluating the recommendations TikTok provides in the For You Page, our approach focuses solely on the new account experience. We chose to repeatedly test fresh accounts to evaluate recommendations in a relatively controlled environment—for accounts with no preexisting information, over the course of one session. However, it is also likely that the recommendation algorithm continually learns users' preferences over repeated sessions. Future work should attempt to examine how recommendations might change over long-term usage.

This approach may also be affected by path dependencies. In other words, the recommendation algorithm may latch onto the first video that an account dwells on, using that behavior to inform subsequent recommendations. This may in turn prevent accounts from seeing news recommendations that they would have encountered had they kept scrolling. However, this possibility is difficult to balance with the desire for realistic user behavior, in which people will often pause to watch videos across multiple areas of interest.

Finally, while we strive in this work to collect data from across TikTok, our sample does not encapsulate every surface. In particular, we do not explore TikTok's search functionality in this work,

which, as noted above, may be a fruitful avenue for news discovery. Future work should examine the extent to which credible news sources appear in search results for news-related queries.

### 5.6.2 Future work

TikTok is a relatively new social platform, and we are just beginning to develop approaches for studying it systematically. To that end, one of the most pressing directions for future work is to develop methods for data collection and analysis.

In one approach, this involves relying primarily on data donations from users. Data donations give a view into real users' experience of the platform, allowing researchers to directly analyze usage data (Boeschoten et al., 2022). Efforts to analyze TikTok through this method have begun to crop up[11]; to the extent that users are willing to share their data, they may generate fruitful datasets for more in-depth analysis of how users engage with TikTok over time.

Another approach involves building software tools that can be distributed openly and leveraged by researchers for subsequent study. Data collection on TikTok involves substantial engineering efforts. Unlike, for example, Twitter and Reddit, it has no external APIs. And unlike many large social platforms, its primary medium is video, rather than text. These features create substantial difficulties for automated data analysis. Encoding visual data and transcribing audio both require computationally expensive state-of-the-art machine learning models. Reconstructing a user's experience of the platform involves subverting TikTok's systems, by successfully emulating human behavior in a browser or mobile application. Developing standardized toolsets for these tasks will open the door for far richer subsequent research.

[11] https://dataskop.net/

### 5.6.3 Conclusion

This study examined the new account experience on TikTok, in an attempt to measure the extent to which the platform recommends news-related accounts and videos. We find a general lack of news coverage in these surfaces, even when providing explicit news interest signals to the algorithm. In examining trending hashtags across news and entertainment videos, we observe an overwhelming preference for film, television, and music content over traditional "hard news" fare. Finally, we find that news producers receive substantially less engagement on their videos than accounts with trending posts. These findings highlight a mechanism by which, in the absence of explicit intervention, an algorithmic recommender can exacerbate news avoidance for all but the most active consumers. They also suggest a need for greater adherence to platform logic on the part of news producers, if they hope to broaden their audiences via algorithmic amplification.

# CHAPTER 6
# DISCUSSION

This dissertation examined approaches to a system-level view to studies of digital media, one that incorporates conceptual and methodological tools from complexity science. Each study focuses on a different set of relationships within the broader system, and how those relationships interact to generate novel outcomes. In this chapter, I present key themes across studies, as well as a roadmap for potential future work.

## 6.1 Unpredictability and understanding

Across these studies, key news processes are at least somewhat unpredictable. In Chapter 3, audiences sort news sources with no discernible pattern. In Chapter 4, audiences exhibit only weakly consistent responses to any given headline writing style. In Chapter 5, user interest in news appears to influence algorithmic recommendations to some extent, but not in a way that surfaces more news content.

It is important to situate the exact level at which this unpredictability occurs. In one view, the conclusions drawn from these findings closely resemble the ecological fallacy created by applying central tendencies to individual cases (Piantadosi et al., 1988). The findings of Chapter 4, for example, exhibit some tension between the broad agreement with prior work we identify in some cases and the unpredictability of individual cases. This view situates the fallacy as the result of distributional characteristics—some fraction of individuals will fall on the outlying portions of a well-defined distribution, and will therefore appear radically different from the expected outcome. However, we can also situate unpredictability within the underlying process that generates the

distribution. In other words, characteristics like the spread of the distribution of interest become measurements of the extent to which a certain process produces consistent outcomes when played out multiple times.

The unpredictability of these processes is a characteristic expected from complex interactions. Martin et al. (2016) posits that there are limits to predictability in many complex systems, as empirical outcomes emerge from a blend of quantifiable explanatory variables ("skill") and seemingly random chance ("luck"). They illustrate this limitation via synthetic viral cascades, in which even perfect knowledge of the process' initial parameters leads to imperfect prediction as outcomes vary. This probabilistic disconnect appears in real-world data as well. Macy et al. (2019) demonstrate the sensitivity of partisan opinions to initial "tipping points", making their ultimate development unpredictable. Salganik et al. (2006) use a similar approach to demonstrate how increased social knowledge increases the unpredictability of music popularity. These studies argue that temporal processes in connected environments are unpredictable because they are not deterministic. Their outcomes vary due to myriad influences, often unobservable because of the lack of an empirical counterfactual.

To some extent, we can apply a similar argument to the studies in this dissertation. News on Reddit appears in a very similar environment to that used in Salganik et al. (2006), one that is ripe for social influence. In addition, news submissions are subject to user behaviors and design decisions outside of the direct purview of an individual story or publisher (Gilbert, 2013). In the case of news headlines, an individual's decision to click on a story is often one choice among many in a sequence of internet browsing decisions (Bentley et al., 2019). That decision might carry little weight for the user, far less than it is ascribed by practitioners or researchers (Kormelink & Meijer, 2018; Makhortykh et al., 2021; Taneja & Yaeger, 2019). And even in the context of that decision making process, individuals are inundated with a flood of "just in time" information that makes

ascribing their choice to any one characteristic difficult (Arapakis et al., 2017; Gal & Simonson, 2021). In the face of so many interacting, shifting influences, it's no wonder that our empirical observation finds very little consistency in collective news behaviors.

At the same time, it is important to consider the extent to which the *context* of these studies shapes their outcome. In particular, Reddit and TikTok both operate using algorithmic recommenders designed to spur some outcomes over others. In Reddit's case, popularity decay is explicitly baked into the platform's sorting mechanism. TikTok, as Chapter 5 explores, applies targeted interventions to on-platform content in service of larger strategic goals. On these platforms, what may look like emergent behavior could actually stem from these types of design decisions (Salganik et al., 2006). But while the extent to which design impacts collective behavior is difficult to untangle, these designed components fundamentally modulate emergence, rather than undermine it. In cases of algorithmic recommenders, users often formulate strategies around how they interact with systems. These strategies center around, for example, maximizing engagement with posted content. But while this strategic behavior occurs in response to designed systems, it is not explicitly informed by that system's mechanisms or priorities. Users may instead develop "folk theories," heuristic representations of their own understanding of a system and how to act in response (DeVito et al., 2018). In this scenario, the control of the system's design is not complete; rather, users' understanding of and interactions with the system create a new channel for emergent behavior.

The ways in which emergent behaviors play out highlight two areas where empirical research can better engage with unpredictability.

First, empirical media research may benefit from augmenting studies of unidirectional relationships of influence with more complex representations. The complexity perspective argues that unidirectional linear relationships cannot adequately capture the dynamics driving outcomes of in-

terest (Ladyman & Wiesner, 2020). Many studies of digital media attempt to model the impact of x on y, for a wide range of empirical relationships: the effect of parts of speech in a headline on readership (Kuiken et al., 2017), the extent to which news outlets adapt their coverage to social media platforms (Hase et al., 2022), audience perceptions of alternative media (Steppat et al., 2021), and many others (Sherry, 2015). This approach produces an unambiguous quantification of a relationship between variables, while limiting the extent to which outside factors can play an explicit role. To give a concrete example, experimental or quasi-experimental research examining media effects finds evidence that exposure to certain kinds of messages can alter perception or awareness in isolation (Berger & Milkman, 2012; Lee & Kim, 2017; León, 1997). On the other hand, research that more broadly encapsulates patterns of media use, alongside the range of other media habits that a person might engage with, demonstrate at best marginal influence from media consumption (Wojcieszak, von Hohenberg, et al., 2022). This type of discrepancy does not invalidate the former approach, but it highlights the potential benefits of the latter—factoring in interactions among variables, measuring how shifting prioritization alters outcomes, and gauging the level of consistency individuals display outside of controlled environments. The primacy of interconnectedness in this approach therefore presents a challenge for lab-based studies. However, if we reject lab-based experimentation on the basis of that limitation, we also lose an important tool for understanding causality and measuring effect sizes. To mitigate this, researchers may require novel approaches to quantifying media effects. Leveraging observational data, methods like regression discontinuity design allow researchers to identify the effect of an exogenous shock on some facet of a system (e.g., measuring the effect of newsroom layoffs on news coverage—Hagar, 2021). In the absence of an application for this kind of statistical approach, researchers might also benefit from novel forms of partnership with industry practitioners. Large newsrooms conduct extensive experimentation—around how their journalism is presented to audiences, but

also in service of deploying novel products, recommendation systems, and tools to communicate with audiences. To the extent that these entities are amenable to academic collaboration, their data offer potential to readily provide measurements of audience behavior, algorithmic recommendation, and other perennial areas of interest for journalism scholars. By plugging into these systems, researchers can leverage ecologically valid, quasi-experimental data that cannot be easily recreated from observation alone.

Second are attempts to claim general findings from examinations of narrow subsets of the media system. Taking Chapter 4 as an example, much research has attempted to construct a set of generalizable recommendations for headline writing best practices. The domains for this type of exploration span email newsletters (Kuiken et al., 2017), digital news portals (Kim et al., 2016), and specific outlets with extensive headline testing data (Banerjee & Urminsky, 2021). In each case, a specific set of textual features is shown to meaningfully influence reader response. However, these findings are also demonstrated in the context of particular sites of news consumption. The logic governing how news gets prioritized, presented, and consumed varies by context, especially when moving between news outlets and social media (Dijck & Poell, 2013; Hurcombe et al., 2019; Welbers & Opgenhaffen, 2018). Similarly, each outlet in these studies carries specific political and social signals for its audience, and specific organizational priorities and strategies (King et al., 2017; Tsfati et al., 2014). These factors can bleed into headline construction in ways that are not directly responsive to data, but that hold significance regardless (Hagar & Diakopoulos, 2019). It is therefore unrealistic to expect that any set of strategies would be repeatably successful across outlets, channels, or media. In these cases, more precise scoping of claims may help us understand the bounds of predictability around a given question.

Given these considerations, how should we approach studying media systems in a way that still provides explanations for key phenomena? I propose two approaches to complement observa-

tional study, which build a potential framework for more comprehensive modeling of unpredictable systems.

First, to return again to C. Anderson (2010), we would benefit from complementing detailed observational work at scale with equally in-depth qualitative research. Much of the unpredictability in these systems results from characteristics or phenomena that are simply not directly measurable via computational means—the motivations of news readers, the strategic reasoning within a newsroom, the profit-motivated tweaks to an algorithm (see Chapter 5). Just as ethnographies and interview studies have elucidated the inner workings of newsroom processes (Boczkowski, 2004; Hagar & Diakopoulos, 2019; Petre, 2015; Shoemaker & Vos, 2009), they and related qualitative methods can help us deeply examine the motivations and behaviors that lead to observed phenomena in digital contexts. Qualitative examination can also feed into computational work, helping to ground models in real-world behavior (Ophir et al., 2020).

Focusing on that computational work, studies of digital systems may benefit from focusing on understanding generative processes, rather than observed phenomena. The study of information cascades provides a useful analogue here. Early studies of cascades in a social media context took a largely descriptive approach, dissecting empirical observations for distinguishing characteristics (Dow et al., 2021). Descriptive work allowed researchers to engineer more general features from empirical samples, and to model the behavior of cascades from those features (Cheng, Adamic, et al., 2014). This modeling work simultaneously highlighted general dynamics of cascades and explored the limits of their predictability from engineered features, motivating researchers to develop generative models of event sequences (Zhou et al., 2021). A recent example of this latter approach is SEISMIC, a point-process based model for predicting tweet activity (Zhao et al., 2015).

For news systems, an analogous approach might drive the kinds of complex, formal modeling that help alleviate the limitations of the unidirectional linear paradigm. Such an approach might

still incorporate strands of this empirical work, but as one step in a process focused on generating generalized understanding of some facet of digital news. In this way, repeated empirical observation becomes a pathway toward elucidating some underlying mechanism, separate from any specific observed case. The studies in this dissertation provide a potential model for the various stages of this approach, in different contexts. Chapter 5 provides an empirical description of several facets of news within a platform, that could provide the basis for more sophisticated modeling of news flow. Chapter 4 attempts a modeling task based on engineered features, across a wider empirical sample than prior descriptive and modeling work. Chapter 2 begins to identify a structural model of freelancer movement, one that can be tested on a broader sample of outlets and generalized. And Chapter 3 posits a generative model of news outlet ranking by popularity. Combining these steps in the study of well-defined structures and processes allows us to retain a connection to the observed phenomena, while also formalizing our theoretical understanding of how they are generated.

Furthermore, this kind of generative modeling work can also be productive for theory building. Computational research can "audition" new candidates for potential causal factors or hypotheses, providing fertile ground for later modeling work (Margolin, 2019). Simulation-based examinations of social systems fulfill a similar role, allowing researchers to work out potential theoretical mechanisms with synthetic agents who follow clear rules (Epstein, 1999). In this framework, the interplay between theoretical work and computational modeling allows them to evolve in tandem.

## 6.2 The political economy of systems

Underlying this work is an essential question of construction: What motivates the relationships we observe among actors? In many cases, they are symbiotic, as in the case of hyperlinks among news publishers (Coddington, 2014; Fu & Shumate, 2016). They may also be a consequence of design,

wherein a platform's social or algorithmic features inherently connect the actors who use them. For the researcher, there is also a mechanistic motivation for associating actors in this way. Actors in the media system, whether publishers, platforms, or consumers, are commonly observed to interact with each other. The object of study reflects the observed phenomena, naturally motivating deployment of the networked paradigm (Qvortrup, 2006).

These examples span *structures* of connection, in which design and position connect actors, as well as *motivation* of connection. The work up to this point has focused largely on the former. However, in examining the latter, we can combine the analytic tools of complex systems with broad-ranging theoretical frameworks of media systems. Interrogating the driving forces of the system, the outcomes they create, and how those outcomes align with actors' motivations allows us to generalize the purpose and operation of the system, and to plug it into related theories.

One framework that considers media systems and the motivations that drive them is the political economy perspective. McChesney (2008) consistently motivates media actions and interactions via the twin forces of regulation and capital. This perspective applies an additional layer of meaning to observed relationships: Journalists rely heavily on official sources because of ownership demands and resource constraints (McChesney, 2003). Hard news and investigation receive less attention because newsrooms have been hollowed out by large corporate owners (McChesney, 2012). Regulators allow some kinds of media and not others implicitly, by the nature of how they construct media markets (McChesney, 1996). Similarly, Hindman (2018) situates news as a market with winners and losers. Some news organizations receive orders of magnitude more attention than others because of their disproportionate institutional resources. Their coverage may also exceed other outlets' in quality and volume, but this is inherently driven by access to more and better talent, more performant technical infrastructure, and a wider range of opportunities for collaboration and partnership. Even in the spread of partisan rhetoric and misinformation, Benkler et al. (2018) finds

motivations tied to capital. The producers of this material disseminate it for financial gain, it gets picked up by more mainstream outlets as it proves its ability to attract attention, and platforms allow its spread because of their profit motive.

The power dynamics explored by the political economy perspective also provide nuance to the unpredictability of digital news. Organizational decision making processes are shaped by interactions among individuals who must balance sometimes competing priorities. That organizational decision making in turn shapes all manner of non-human components within the system: news coverage, the products through which news is conveyed and consumed, and algorithmic recommenders. As such, there is a further complex decision making process latent in many of the interactions examined in this work, one that partially dictates the extent to which an agent's actions can shape a given outcome (Trielli & Diakopoulos, 2019b).

There are many other frameworks (such as social or cognitive factors—see Reese, 2001) that a researcher could apply to generate potential motives for news actors. Political economy is explored in detail here because it fits well with a system-level view, and because it is deployed in many of the studies in this dissertation. In Chapter 2, the prevalence of freelancing in journalism is driven largely by the erosion of full-time newsroom work (Rosenkranz, 2018). Writers' strategic approaches to outlet and topic selection are critical, because they represent strategies for successfully navigating a tenuous labor market (Leung, 2014). At an organizational level, Chapters 3 and 4 address the extent to which an outlet captures audience attention, a critical aspect to generating advertising and subscription revenue. And Chapter 5 explicitly traces the political economy motivations of TikTok in framing itself as an entertainment platform, to the apparent detriment of news producers.

In a narrow sense, the primacy of financial motivations in these systems suggests that the study of (U.S.) media must involve the study of capital and how it is deployed. Researchers must carry

out the messy work of aligning media activity with media business, theoretically and empirically. Computational studies of news generally involve multi million dollar corporations, entities that are publicly traded or owned by catastrophically wealthy individuals. Some efforts have attempted to catalogue these entities, tracking the ownership structure of media organizations ("Index of US Mainstream Media Ownership", 2021) or chronicling growing trends of consolidation within the industry (McChesney, 1996; Winseck, 2022). These high-level overviews could be augmented with publicly-available information about, for example, earnings and corporate strategy, to provide a clearer picture of the motivations of news media as a business.

In thinking about how to systemically capture the influence of regulation and capital in computational work, the pressing question is not one of data access in many cases. Publicly-traded media companies produce large amounts of legally-mandated financial disclosures, and financial services firms like Morningstar publish corresponding analysis. Documents concerning media regulation—legislation, legal opinions, and analysis—are similarly accessible in many jurisdictions. The challenge lies in processing all this information, prioritizing it relative to computational inquiry, and structuring it in such a way that it can be modeled. To that effect, translating information largely stored in unstructured documents may be an area where recent advances in large language models (LLMs) prove impactful. For example, BloombergGPT, Bloomberg's novel LLM, is fine tuned to provide analysis, suggest headlines and stories, and answer questions based on massive amounts of financial data (S. Wu et al., 2023). In doing so, this model provides an interface to data that is otherwise often unstructured. Similar approaches to regulatory documents, or to news coverage and analysis around media ownership, could provide a translation layer between unstructured text and computational modeling, enabling researchers to encode novel types of influence in their empirical work. Further connecting these facets to the day-to-day observation of actual news activities is an open area, but it is one that is critical to deepening our understanding of digital media.

Defining motivation also presents us with an opportunity to put clear bounds on a system of interest. Complex systems are porous. They can extend out to the full set of interacting actors, which, in the case of news media, encompasses a vast set. However, by deploying well-defined theoretical conceptions of consequential relationships within media, we can generate targeted and coherent sub-systems, within which we can begin to define potentially consequential dynamics for the system as a whole. For example, our driving conceptual diagram (Figure 1.1) contains many of the actors who drive processes related to the day-to-day dissemination of news. However, taking cues from McChesney (1996), it may be worth incorporating the relationship between media regulators and producers into our diagram, and interrogating how that relationship impacts news audiences. Theoretically motivated inclusion criteria prevent conceptual systems driven purely by mechanistic interaction in empirical settings, instead encouraging purposeful scoping and mapping.

### 6.3 Defining and determining position

The studies in this dissertation are concerned with determining the relative *position* of actors in the system. Beyond the connections they form, actors may differ in ways that are consequential to how we consider them as members of the media environment. This positioning happens along a number of dimensions:

*Hierarchy:* The idea of nested hierarchy, in which systems collapse down into coherent components, is central to complexity (Simon, 1991). In these studies, we see a similar unfolding depending on the granularity of studies' focus. Chapter 5 attempts to encapsulate a broad range of actors, situated within a broader platform that dictates the shape of their interactions. Chapters 2 and 4 zoom in on particular interactions, revolving around singular artifacts and moments in time, largely within the context of news organizations. Chapter 3 takes an intermediate lens, of a subset of audience evaluating a subset of news, again on a particular social media platform. Hierarchy

allows us to consider multiple fidelities in space and time, modulating the dynamic at play among actors while retaining their essential relationship.

*Bidirectionality:* Central to the networked model of these actors is the idea that influence flows in all directions (Waldherr et al., 2021). This allows us to examine different kinds of influence in turn. Chapters 2 and 4 illustrate this, in that we first consider journalists' impact on audiences by way of content, then of audiences' influence on journalists through the same mode. We can also consider this multidirectional influence all at once, as in the case of Chapter 5. In this context, a feedback loop among all actors is consequential to the ultimate lack of news exposure on the platform. These actors are not merely the senders or receivers of influence. They can flexibly take on the role of either, or they can play a part in a more complicated structure.

*Nonlinearity:* Many of the phenomena examined here follow a power law distribution, in which the most successful actors receive orders of magnitude more reward than others (Allison et al., 1982). As a result, being the *largest* actor in a set of connections carries more weight than a linear relationship might suggest. Similarly, research has demonstrated the role of novelty in a variety of media processes, including consumption and distribution (Gleeson, Cellai, et al., 2014; Harcup & O'Neill, 2017). This makes the *newest* actor in a system disproportionately consequential. These characteristics are relative and change over time, making position a permanently contextual facet of the system.

*Probability:* Finally, complex systems are inherently probabilistic, meaning their outcomes vary according to chance (Ladyman & Wiesner, 2020). This prevents us from too firmly establishing actors' positions, especially at granular levels of focus. We see this clearly in Chapter 3. The relative popularity of *rankings* in the distribution of user attention is stable, but the popularity of individual *outlets* is anything but. This variation emphasizes the fact that positions are snapshots produced by a particular set of conditions, and are merely one permutation of a possible range of

configurations.

Together, these factors work to determine what the underlying structure of the system *looks like*. Through a combination of conceptual understanding and empirical measurement, we can leverage these dimensions to more precisely place media actors in relation to each other, situating power, causality, and consequence accordingly.

## 6.4 Implications for practitioners

For practitioners in the media industry, this work presents challenging conclusions. It argues that much of what seems solid and knowable about news may in fact not be so, that the heuristics that journalists and editors deploy to broaden the reach of their coverage may only be a very small slice of a much larger system of influence. In response, it is worth reflecting on the form and function of analysis within newsrooms.

In form, newsroom analysis seem to fall into many of the traps described here. Efforts to understand data are at best quasi-experimental, seemingly giving newsrooms disproportionate confidence in their findings (Hagar & Diakopoulos, 2019). They are focused largely on specific cases, which then get extrapolated out to general guidance without the intermediate work of modeling underlying processes (Hagar & Diakopoulos, 2019; Petre, 2015). This leaves newsrooms without solid computational work to guide their decision making, which practitioners must respond to with rigorous, longer-term modeling work.

Mitigating this issue requires that practitioners adapt their behavior to the demands of a complex system. Rather than a tool for providing generalizable insights, experimentation should be viewed as a form of optimization toward some specified outcome. This shift necessitates that practitioners do not view any particular case as instructive or explanatory, because of the myriad interacting influences that shape its results. It also demands strict definition of outcomes against

which a system might optimize—increasing the click through rate on headlines, for example. Pairing these criteria enables a model of newsroom experimentation that improves desirable outcomes, in a largely automated way, without running the risk of steering newsroom strategy with erroneous optimization.

At the same time, though, such an approach raises new issues. It leaves a hole in the newsroom's ability to provide journalists with insights about their work. The function of newsroom analysis is also often caught among competing priorities. The business needs of the organization are often well served by computational modeling, as advertising, subscription, and audience engagement are relatively straightforward to quantify (C. Anderson, 2011). Less intuitive is measuring the public value of an investigation, for example, or the utility to the audience of ongoing local coverage. These considerations, in other words, are invisible to the system, which implicitly guides decision making away from them and toward measurable outcomes (Zamith, 2018). One potential approach to address these issues is incorporating the kind of qualitative work described above, which allows newsrooms to consider computational measurement alongside journalistic judgment and audience response. Another might involve rigorous, repeated hypothesis testing within the framework of automated experimentation. This approach still falls victim to the complex nature of news processes, but it does at least provide an understanding of how certain interventions shape outcomes relative to a baseline approach *within a narrow context*. This approach also opens the way for practitioners to consider the extent to which their testing environment might shift, in response to changing news cycles, newsroom strategy, or audience preferences.

## 6.5 Limitations

Given the threads for future empirical and theoretical development evinced in the preceding sections, there are several limitations to this work worth expanding upon.

First is the limited set of actors and connections that it addresses. As explored in Chapter 1, the scoping of this dissertation reflects a desire to interrogate major players in a particular set of news processes. However, those actors do not provide a full representation of the influences shaping digital news media. Notably, actors like regulators and corporate owners—consequential from a political economy perspective—are absent from these analyses. Similarly, this work addresses only a subset of the possible connections drawn among the included actors. Both of these selections on the part of the researcher are theory-driven efforts to scope one possible permutation of a conceptual system. They craft one lens through which we can view the network shaping news processes. By the same token, they are *only* one such permutation. The framework presented in Figure 1.1 contains 5 abstract actors, and maps out 4 relationships in which 2 actors influence a third. In full, there are 30 such possible relationships. As the set of actors considered expands, that number grows rapidly.

The question, then, is how we should conceptualize these missing connections. In one view, full coverage of a system's influences provide the clearest picture of how it functions. Not only is such a comprehensive approach impractical for its sheer scope, though, it also ignores the pruning that prior work can motivate. Some relationships are a priori more important than others, based on existing theory, and should be the focus of limited research resources. Returning to the idea of permutations, then, might provide a more fruitful path forward. Alternative conceptualizations of the processes examined here, also motivated by theory, might bring forward additional nuance or competing explanations for observed phenomena. One avenue for future work might therefore center around the process of generating candidate frameworks for later empirical analysis and systematically evaluating their relative strengths and weaknesses.

Turning to methodology, while the overarching approach of this work drives toward generalizability, it does so by sacrificing some nuance. As Ladyman and Wiesner (2020) notes, models

of complex systems often rely on idealized abstractions, smoothing away variations in empirical observations. As a result, the modeling work of this approach and the conclusions it generates may not apply in specific empirical contexts. The structure modeled in Chapter 2, for example, applies broadly across a large body of writers. It does not necessarily hold for subsets of outlets, or specific types of writers. Two strategies might help address this concern. First, as explored in Section 6.1, accompanying qualitative work can help emphasize the heterogeneity of systems, providing an invaluable companion to computational modeling. Second, it is critical for researchers to clearly scope their claims and the scale at which modeling is occurring, to distinguish the contexts in which an analysis might apply.

## 6.6 Future work

This work proposes a reimaginging of our approach to studying news systems. Underpinning the potential advancements described above is a need for increased research engineering resources, and for a modified conceptual positioning of empirical analysis. To enable these shifts, we require advancement of three research streams.

First, researchers need access to ongoing monitoring of key news processes. In part, this represents a shift in data collection—since it is unclear what fidelity or perspective might be most fruitful for understanding a given outcome in the moment, we should endeavor to capture as much information as possible for later reconstruction. Some work in this vein is currently possible: Researchers can piece together some of the stories published during a given time frame, using resources like Media Cloud and GDELT. They can reconstruct how news traveled through social media, using (a waning number of) open APIs (Weatherbed, 2023). And some researchers can measure direct attention to websites or web pages, if they have access to, e.g., Comscore panel data. However, this piecemeal approach is both incomplete and heavily reliant on non-research

actors. Platforms are increasingly reluctant to share data. Some news data are available, but many examinations remain out of reach without bespoke data collection. For example, Waldherr et al. (2021) uses the evolution of online protests as a running example of how novel theoretical frameworks might be applied in communication research. This type of work might require access to specific news coverage, to social media posts across multiple platforms, and even to information shared in semi-private messaging contexts like Discord or Telegram. Such data are not readily accessible, hampering efforts by researchers without the resources to collect them.

Ongoing monitoring could also represent a novel conceptual paradigm in how researchers address the system-level unpredictability explored above. Rather than basing sampling on a particular window of data collection, researchers could construct longitudinal samples over an extended time period, or purposively sample around, e.g., a salient news cycle. This approach would provide a broader lens through which to view variations across empirical contexts.

Second, the political economy lens explored above suggests a need for access to data about media companies and the markets in which they operate. Such an endeavor requires novel data collection and transformation approaches. More broadly, researchers evaluating these systems increasingly require accessible state-of-the-art engineering toolsets. Platforms like TikTok, which are growing increasingly popular (Vogels et al., 2022) demand far more engineering labor for data collection and analysis. They involve working with multimedia data, collecting post and account information without an open API, and working around stringent platform security measures. Even for text-based resources, state-of-the-art methods involve sophisticated machine learning models that require significant computational resources. Easing this burden as much as possible by providing open sources tools to conduct rigorous computational research at scale will help propel the field forward.

Finally, this work advocates for an approach to news media research that focuses more on

modeling generalizable underlying processes than on measuring effects in constrained samples. This paradigm shifts us toward repeatedly testing and iterating our understanding across contexts, in an effort to do the kind of relative conceptual positioning described above. This suggestion falls in line with recent calls for the field of communication studies to embrace complexity as a driving framework, and to deploy it in conjunction with empirical work (Waldherr et al., 2021). Similarly, the aim expressed here is not to supplant or displace more traditional empirical studies of the news media; rather, it is to offer an additional step, an extension to our current practices that has the potential to formalize and codify our knowledge of these systems. In doing so, we can begin to construct a standardized approach to grappling with this constantly shifting landscape.

## 6.7 Conclusion

The four papers of this dissertation aim to illuminate the interconnections that drive news media and their consequences for audiences, platforms, and producers. No part of news works in isolation, and the relationships formed among the heterogeneous and shifting actors at play have meaningful consequences for all other parts of the system.

Future work should embrace the complexity at play in this space. We now have computational and theoretical tools to contextualize the complex inner workings of news processes. We have the conceptual and analytic vocabulary to trace paths of consequence among multiple actors, in multiple directions, across differing magnitudes of time. Embracing these tools lets us better discern how news works, from the perspective of all the actors who help shape it.

# REFERENCES

Abdollahpouri, H., Malthouse, E., Konstan, J., Mobasher, B., & Gilbert, J. (2021). Toward the Next Generation of News Recommender Systems. *Companion Proceedings of the Web Conference 2021*, 402–406.

Adamic, L., & Glance, N. (2005). The Political Blogosphere and the 2004 U.S. Election: Divided They Blog. *LinkKDD '05*, 36–43.

Allison, P. D., Long, J. S., & Krauze, T. K. (1982). Cumulative Advantage and Inequality in Science. *American Sociological Review*, *47*(5), 615.

Anderson, C. (2010). Journalistic Networks and the Diffusion of Local News: The Brief, Happy News Life of the "Francisville Four". *Political Communication*, *27*(3), 289–309.

Anderson, C. (2011). Between creative and quantified audiences: Web metrics and changing patterns of newswork in local US newsrooms. *Journalism: Theory, Practice & Criticism*, *12*(5), 550–566.

Anderson, S. P., & McLaren, J. (2012). Media Mergers and Media Bias with Rational Consumers. *Journal of the European Economic Association*, *10*(4), 831–859.

Anspach, N. M. (2017). The new personal influence: How our facebook friends influence the news we read. *Political communication*, *34*(4), 590–606.

Antunovic, D., Grzeslo, J., & Hoag, A. (2019). "Ice Cream is Worse, and Joblessness is not an Option" Gendered experiences of freelancing. *Journalism Practice*, *13*(1), 52–67.

Arapakis, I., Cambazoglu, B. B., & Lalmas, M. (2017). On the feasibility of predicting popular news at cold start. *Journal of the Association for Information Science and Technology*, *68*(5), 1149–1164.

Arthur, W. B. (1999). Complexity and the economy. *Science*, *284*(5411), 107–109.

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can

increase political polarization. *Proceedings of the National Academy of Sciences*, *115*(37), 9216–9221.

Bail, C. A., Brown, T. W., & Mann, M. (2017). Channeling hearts and minds: Advocacy organizations, cognitive-emotional currents, and public conversation. *American Sociological Review*, *82*(6), 1188–1213.

Baker-White, E. (2023). TikTok's Secret 'Heating' Button Can Make Anyone Go Viral.

Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, *348*(6239), 1130–1132.

Bandy, J., & Diakopoulos, N. (2020). Auditing news curation systems: A case study examining algorithmic and editorial logic in apple news. *Proceedings of the International AAAI Conference on Web and Social Media*, *14*, 36–47.

Bandy, J., & Diakopoulos, N. (2021). More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proc. ACM Hum.-Comput. Interact.*, *5*(CSCW).

Banerjee, A., & Urminsky, O. (2021). The Language That Drives Engagement: A Systematic Large-scale Analysis of Headline Experiments. *SSRN Electronic Journal*.

Barbieri, N., & Manco, G. (2011). An Analysis of Probabilistic Methods for Top-N Recommendation in Collaborative Filtering. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 172–187). Springer Berlin Heidelberg.

Barnhurst, K. G., & Mutz, D. (1997). American Journalism and the Decline in Event-Centered Reporting. *Journal of Communication*, *47*(4), 27–53.

Barry, C., & Lardner, M. (2011). A Study of First Click Behaviour and User Interaction on the Google SERP. In J. Pokorny, V. Repa, K. Richta, W. Wojtkowski, H. Linger, C. Barry, & M. Lang (Eds.), *Information Systems Development* (pp. 89–99). Springer New York.

Barry, R., West, J., & Wells, G. (2021). Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires.

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016). Seven-in-Ten Reddit Users Get News on the Site.

Becker, L. B., & Vlad, T. (2009). News Organizations and Routines. In *The Handbook of Journalism Studies* (pp. 59–72). Taylor & Francis.

Belyaeva, E., Košmerlj, A., Mladenić, D., & Leban, G. (2018). Automatic Estimation of News Values Reflecting Importance and Closeness of News Events. *Informatica*, *42*(4), 527–533.

Benkler, Y., Faris, R., & Roberts, H. (2018). *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.

Bentley, F., Quehl, K., Wirfs-Brock, J., & Bica, M. (2019). Understanding Online News Behaviors. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 1–11.

Berger, J., & Milkman, K. L. (2012). What Makes Online Content Viral? *Journal of Marketing Research*, *49*(2), 192–205.

Bernhardt, D., Krasa, S., & Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics*, *92*(5-6), 1092–1104.

Berton, F., Devicienti, F., & Pacelli, L. (2011). Are temporary jobs a port of entry into permanent employment?: Evidence from matched employer-employee. *International Journal of Manpower*, *32*(8), 879–899.

Bimber, B., & Gil de Zúñiga, H. (2020). The unedited public sphere. *New Media & Society*, *22*(4), 700–715.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2002). Latent Dirichlet Allocation. *Advances in Neural Information Processing Systems*, 601–608.

Blex, C., & Yasseri, T. (2022). Positive algorithmic bias cannot stop fragmentation in homophilic networks. *The Journal of Mathematical Sociology*, *46*(1), 80–97.

Blom, J. N., & Hansen, K. R. (2015). Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, *76*, 87–100.

Boczkowski, P. J. (2004). The Processes of Adopting Multimedia and Interactivity in Three Online Newsrooms. *Journal of Communication*, *54*(2), 197–213.

Boczkowski, P. J., & de Santos, M. (2007). When More Media Equals Less News: Patterns of Content Homogenization in Argentina's Leading Print and Online Newspapers. *Political Communication*, *24*(2), 167–180.

Boczkowski, P. J., Mitchelstein, E., & Matassi, M. (2018). "news comes across when i'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New media & society*, *20*(10), 3523–3539.

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. *Computational Communication Research*, *4*(2), 388–423.

Bolsen, T., & Shapiro, M. A. (2018). The US News Media, Polarization on Climate Change, and Pathways to Effective Communication. *Environmental Communication*, *12*(2), 149–163.

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the Tone? Easy Doesn't Do It: Analyzing Performance and Agreement Between Off-the-Shelf Sentiment Analysis Tools. *Communication Methods and Measures*, *14*(2), 83–104.

boyd, d., & Crawford, K. (2012). CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Bright, J. (2016). The Social News Gap: How News Reading and News Sharing Diverge: The Social News Gap. *Journal of Communication*, *66*(3), 343–365.

Bruns, A. (2003). Gatewatching, Not Gatekeeping: Collaborative Online News. *Media International Australia*, *107*(1), 31–44.

Bruns, A. (2019). *Are Filter Bubbles Real?* Polity Press.

Bruns, A., & Burgess, J. (2012). Researching news discussion on twitter: New methodologies. *Journalism studies*, *13*(5-6), 801–814.

Buhl, F., Günther, E., & Quandt, T. (2019). Bad News Travels Fastest: A Computational Approach to Predictors of Immediacy in Digital Journalism Ecosystems. *Digital Journalism*.

Burt, R. (1992). *Structural holes*. Harvard University Press.

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 1–12.

Castaldo, M., Venturini, T., Frasca, P., & Gargiulo, F. (2022). Junk news bubbles modelling the rise and fall of attention in online arenas. *New Media & Society*, *24*(9), 2027–2045.

Castellano, C., Fortunato, S., & Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of Modern Physics*, *81*(2), 591–646.

Chan, C.-h., Bajjalieh, J., Auvil, L., Wessler, H., Althaus, S., Welbers, K., van Atteveldt, W., & Jungblut, M. (2021). Four best practices for measuring news sentiment using 'off-the-shelf' dictionaries: A large-scale p-hacking experiment. *Computational Communication Research*, *3*(1), 1–27.

Chandrasekharan, E., Samory, M., Jhaver, S., Charvat, H., Bruckman, A., Lampe, C., Eisenstein, J., & Gilbert, E. (2018). The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1–25.

Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., & Leskovec, J. (2014). Can cascades be predicted? *Proc. of the 23rd International Conference on World Wide Web*, 925–936.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2014). How Community Feedback Shapes User Behavior. *Eighth International AAAI Conference on Weblogs and Social Media*, 41–50.

Chiricos, T., Eschholz, S., & Gertz, M. (1997). Crime, News and Fear of Crime: Toward an Identification of Audience Effects. *Social Problems*, *44*(3), 342–357.

Christin, A., & Petre, C. (2020). Making Peace with Metrics: Relational Work in Online News Production. *Sociologica*, *14*, 133–156.

Cobb, G. W., & Chen, Y.-P. (2003). An Application of Markov Chain Monte Carlo to Community Ecology. *The American Mathematical Monthly*, *110*(4), 265–288.

Coddington, M. (2014). Normalizing the Hyperlink: How bloggers, professional journalists, and institutions shape linking values. *Digital Journalism*, *2*(2), 140–155.

Coddington, M., Lewis, S. C., & Belair-Gagnon, V. (2021). The Imagined Audience for News: Where Does a Journalist's Perception of the Audience Come From? *Journalism Studies*, *22*(8), 1028–1046.

Cohen, N. S. (2019). At work in the digital newsroom. *Digital Journalism*, *7*(5), 571–591.

Cohen, N. S., Hunter, A., & O'Donnell, P. (2019). Bearing the burden of corporate restructuring: Job loss and precarious employment in canadian journalism. *Journalism Practice*, *13*(7), 817–833.

Coleman, J. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology*, *94*, S95–S120.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2017). Very Deep Convolutional Networks for Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 1107–1116.

Couldry, N., & Curran, J. (Eds.). (2003). *Contesting media power: Alternative media in a networked world*. Rowman & Littlefield.

Damstra, A., Vliegenthart, R., Boomgaarden, H., Glüer, K., Lindgren, E., Strömbäck, J., & Tsfati, Y. (2023). Knowledge and the News: An Investigation of the Relation Between News Use, News Avoidance, and the Presence of (Mis)beliefs. *The International Journal of Press/Politics*, *28*(1), 29–48.

Davidson, R., & Meyers, O. (2016). Toward a Typology of Journalism Careers: Conceptualizing Israeli Journalists' Occupational Trajectories: Journalism Careers. *Communication, Culture & Critique*, *9*(2), 193–211.

DeVito, M. A., Birnholtz, J., Hancock, J. T., French, M., & Liu, S. (2018). How People Form Folk Theories of Social Media Feeds and What it Means for How We Study Self-Presentation. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

di Buono, M. P., Šnajder, J., Dalbelo Basic, B., Glavaš, G., Tutek, M., & Milic-Frayling, N. (2017). Predicting News Values from Headline Text and Emotions. *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, 1–6.

Diakopoulos, N. (2019). *Automating the news: How algorithms are rewriting the media*. Harvard University Press.

Diakopoulos, N., Bandy, J., & Dambanemuya, H. (2021). Auditing human-machine communication systems using simulated humans. *Handbook of Human-Machine Communication*.

Dijck, J. V., & Poell, T. (2013). Understanding Social Media Logic. *Media and Communication*, *1*(1), 2–14.

DiPrete, T. A., & Eirich, G. M. (2006). Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments. *Annual Review of Sociology*, *32*(1), 271–297.

Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of pragmatics*, *35*(5), 695–721.

Dow, P. A., Adamic, L., & Friggeri, A. (2021). The anatomy of large facebook cascades. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 145–154.

Dvir-Gvirsman, S. (2017). Media audience homophily: Partisan websites, audience identity and polarization processes. *New Media & Society*, *19*(7), 1072–1091.

Edgerly, S., & Vraga, E. K. (2020). That's Not News: Audience Perceptions of "News-ness" and Why It Matters. *Mass Communication and Society*, *23*(5), 730–754.

Edstrom, M., & Ladendorf, M. (2012). Freelance Journalists as a Flexible Workforce in Media Industries. *Journalism Practice*, *6*(5-6), 711–721.

Ehrenberg, A. S., Goodhardt, G. J., & Barwise, T. P. (1990). Double jeopardy revisited. *Journal of marketing*, *54*(3), 82–91.

England, G. W., Thomas, M., & Paterson, D. G. (1953). Reliability of the original and the simplified Flesch reading ease formulas. *Journal of Applied Psychology*, *37*(2), 111–113.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Fast, E., Chen, B., & Bernstein, M. S. (2016). Empath: Understanding Topic Signals in Large-Scale Text. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657.

Fischer, P., Jonas, E., Frey, D., & Schulz-Hardt, S. (2005). Selective exposure to information: The impact of information limits. *European Journal of Social Psychology*, *35*(4), 469–492.

Flaxman, S., Goel, S., & Rao, J. M. (2016). Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*, *80*(S1), 298–320.

Fletcher, R., Kalogeropoulos, A., & Nielsen, R. K. (2021). More diverse, more politically varied: How social media, search engines and aggregators shape news repertoires in the United Kingdom. *New Media & Society*, 146144482110273.

Fletcher, R., & Nielsen, R. K. (2017). Are news audiences increasingly fragmented? A cross-national comparative analysis of cross-platform news audience fragmentation and duplication. *Journal of Communication*, *67*(4), 476–498.

Fletcher, R., & Nielsen, R. K. (2018). Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society*, *20*(7), 2450–2468.

Forman-Katz, N., & Matsa, K. E. (2022). News Platform Fact Sheet.

Fu, J. S., & Shumate, M. (2016). Hyperlinks as Institutionalized Connective Public Goods for Collective Action Online. *Journal of Computer-Mediated Communication*, *21*(4), 298–311.

Fürst, S. (2020). In the service of good journalism and audience interests? how audience metrics affect news quality. *Media and Communication*, *8*(3), 270–280.

Gal, D., & Simonson, I. (2021). Predicting consumers' choices in the age of the internet, ai, and almost perfect tracking: Some things change, the key challenges do not. *Consumer Psychology Review*, *4*(1), 135–152.

Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the congo, cuba and cyprus crises in four norwegian newspapers. *Journal of peace research*, *2*(1), 64–90.

Garrett, R. K., & Stroud, N. J. (2014). Partisan paths to exposure diversity: Differences in pro-and counterattitudinal news consumption. *Journal of Communication*, *64*(4), 680–701.

Gentile, A. L., Gruhl, D., Ristoski, P., & Welch, S. (2019). Explore and exploit. Dictionary expansion with human-in-the-loop. In P. Hitzler, M. Fernández, K. Janowicz, A. Zaveri, A. J. Gray, V. Lopez, A. Haller, & K. Hammar (Eds.), *The semantic web* (pp. 131–145). Springer International Publishing.

Gilbert, E. (2013). Widespread underprovision on Reddit. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 803–808.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society*, *12*(3), 347–364.

Gillespie, T. (2014). The Relevance of Algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media Technologies* (pp. 167–194). The MIT Press.

Gionis, A., Mannila, H., Mielikäinen, T., & Tsaparas, P. (2007). Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*, *1*(3), 14–es.

Gleeson, J. P., Cellai, D., Onnela, J.-P., Porter, M. A., & Reed-Tsochas, F. (2014). A simple generative model of collective online behavior. *Proceedings of the National Academy of Sciences*, *111*(29), 10411–10415.

Gleeson, J. P., Ward, J. A., O'Sullivan, K. P., & Lee, W. T. (2014). Competition-Induced Criticality in a Model of Meme Popularity. *Physical Review Letters*, *112*(4), 048701.

Gollmitzer, M. (2014). Precariously Employed Watchdogs?: Perceptions of working conditions among freelancers and interns. *Journalism Practice*, *8*(6), 826–841.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Guess, A., & Coppock, A. (2020). Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments. *British Journal of Political Science*, *50*(4), 1497–1515.

Guess, A., Nyhan, B., Lyons, B., & Reifler, J. (2018). Avoiding the echo chamber about echo chambers. *Knight Foundation*, *2*, 1–25.

Günther, E., & Quandt, T. (2018). Word counts and topic models: Automated text analysis methods for digital journalism research. In *Rethinking research methods in an age of digital journalism* (pp. 75–88). Routledge.

Hagar, N. (2021). Newsroom Layoffs Decrease News Coverage Diversity. *ICA 2021*.

Hagar, N., Bandy, J., Trielli, D., Wang, Y., & Diakopoulos, N. (2020). Defining Local News: A Computational Approach. *Computation + Journalism Symposium*.

Hagar, N., & Diakopoulos, N. (2019). Optimizing Content with A/B Headline Testing: Changing Newsroom Practices. *Media and Communication*, *7*(1), 117–127.

Hagar, N., Diakopoulos, N., & DeWilde, B. (2021). Anticipating Attention: On the Predictability of News Headline Tests. *Digital Journalism*.

Hagar, N., & Shaw, A. (2022). Concentration without cumulative advantage: The distribution of news source attention in online communities. *Journal of Communication*, *72*(6), 675–686.

Hagar, N., Wachs, J., & Horvát, E.-Á. (2021). Writer movements between news outlets reflect political polarization in media. *New Media & Society*.

Harcup, T., & O'Neill, D. (2017). What is News?: News values revisited (again). *Journalism Studies*, *18*(12), 1470–1488.

Harder, R. A., Sevenans, J., & Van Aelst, P. (2017). Intermedia Agenda Setting in the Social Media Age: How Traditional Players Dominate the News Agenda in Election Times. *The International Journal of Press/Politics*, *22*(3), 275–293.

Hase, V., Boczek, K., & Scharkow, M. (2022). Adapting to affordances and audiences? a cross-platform, multi-modal analysis of the platformization of news on facebook, instagram, tiktok, and twitter. *Digital Journalism*, 1–22.

Hasell, A., & Weeks, B. E. (2016). Partisan provocation: The role of partisan news use and emotional responses in political information sharing in social media. *Human Communication Research*, *42*(4), 641–661.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In *The Elements of Statistical Learning* (pp. 587–604). Springer.

Hayes, K., & Silke, H. (2018). The Networked Freelancer?: Digital labour and freelance journalism in the age of social media. *Digital Journalism*, *6*(8), 1018–1028.

Heatherly, K. A., Lu, Y., & Lee, J. K. (2017). Filtering out the other side? Cross-cutting and like-minded discussions on social networking sites. *New Media & Society*, *19*(8), 1271–1289.

Helberger, N. (2019). On the Democratic Role of News Recommenders. *Digital Journalism*, *7*(8), 993–1012.

Hernández-Serrano, M.-J., Renés-Arellano, P., Graham, G., & Greenhill, A. (2017). From prosumer to prodesigner: Participatory news consumption. *Comunicar. Media Education Research Journal*, *25*(1).

Hindman, M. (2015). Building Better Models: Prediction, Replication, and Machine Learning in the Social Sciences. *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 48–62.

Hindman, M. (2018). *The internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton University Press.

Hindman, M., & Rogers, B. (2018). The Dynamics of Web Traffic. In *The internet trap: How the digital economy builds monopolies and undermines democracy*. Princeton University Press.

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength natural language processing in python.

Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, *54*(1), 229–247.

Horvát, E.-Á., Wachs, J., Wang, R., & Hannák, A. (2018). The role of novelty in securing investors for equity crowdfunding campaigns. *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Howarth, J. (2022). TikTok User Age, Gender, & Demographics (2023).

Hurcombe, E., Burgess, J., & Harrington, S. (2019). What's newsworthy about 'social news'? Characteristics and potential of an emerging genre. *Journalism*, 378–394.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth international AAAI conference on weblogs and social media*.

Ifantidou, E. (2009). Newspaper headlines and relevance: Ad hoc concepts in ad hoc contexts. *Journal of Pragmatics*, *41*(4), 699–720.

Index of US Mainstream Media Ownership. (2021).

Iyengar, S., & Hahn, K. S. (2009). Red Media, Blue Media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication*, *59*(1), 19–39.

Jones, R. (2023). Meta Has a TikTok Problem, and It's Hiring BuzzFeed to Help.

Joo, J., & Steinert-Threlkeld, Z. C. (2022). Image as data: Automated content analysis for visual presentations of political actors and events. *Computational Communication Research*, *4*(1).

Jungherr, A., Posegga, O., & An, J. (2019). Discursive Power in Contemporary Media Systems: A Comparative Framework. *The International Journal of Press/Politics*, *24*(4), 404–425.

Kaiser, J., Rauchfleisch, A., & Bourassa, N. (2019). Connecting the (Far-)Right Dots: A Topic Modeling and Hyperlink Analysis of (Far-)Right Media Coverage during the US Elections 2016. *Digital Journalism*, *8*(3), 422–441.

Karnowski, V., Leiner, D. J., Sophie Kümpel, A., & Leonhard, L. (2020). Worth to Share? How Content Characteristics and Article Competitiveness Influence News Sharing on Social Network Sites. *Journalism & Mass Communication Quarterly*.

Kaye, B. (2022). Australia says law making Facebook and Google pay for news has worked. *Reuters*.

Kenter, T., & de Rijke, M. (2015). Short Text Similarity with Word Embeddings. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, 1411–1420.

Kessler, S. H., & Engelmann, I. (2019). Why do we click? investigating reasons for user selection on a news aggregator website. *Communications*, *44*(2), 225–247.

Kim, J. H., Mantrach, A., Jaimes, A., & Oh, A. (2016). How to compete online for news audience: Modeling words that attract clicks. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1645–1654.

King, G., Schneer, B., & White, A. (2017). How the news media activate public expression and influence national agendas. *Science*, *358*(6364), 776–780.

Klinenberg, E. (2005). Convergence: News Production in a Digital Age. *The ANNALS of the American Academy of Political and Social Science*, *597*(1), 48–64.

Klingenstein, S., Hitchcock, T., & DeDeo, S. (2014). The civilizing process in london's old bailey. *Proceedings of the National Academy of Sciences*, *111*(26), 9419–9424.

Klug, D. (2020). "Jump in and be Part of the Fun". How U.S. News Providers Use and Adapt to TikTok. *Midwest Popular Culture Association/Midwest American Culture Association Annual Conference*.

Knobloch-Westerwick, S., & Meng, J. (2009). Looking the Other Way: Selective Exposure to Attitude-Consistent and Counterattitudinal Political Information. *Communication Research*, *36*(3), 426–448.

Kormelink, T. G., & Meijer, I. C. (2018). What clicks actually mean: Exploring digital news user practices. *Journalism*, *19*(5), 668–683.

Kreiss, D., & Mcgregor, S. C. (2018). Technology Firms Shape Political Communication: The Work of Microsoft, Facebook, Twitter, and Google With Campaigns During the 2016 U.S. Presidential Cycle. *Political Communication*, *35*(2), 155–177.

Kuiken, J., Schuth, A., Spitters, M., & Marx, M. (2017). Effective Headlines of Newspaper Articles in a Digital Environment. *Digital Journalism*, *5*(10), 1300–1314.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network or a news media? *Proceedings of the 19th international conference on World wide web*, 591–600.

Ladyman, J., & Wiesner, K. (2020). *What is a Complex System?* Yale University Press.

Lamberson, P., & Soroka, S. (2018). A model of attentiveness to outlying news. *Journal of Communication*, *68*(5), 942–964.

Lampe, C., & Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 543–550.

Lamprinidis, S., Hardt, D., & Hovy, D. (2018). Predicting News Headline Popularity with Syntactic and Semantic Knowledge Using Multi-Task Learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 659–664.

Latzer, M., & Just, N. (2020). Governance by and of algorithms on the internet: Impact and consequences. In *Oxford research encyclopedia of communication*. Oxford University Press.

Lau, J. H., & Baldwin, T. (2016). An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. (2009). Computational social science. *Science*, *323*(5915), 721–723.

Leavitt, A., & Clark, J. A. (2014). Upvoting hurricane Sandy: Event-based news production processes on a social news site. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1495–1504.

Lebow, S. (2022). Why time spent with TikTok is on the decline.

Lee, J. K., & Kim, E. (2017). Incidental exposure to news: Predictors in the social media setting and effects on information gain online. *Computers in Human Behavior*, *75*, 1008–1015.

León, J. A. (1997). The effects of headlines and summaries on news comprehension and recall. *Reading and Writing*, *9*(2), 85–106.

Lerman, K., & Hogg, T. (2010). Using a Model of Social Dynamics to Predict Popularity of News. *Proceedings of the 19th International Conference on World Wide Web*, 621–630.

Leung, M. D. (2014). Dilettante or Renaissance Person? How the Order of Job Experiences Affects Hiring in an External Labor Market. *American Sociological Review*, *79*(1), 136–158.

Lewis, S. C., & Westlund, O. (2015). Actors, Actants, Audiences, and Activities in Cross-Media News Work: A matrix and a research agenda. *Digital Journalism*, *3*(1), 19–37.

Li, Q., Wang, J., Chen, Y. P., & Lin, Z. (2010). User comments for news recommendation in forum-based social media. *Information Sciences*, *180*(24), 4929–4939.

Liu, Z., Zou, L., Zou, X., Wang, C., Zhang, B., Tang, D., Zhu, B., Zhu, Y., Wu, P., Wang, K., & Cheng, Y. (2022). Monolith: Real Time Recommendation System With Collisionless Embedding Table.

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, *10*(1), 1759.

Lowrey, W. (2006). Mapping the journalism–blogging relationship. *Journalism: Theory, Practice & Criticism*, *7*(4), 477–500.

Lowrey, W. (2012). Journalism Innovation and the Ecology of News Production: Institutional Tendencies. *Journalism & Communication Monographs*, *14*(4), 214–287.

Macy, M., Deri, S., Ruch, A., & Tong, N. (2019). Opinion cascades and the unpredictability of partisan polarization. *Science Advances*, *5*(8), eaax0754.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., et al. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2-3), 93–118.

Maier, S. (2010). All the News Fit to Post? Comparing News Content on the Web to Newspapers, Television, and Radio. *Journalism & Mass Communication Quarterly*, *87*(3-4), 548–562.

Makhortykh, M., de Vreese, C., Helberger, N., Harambam, J., & Bountouridis, D. (2021). We are what we click: Understanding time and content-based habits of online news readers. *new media & society*, *23*(9), 2773–2800.

Margolin, D. B. (2019). Computational Contributions: A Symbiotic Approach to Integrating Big, Observational Data Studies into the Communication Field. *Communication Methods and Measures*, *13*(4), 229–247.

Marín-Sanchiz, C.-R., Carvajal, M., & González-Esteban, J.-L. (2021). Survival Strategies in Freelance Journalism: An Empowering Toolkit to Improve Professionals' Working Conditions. *Journalism Practice*, 1–24.

Martin, T., Hofman, J. M., Sharma, A., Anderson, A., & Watts, D. J. (2016). Exploring Limits to Prediction in Complex Social Systems. *Proceedings of the 25th International Conference on World Wide Web - WWW '16*, 683–694.

Matsa, K. E. (2022). More Americans are getting news on TikTok, bucking the trend on other social media sites.

McCallum, A. K. (2002). *Mallet: A machine learning for language toolkit* (tech. rep.). UMass.

McChesney, R. W. (1996). The Internet and U. S. Communication Policy-Making in Historical and Critical Perspective. *Journal of Communication*, *46*(1), 98–124.

McChesney, R. W. (2003). The Problem of Journalism: A political economic contribution to an explanation of the crisis in contemporary US journalism. *Journalism Studies*, *4*(3), 299–329.

McChesney, R. W. (2008). *The Political Economy of Media: Enduring Issues, Emerging Dilemmas*. Monthly Review Press.

McChesney, R. W. (2012). FAREWELL TO JOURNALISM?: Time for a rethinking. *Journalism Studies*, *13*(5-6), 682–694.

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

McCombs, M. E., & Winter, J. P. (1981). Defining Local News. *Newspaper Research Journal*, *3*(1), 16–21.

McLachlan, S. (2021). Do "For You Page" Hashtags Actually Work on TikTok?

Messing, S., & Westwood, S. J. (2014). Selective Exposure in the Age of Social Media: Endorsements Trump Partisan Source Affiliation When Selecting News Online. *Communication Research*, *41*(8), 1042–1063.

Miller, J. H., & Page, S. E. (2007). *Complex Adaptive Systems: An Introduction to Computational Models of Social Life*. Princeton University Press.

Mitchell, A., Jurkowitz, M., Oliphant, J. B., & Shearer, E. (2020). Americans Who Mainly Get Their News on Social Media Are Less Engaged, Less Knowledgeable.

Mitchelstein, E., & Boczkowski, P. J. (2009). Between tradition and change: A review of recent research on online news production. *Journalism: Theory, Practice & Criticism*, *10*(5), 562–586.

Mohr, J. W., & Bogdanov, P. (2013). Introduction—topic models: What they are and why they matter [Topic Models and the Cultural Sciences]. *Poetics*, *41*(6), 545–569.

Montgomery, J. M., & Nyhan, B. (2017). The Effects of Congressional Staff Networks in the US House of Representatives. *The Journal of Politics*, *79*(3), 745–761.

Mullainathan, S., & Shleifer, A. (2005). The Market for News. *American Economic Review*, *95*(4), 1031–1053.

Munger, K. (2020). All the News That's Fit to Click: The Economics of Clickbait Media. *Political Communication*, *37*(3), 376–397.

Munroe, R. (2009). Reddit's new comment sorting system.

Navlakha, M. (2023). TikTok knows you're using it as a search engine. It's even made an ad.

Nelson, J. L., & Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New media & society*, *20*(10), 3720–3737.

Nelson, J. L., & Webster, J. G. (2017). The Myth of Partisan Selective Exposure: A Portrait of the Online Political News Audience. *Social Media + Society*, *3*(3).

Newman, M. E. J. (2001). Scientific collaboration networks. I. Network construction and fundamental results. *Physical Review E*, *64*(1), 016131.

Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, *45*(2), 167–256.

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, *103*(23), 8577–8582.

Newman, M. E. J. (2010). *Networks: An introduction* [OCLC: ocn456837194]. Oxford University Press.

Newman, N. (2022). *How Publishers are Learning to Create and Distribute News on TikTok* (tech. rep.). Reuters Institue for the Study of Journalism.

Nielsen, F. A. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages.*, *718 in CEUR Workshop Proceedings*, 93–98.

Nielsen, R., & Ganter, S. A. (2018). Dealing with digital intermediaries: A case study of the relations between publishers and platforms. *New Media & Society*, *20*(4), 1600–1617.

Nygren, G., Leckner, S., & Tenor, C. (2018). Hyperlocals and legacy media: Media ecologies in transition. *Nordicom Review*, *39*(1), 33–49.

Ophir, Y., Walter, D., & Marchant, E. R. (2020). A collaborative way of knowing: Bridging computational communication research and grounded theory ethnography. *Journal of Communication*, *70*(3), 447–472.

Panigrahi, A., Simhadri, H. V., & Bhattacharyya, C. (2019). Word2Sense: Sparse Interpretable Word Embeddings. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5692–5705.

Parks, P. (2019). Textbook News Values: Stable Concepts, Changing Choices. *Journalism & Mass Communication Quarterly*, *96*(3), 784–810.

Patel, S. (2020). Reddit Claims 52 Million Daily Users, Revealing a Key Figure for Social-Media Platforms - WSJ.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The Ecological Fallacy. *American Journal of Epidemiology*, *127*(5), 893–905.

Primo, A., & Zago, G. (2015). Who And What Do Journalism?: An actor-network perspective. *Digital Journalism*, *3*(1), 38–52.

Prior, M. (2007). *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press.

Qvortrup, L. (2006). Understanding New Digital Media: Medium Theory or Complexity Theory? *European Journal of Communication*, *21*(3), 345–356.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Rao, A. R., Jana, R., & Bandyopadhyay, S. (1996). A Markov Chain Monte Carlo Method for Generating Random (0, 1)-Matrices with Given Marginals. *Sankhyā: The Indian Journal of Statistics, Series A*, *58*(2), 225–242.

Ratkiewicz, J., Fortunato, S., Flammini, A., Menczer, F., & Vespignani, A. (2010). Characterizing and Modeling the Dynamics of Online Popularity. *Physical Review Letters*, *105*(15).

Rayson, S. (2017). We Analyzed 100 Million Headlines. Here's What We Learned (New Research).

Reese, S. D. (2001). Understanding the Global Journalist: A hierarchy-of-influences approach. *Journalism Studies*, *2*(2), 173–187.

Řehůřek, R., & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, N., & Gurevych, I. (2019a). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3980–3990.

Reimers, N., & Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Resnick, P., Garrett, R. K., Kriplean, T., Munson, S. A., & Stroud, N. J. (2013). Bursting your (filter) bubble: Strategies for promoting diverse exposure. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion - CSCW '13*, 95.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141.

Roberts, H., Bhargava, R., Valiukas, L., Jen, D., Malik, M. M., Bishop, C. S., Ndulue, E. B., Dave, A., Clark, J., Etling, B., et al. (2021). Media cloud: Massive open source collection of global news on the open web. *Proceedings of the International AAAI Conference on Web and Social Media*, *15*, 1034–1045.

Rosenkranz, T. (2016). Becoming entrepreneurial: Crisis, ethics and marketization in the field of travel journalism. *POETICS*, *54*, 54–65.

Rosenkranz, T. (2018). From Contract to Speculation: New Relations of Work and Production in Freelance Travel Journalism. *Work, Employment and Society*, *33*(4), 613–630.

Ross, A. A. (2017). "If Nobody Gives a Shit, is it Really News?": Changing standards of news production in a learning newsroom. *Digital Journalism*, *5*(1), 82–99.

Rubin, A. M. (1993). Audience activity and media use. *Communication Monographs*, *60*(1), 98–105.

Russell, J. (2018). Short video service Musical.ly is merging into sister app TikTok.

Russell Neuman, W., Guggenheim, L., Mo Jang, S., & Bae, S. Y. (2014). The Dynamics of Public Attention: Agenda-Setting Theory Meets Big Data: Dynamics of Public Attention. *Journal of Communication*, *64*(2), 193–214.

Ryan, M. (2001). Journalistic Ethics, Objectivity, Existential Journalism, Standpoint Epistemology, and Public Journalism. *Journal of Mass Media Ethics*, *16*(1), 3–22.

Salamon, E. (2020). Digitizing freelance media labor: A class of workers negotiates entrepreneurialism and activism. *New Media & Society*, *22*(1), 105–122.

Salganik, M. J., Dodds, P. S., & Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, *311*(5762), 854–856.

Sawyer, R. K. (2005). *Social Emergence: Societies as Complex Systems*. Cambridge University Press.

Schiffer, A. J. (2006). Assessing Partisan Bias in Political News: The Case(s) of Local Senate Election Coverage. *Political Communication*, *23*(1), 23–39.

Schlauch, W. E., Horvát, E. Á., & Zweig, K. A. (2015). Different flavors of randomness: Comparing random graph models with fixed degree sequences. *Social Network Analysis and Mining*, *5*(36).

Schrodt, P. A. (2010). Automated production of high-volume, real-time political event data. *Apsa 2010 annual meeting paper*.

Schudson, M. (2001). The objectivity norm in American journalism*. *Journalism: Theory, Practice & Criticism*, *2*(2), 149–170.

Schudson, M., & Anderson, C. (2009). Objectivity, Professionalism, and Truth Seeking in Journalism. In *The Handbook of Journalism Studies* (pp. 88–101). Routledge.

Scolari, C. A. (2012). Media ecology: Exploring the metaphor to expand the theory. *Communication Theory*, *22*(2), 204–225.

Shaker, L. (2014). Dead Newspapers and Citizens' Civic Engagement. *Political Communication*, *31*(1), 131–148.

Shaw, A. (2012). Centralized and decentralized gatekeeping in an open online collective. *Politics & Society*, *40*(3), 349–388.

Sherry, J. (2015). The Complexity Paradigm for Studying Human Communication: A Summary and Integration of Two Fields. *Review of Communication Research*, *3*, 22–65.

Shi, F., Shi, Y., Dokshin, F. A., Evans, J. A., & Macy, M. W. (2017). Millions of online book co-purchases reveal partisan differences in the consumption of science. *Nature Human Behaviour*, *1*(4), 0079.

Shin, J., Madotto, A., & Fung, P. (2018). Interpreting Word Embeddings with Eigenvector Analysis. *Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310.

Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. Routledge
OCLC: ocn258331230.

Shulman, B., Sharma, A., & Cosley, D. (2016). Predictability of Popularity: Gaps between Prediction and Understanding. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*.

Simon, H. A. (1991). The architecture of complexity. In *Facets of systems science* (pp. 457–476). Springer.

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.

Singer, J. B. (2005). The political j-blogger: 'Normalizing' a new media form to fit old norms and practices. *Journalism: Theory, Practice & Criticism*, *6*(2), 173–198.

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892.

Steensen, S., Ferrer-Conill, R., & Peters, C. (2020). (Against a) Theory of Audience Engagement with News. *Journalism Studies*, *21*(12), 1662–1680.

Steppat, D., Castro, L., & Esser, F. (2021). What news users perceive as 'alternative media'varies between countries: How media fragmentation and polarization matter. *Digital Journalism*, 1–21.

Storey, J., Salaman, G., & Platman, K. (2005). Living with enterprise in an enterprise economy: Freelance and contract workers in the media. *Human Relations*, *58*(8), 1033–1054.

Stroud, N. J. (2010). Polarization and Partisan Selective Exposure. *Journal of Communication*, *60*(3), 556–576.

Sylwester, K., & Purver, M. (2015). Twitter language use reflects psychological differences between democrats and republicans. *PloS ONE*, *10*(9), e0137422.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of the ACM*, *53*(8), 80–88.

Szymanski, T., Orellana-Rodriguez, C., & Keane, M. T. (2016). Helping News Editors Write Better Headlines: A Recommender to Improve the Keyword Contents & Shareability of News Headlines. *Natural Language Processing Meets Journalism. IJCAI.*

Taneja, H., & Webster, J. G. (2016). How Do Global Audiences Take Shape? The Role of Institutions and Culture in Patterns of Web Use: The Role of Institutions and Culture in Patterns of Web Use. *Journal of Communication*, *66*(1), 161–182.

Taneja, H., Webster, J. G., Malthouse, E. C., & Ksiazek, T. B. (2012). Media consumption across platforms: Identifying user-defined repertoires. *New media & society*, *14*(6), 951–968.

Taneja, H., & Wu, A. X. (2018). Pathways to Fragmentation: User Flows and Web Distribution Infrastructures. *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, 255–259.

Taneja, H., & Yaeger, K. (2019). Do People Consume the News they Trust? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*, 540:1–540:10.

Tatar, A., Antoniadis, P., de Amorim, M. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*(1), 174.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24–54.

Tetlock, P., & Gardner, D. (2015). *Superforecasting: The Art and Science of Prediction*. Crown Publishers.

Tewksbury, D., & Riles, J. M. (2015). Polarization as a Function of Citizen Predispositions and Exposure to News on the Internet. *Journal of Broadcasting & Electronic Media*, *59*(3), 381–398.

Thompson, A. (2017). All the news: 143,000 articles from 15 American publications.

Thorson, K. (2020). Attracting the news: Algorithms, platforms, and reframing incidental exposure. *Journalism*, *21*(8), 1067–1082.

Thorson, K., Cotter, K., Medeiros, M., & Pak, C. (2021). Algorithmic inference, political interest, and exposure to news and politics on Facebook. *Information, Communication & Society*, *24*(2), 183–200.

Thorson, K., Medeiros, M., Cotter, K., Chen, Y., Rodgers, K., Bae, A., & Baykaldi, S. (2020). Platform Civics: Facebook in the Local Information Infrastructure. *Digital Journalism*, *8*(10), 1231–1257.

Thorson, K., & Wells, C. (2016). Curated flows: A framework for mapping media exposure in the digital age. *Communication Theory*, *26*(3), 309–328.

TikTok users worldwide (2020-2025). (2023).

Toff, B., & Kalogeropoulos, A. (2020). All the News That's Fit to Ignore. *Public Opinion Quarterly*, *84*(S1), 366–390.

Trielli, D., & Diakopoulos, N. (2019a). Search as News Curator: The Role of Google in Shaping Attention to News Information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*.

Trielli, D., & Diakopoulos, N. (2019b). Search as news curator: The role of google in shaping attention to news information. *Proceedings of the 2019 CHI Conference on human factors in computing systems*, 1–15.

Trielli, D., & Diakopoulos, N. (2022). Partisan search behavior and google results in the 2018 us midterm elections. *Information, Communication & Society*, *25*(1), 145–161.

Trilling, D., Tolochko, P., & Burscher, B. (2017). From Newsworthiness to Shareworthiness: How to Predict News Sharing Based on Article Characteristics. *Journalism & Mass Communication Quarterly*, *94*(1), 38–60.

Tsfati, Y., Stroud, N. J., & Chotiner, A. (2014). Exposure to Ideological News and Perceived Opinion Climate: Testing the Media Effects Component of Spiral-of-Silence in a Fragmented Media Landscape. *The International Journal of Press/Politics*, *19*(1), 3–23.

Tumminello, M., Micciché, S., Lillo, F., Piilo, J., & Mantegna, R. N. (2011). Statistically Validated Networks in Bipartite Complex Systems (E. Ben-Jacob, Ed.). *PLoS ONE*, *6*(3), e17994.

Valenzuela, S., Piña, M., & Ramírez, J. (2017). Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing: Framing Effects on News Sharing. *Journal of Communication*, *67*(5), 803–826.

van Atteveldt, W., Margolin, D., Shen, C., Trilling, D., & Weber, R. (2019). A Roadmap for Computational Communication Research. *Computational Communication Research*, *1*(1), 1–11.

Van Dijck, J. (2014). Datafication, dataism and dataveillance: Big data between scientific paradigm and ideology. *Surveillance & society*, *12*(2), 197–208.

Vasterman, P. L. (2005). Media-Hype: Self-Reinforcing News Waves, Journalistic Standards and the Construction of Social Problems. *European Journal of Communication*, *20*(4), 508–530.

Vázquez-Herrero, J., Negreira-Rey, M.-C., & López-García, X. (2020). Let's dance the news! How the news media are adapting to the logic of TikTok. *Journalism*, 146488492096909.

Vedres, B., & Stark, D. (2010). Structural folds: Generative disruption in overlapping groups. *American Journal of Sociology*, *115*(4), 1150–1190.

Vermeer, S., Trilling, D., Kruikemeier, S., & de Vreese, C. (2020). Online News User Journeys: The Role of Social Media, News Websites, and Topics. *Digital Journalism*, *8*(9), 1114–1141.

Villi, M., Aharoni, T., Tenenboim-Weinblatt, K., Boczkowski, P. J., Hayashi, K., Mitchelstein, E., Tanaka, A., & Kligler-Vilenchik, N. (2022). Taking a Break from News: A Five-nation Study of News Avoidance in the Digital Era. *Digital Journalism*, *10*(1), 148–164.

Vogels, E. A., Gelles-Watnick, R., & Massarat, N. (2022). Teens, Social Media and Technology 2022.

Waldherr, A. (2014). Emergence of News Waves: A Social Simulation Approach: Emergence of News Waves. *Journal of Communication*, *64*(5), 852–873.

Waldherr, A., Geise, S., Mahrt, M., Katzenbach, C., & Nuernbergk, C. (2021). Toward a stronger theoretical grounding of computational communication science: How macro frameworks shape our research agendas. *Computational Communication Research*, *3*(2), 1–28.

Wall, M. (2015). Citizen journalism: A retrospective on what we know, an agenda for what we don't. *Digital journalism*, *3*(6), 797–813.

Weatherbed, J. (2023). Twitter is replacing free access to its API with a new paid tier.

Weaver, D. H., & Wilhoit, G. C. (1986). *The American journalist: A portrait of U.S. news people and their work*. Indiana University Press.

Weber, M. S. (2012). Newspapers and the Long-Term Implications of Hyperlinking. *Journal of Computer-Mediated Communication*, *17*(2), 187–201.

Webster, J. G. (2009). The role of structure in media choice. In *Media choice* (pp. 235–247). Routledge.

Webster, J. G. (2011). The duality of media: A structurational theory of public attention. *Communication Theory*, *21*(1), 43–66.

Webster, J. G. (2016). The Marketplace of Attention. In *The marketplace of attention: How audiences take shape in a digital age*. The MIT Press
OCLC: 962400132.

Webster, J. G., & Ksiazek, T. B. (2012). The Dynamics of Audience Fragmentation: Public Attention in an Age of Digital Media. *Journal of Communication*, *62*(1), 39–56.

Welbers, K., & Opgenhaffen, M. (2018). Presenting News on Social Media: Media logic in the communication style of newspapers on Facebook. *Digital Journalism*, 45–62.

Whateley, D. (2022). How TikTok is changing the music industry.

White, D. M. (1950). The "gate keeper": A case study in the selection of news. *Journalism quarterly*, *27*(4), 383–390.

Wihbey, J., Joseph, K., & Lazer, D. (2019). The social silos of journalism? Twitter, news media and partisan segregation. *New Media & Society*, *21*(4), 815–835.

Winseck, D. (2022). Media and internet concentration in canada, 1984–2021.

Wojcieszak, M., Menchen-Trevino, E., Goncalves, J. F., & Weeks, B. (2022). Avenues to news and diverse news exposure online: Comparing direct navigation, social media, news aggregators, search queries, and article hyperlinks. *The International Journal of Press/Politics*, *27*(4), 860–886.

Wojcieszak, M., von Hohenberg, B. C., Casas, A., Menchen-Trevino, E., de Leeuw, S., Gonçalves, A., & Boon, M. (2022). Null effects of news exposure: A test of the (un) desirable effects of a 'news vacation'and 'news binging'. *Humanities and Social Sciences Communications*, *9*(1), 1–10.

Wright, M. N., & Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*.

Wu, A. X., Taneja, H., & Webster, J. G. (2021). Going with the flow: Nudging attention online. *New Media & Society*, *23*(10), 2979–2998.

Wu, F., Wilkinson, D. M., & Huberman, B. A. (2009). Feedback loops of attention in peer production. *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, 409–415.

Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A Large Language Model for Finance.

Xu, W. W., Sang, Y., & Kim, C. (2020). What Drives Hyper-Partisan News Sharing: Exploring the Role of Source, Style, and Content. *Digital Journalism*, *8*(4), 486–505.

Zamith, R. (2018). Quantified Audiences in News Production: A synthesis and research agenda. *Digital Journalism*, *6*(4), 418–435.

Zhao, Q., Erdogdu, M. A., He, H. Y., Rajaraman, A., & Leskovec, J. (2015). SEISMIC: A Self-Exciting Point Process Model for Predicting Tweet Popularity. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1513–1522.

Zhou, F., Xu, X., Trajcevski, G., & Zhang, K. (2021). A survey of information cascade analysis: Models, predictions, and recent advances. *ACM Computing Surveys (CSUR)*, *54*(2), 1–36.

Zillmann, D., Chen, L., Knobloch, S., & Callison, C. (2004). Effects of Lead Framing on Selective Exposure to Internet News Reports. *Communication Research*, 24.

Zweig, K. A., & Kaufmann, M. (2011). A systematic approach to the one-mode projection of bipartite graphs. *Social Network Analysis and Mining*, *1*(3), 187–218.

# APPENDIX A

# STUDY 1 SUPPLEMENT

This appendix presents supplementary analysis and data for Chapter 2.

## A.1 Network robustness

### A.1.1 Z-score threshold

We recognize that there are potential sensitivities in our analysis to the Z-score threshold we have chosen. In particular, it is possible that edges were barely excluded at the $Z > 1.96$ threshold that would change some of the clustering dynamic we observe, or that edges crucial to the structure of the clusters would disappear at a stricter threshold. To account for these scenarios, we construct additional networks with edges selected by varying Z-scores. In addition to the network we present in the main text, which filters out edges based on a $Z > 1.96$ threshold, we also check 1.64, 2.58, and 3.29. No additional analyses break down the clustering dynamic we observe in our initial network. We present the underlying values for each network edge in Table A.1.

At $Z > 1.64$, one edge gets added to the left-leaning cluster (between *The Guardian* and *Vox*). Neither the clusters themselves or the underlying structures–a dense right-leaning cluster versus a more chain-like left-leaning cluster–are disturbed.

At $Z > 2.58$, eight of the 24 significant edges in our initial projection are lost. Four of these are negative edges across clusters, and three are in the left-leaning cluster. This means that the structure of the left leaning cluster becomes weaker, but, while the two clusters no longer have a significant negative association, they are still not connected.

At $Z > 3.29$, eight additional edges are lost. Once again, we see the weakening of the left-leaning cluster, as well as the loss of negative edges between clusters. Strikingly, the right-leaning cluster remains almost entirely intact, even at this level.

| From | To | Shared journalists | p | Z-score |
|---|---|---|---|---|
| NationalReview | NewYorkPost | 43 | 0.00 | 9.26 |
| NationalReview | WashingtonPost | 24 | 0.86 | -0.18 |
| NationalReview | Atlantic | 6 | 0.01 | -2.50 |
| NationalReview | BuzzfeedNews | 1 | 0.25 | -1.14 |
| NationalReview | Breitbart | 18 | 0.00 | 7.41 |
| NationalReview | FoxNews | 13 | 0.00 | 3.52 |
| NationalReview | Guardian | 2 | 0.00 | -3.07 |
| NationalReview | Vox | 1 | 0.00 | -2.88 |
| NewYorkPost | NewYorkTimes | 2 | 0.01 | -2.57 |
| NewYorkPost | WashingtonPost | 16 | 0.04 | -2.08 |
| NewYorkPost | Atlantic | 2 | 0.00 | -3.60 |
| NewYorkPost | BuzzfeedNews | 1 | 0.30 | -1.03 |
| NewYorkPost | Breitbart | 8 | 0.03 | 2.24 |
| NewYorkPost | FoxNews | 22 | 0.00 | 8.23 |
| NewYorkPost | Guardian | 4 | 0.01 | -2.48 |
| NewYorkPost | NPR | 1 | 0.00 | -3.20 |
| NewYorkTimes | WashingtonPost | 17 | 0.54 | 0.61 |
| NewYorkTimes | Atlantic | 10 | 0.57 | 0.57 |
| NewYorkTimes | BuzzfeedNews | 2 | 0.78 | 0.28 |
| NewYorkTimes | Guardian | 16 | 0.00 | 4.15 |
| NewYorkTimes | NPR | 13 | 0.00 | 2.89 |
| NewYorkTimes | Vox | 4 | 0.56 | -0.58 |

| | | | | |
|---|---|---|---|---|
| WashingtonPost | Atlantic | 28 | 0.38 | 0.88 |
| WashingtonPost | BuzzfeedNews | 3 | 0.44 | -0.78 |
| WashingtonPost | Breitbart | 5 | 0.18 | -1.33 |
| WashingtonPost | FoxNews | 2 | 0.01 | -2.79 |
| WashingtonPost | Guardian | 17 | 0.97 | -0.04 |
| WashingtonPost | NPR | 20 | 0.37 | 0.89 |
| WashingtonPost | Vox | 24 | 0.00 | 3.08 |
| WashingtonPost | CNN | 8 | 0.01 | 2.45 |
| WashingtonPost | TalkingPointsMemo | 1 | 0.34 | 0.95 |
| Atlantic | BuzzfeedNews | 3 | 0.74 | 0.33 |
| Atlantic | Guardian | 15 | 0.05 | 1.97 |
| Atlantic | NPR | 21 | 0.00 | 4.16 |
| Atlantic | Vox | 16 | 0.00 | 3.28 |
| Atlantic | CNN | 3 | 0.35 | 0.94 |
| BuzzfeedNews | Guardian | 4 | 0.10 | 1.64 |
| BuzzfeedNews | NPR | 1 | 0.49 | -0.69 |
| BuzzfeedNews | Vox | 1 | 0.67 | -0.43 |
| BuzzfeedNews | CNN | 2 | 0.01 | 2.77 |
| Breitbart | FoxNews | 6 | 0.00 | 3.67 |
| Guardian | NPR | 8 | 0.81 | 0.24 |
| Guardian | Vox | 10 | 0.06 | 1.90 |
| NPR | Vox | 11 | 0.02 | 2.32 |
| NPR | CNN | 1 | 0.73 | -0.35 |
| Vox | CNN | 1 | 0.93 | -0.09 |

| | | | | |
|---|---|---|---|---|
| CNN | TalkingPointsMemo | 1 | 0.00 | 4.98 |

Table A.1: Underlying frequencies of shared contributors between news outlets and p-values/Z-scores generated by SICOP

### A.1.2 Outlets included

Because our data are collected secondhand, they are subject to a couple potential biases that might affect our results. First, the selection of outlets within the sample may shape the network structure. Second, unobserved inconsistencies in data collection across outlets could distort the contributor publishing histories we use to evaluate edge significance. To better understand the potential impact of these factors, we removed one outlet at a time from our sample, then ran SICOP again on each set of 12 remaining outlets. Rather than just removing a node from our final network, this procedure allows new edge weights to be calculated without the influence of a particular outlet.

In no case did the overall network structure we observe drastically change as a result of this procedure. We still see a loose collection of left-/center-leaning outlets, and a dense cluster of right-leaning ones. In some cases, the left/center cluster breaks into multiple clusters, or into one cluster with isolates (e.g., Fig. A.1 a). This follows naturally from the cluster's observed chain-like structure—removing central nodes causes the chain to break. However, none of these iterations change the looseness of the overall structure. Similarly, in the right-leaning cluster, removing any one of the four nodes simply causes edges to form among the other three (e.g., Fig. A.1b). Most importantly, in no iteration do any significant edges form between clusters. Thus, while the particular structure of the left/center cluster does show some sensitivity to the outlets included, the overall division and the characteristics of the clusters within this network remain consistent.


(a) (b)

Figure A.1: Outlet-outlet projections with a) NPR removed and b) the New York Post removed. These cases are representative of the minor changes in network structure created by dropping outlets from our sample, maintaining the characteristics we highlight in our findings.

## A.2 LIWC

Here we report the full list of LIWC (and VADER) features compared between the two primary groups of articles. The entries are sorted by AUC.

| Feature | M-W U | P-val. | AUC | CntrLeft Avg. | Right Avg. | Diff. | Bonf. P-val. |
|---|---|---|---|---|---|---|---|
| hear | 3144691.0 | 0.000 | 0.652 | 1.08 | 0.71 | 0.37 | 0.000 |
| affect | 3242798.0 | 0.000 | 0.641 | 4.33 | 5.10 | -0.77 | 0.000 |
| percept | 3260857.5 | 0.000 | 0.639 | 2.38 | 1.87 | 0.51 | 0.000 |
| negemo | 3438952.5 | 0.000 | 0.619 | 1.88 | 2.36 | -0.48 | 0.000 |
| certain | 3476109.5 | 0.000 | 0.615 | 1.06 | 1.30 | -0.24 | 0.000 |
| focuspast | 3616885.0 | 0.000 | 0.600 | 4.09 | 3.45 | 0.64 | 0.000 |
| anger | 3625115.0 | 0.000 | 0.599 | 0.69 | 0.91 | -0.22 | 0.000 |
| relativ | 3630524.5 | 0.000 | 0.598 | 13.63 | 12.72 | 0.91 | 0.000 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| negate | 3656126.5 | 0.000 | 0.595 | 1.16 | 1.36 | -0.20 | 0.000 |
| posemo | 3780913.0 | 0.000 | 0.582 | 2.37 | 2.67 | -0.30 | 0.000 |
| QMark | 3797283.5 | 0.000 | 0.580 | 0.18 | 0.25 | -0.07 | 0.000 |
| space | 3797893.0 | 0.000 | 0.580 | 7.28 | 6.82 | 0.46 | 0.000 |
| sad | 3818021.5 | 0.000 | 0.577 | 0.29 | 0.37 | -0.08 | 0.000 |
| discrep | 3839167.0 | 0.000 | 0.575 | 1.19 | 1.37 | -0.18 | 0.000 |
| differ | 3877426.5 | 0.000 | 0.571 | 2.73 | 2.97 | -0.24 | 0.000 |
| Comma | 3878505.0 | 0.000 | 0.571 | 5.73 | 5.39 | 0.34 | 0.000 |
| risk | 3897427.0 | 0.000 | 0.569 | 0.71 | 0.82 | -0.11 | 0.000 |
| prep | 3902945.5 | 0.000 | 0.568 | 14.57 | 14.18 | 0.39 | 0.000 |
| conj | 3913954.5 | 0.000 | 0.567 | 5.31 | 5.54 | -0.23 | 0.000 |
| vaderCompound | 3928249.5 | 0.000 | 0.565 | 0.25 | 0.11 | 0.14 | 0.000 |
| power | 3936100.0 | 0.000 | 0.564 | 4.23 | 4.55 | -0.32 | 0.000 |
| cogproc | 3950743.5 | 0.000 | 0.563 | 9.58 | 10.09 | -0.51 | 0.000 |
| time | 3982288.5 | 0.000 | 0.559 | 4.69 | 4.38 | 0.31 | 0.000 |
| motion | 3997062.0 | 0.000 | 0.558 | 1.74 | 1.60 | 0.14 | 0.000 |
| auxverb | 4016272.5 | 0.000 | 0.555 | 6.77 | 7.05 | -0.28 | 0.000 |
| see | 4063008.0 | 0.000 | 0.550 | 0.87 | 0.77 | 0.10 | 0.000 |
| Quote | 4110413.5 | 0.000 | 0.545 | 2.48 | 2.28 | 0.20 | 0.000 |
| drives | 4116632.5 | 0.000 | 0.544 | 8.46 | 8.79 | -0.33 | 0.000 |
| focuspresent | 4132848.0 | 0.000 | 0.543 | 7.19 | 7.46 | -0.27 | 0.000 |
| assent | 4145328.5 | 0.000 | 0.541 | 0.06 | 0.09 | -0.03 | 0.000 |
| adj | 4152054.5 | 0.000 | 0.540 | 4.77 | 4.94 | -0.17 | 0.000 |
| death | 4158694.5 | 0.000 | 0.540 | 0.24 | 0.27 | -0.03 | 0.000 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| social | 4164598.0 | 0.000 | 0.539 | 8.96 | 8.55 | 0.41 | 0.000 |
| relig | 4164692.0 | 0.000 | 0.539 | 0.33 | 0.41 | -0.08 | 0.000 |
| i | 4164564.0 | 0.000 | 0.539 | 0.71 | 0.53 | 0.18 | 0.000 |
| tbPolarity | 4178439.5 | 0.000 | 0.538 | 0.09 | 0.08 | 0.01 | 0.000 |
| reward | 4181480.0 | 0.000 | 0.537 | 1.00 | 1.07 | -0.07 | 0.000 |
| work | 4188183.5 | 0.000 | 0.536 | 4.60 | 4.31 | 0.29 | 0.000 |
| home | 4195092.5 | 0.000 | 0.536 | 0.40 | 0.33 | 0.07 | 0.000 |
| Apostro | 4190130.0 | 0.000 | 0.536 | 2.22 | 2.31 | -0.09 | 0.000 |
| SemiC | 4208843.5 | 0.000 | 0.534 | 0.03 | 0.02 | 0.01 | 0.000 |
| they | 4219078.0 | 0.000 | 0.533 | 0.86 | 0.95 | -0.09 | 0.000 |
| money | 4232751.5 | 0.000 | 0.532 | 1.12 | 1.23 | -0.11 | 0.001 |
| quant | 4235906.0 | 0.000 | 0.531 | 2.16 | 2.24 | -0.08 | 0.001 |
| adverb | 4241383.5 | 0.000 | 0.531 | 3.54 | 3.64 | -0.10 | 0.002 |
| ingest | 4234582.0 | 0.000 | 0.531 | 0.26 | 0.19 | 0.07 | 0.000 |
| netspeak | 4249819.0 | 0.000 | 0.530 | 0.21 | 0.18 | 0.03 | 0.001 |
| interrog | 4244529.0 | 0.000 | 0.530 | 1.44 | 1.39 | 0.05 | 0.002 |
| AllPunc | 4255325.5 | 0.000 | 0.529 | 17.26 | 16.84 | 0.42 | 0.004 |
| OtherP | 4258541.5 | 0.000 | 0.529 | 0.28 | 0.30 | -0.02 | 0.002 |
| leisure | 4274775.5 | 0.000 | 0.527 | 1.03 | 0.90 | 0.13 | 0.013 |
| function | 4314285.5 | 0.001 | 0.523 | 46.19 | 46.21 | -0.02 | 0.110 |
| tentat | 4330995.5 | 0.003 | 0.521 | 2.18 | 2.25 | -0.07 | 0.241 |
| anx | 4336993.5 | 0.004 | 0.520 | 0.36 | 0.39 | -0.03 | 0.306 |
| verb | 4352186.5 | 0.007 | 0.518 | 12.57 | 12.28 | 0.29 | 0.606 |
| affiliation | 4354677.0 | 0.008 | 0.518 | 1.86 | 1.81 | 0.05 | 0.671 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| achieve | 4353813.5 | 0.008 | 0.518 | 1.65 | 1.70 | -0.05 | 0.648 |
| bio | 4350804.0 | 0.007 | 0.518 | 1.29 | 1.04 | 0.25 | 0.571 |
| feel | 4360879.0 | 0.010 | 0.517 | 0.32 | 0.29 | 0.03 | 0.835 |
| article | 4366549.5 | 0.012 | 0.517 | 8.54 | 8.42 | 0.12 | 1.074 |
| informal | 4368808.5 | 0.013 | 0.516 | 0.38 | 0.38 | 0.00 | 1.152 |
| shehe | 4373985.5 | 0.017 | 0.516 | 1.89 | 1.77 | 0.12 | 1.420 |
| we | 4393997.0 | 0.033 | 0.514 | 0.59 | 0.63 | -0.04 | 2.829 |
| sexual | 4407971.5 | 0.024 | 0.512 | 0.11 | 0.13 | -0.02 | 2.072 |
| Period | 4411906.5 | 0.058 | 0.512 | 5.36 | 5.35 | 0.01 | 5.019 |
| number | 4406451.0 | 0.050 | 0.512 | 2.05 | 2.36 | -0.31 | 4.258 |
| swear | 4411774.5 | 0.008 | 0.512 | 0.03 | 0.03 | 0.00 | 0.656 |
| nonflu | 4405019.5 | 0.031 | 0.512 | 0.08 | 0.08 | 0.00 | 2.694 |
| focusfuture | 4421046.5 | 0.076 | 0.511 | 1.00 | 1.01 | -0.01 | 6.530 |
| cause | 4422613.0 | 0.079 | 0.511 | 1.60 | 1.62 | -0.02 | 6.820 |
| Colon | 4426143.5 | 0.086 | 0.510 | 0.40 | 0.36 | 0.04 | 7.421 |
| compare | 4457029.5 | 0.184 | 0.507 | 2.65 | 2.66 | -0.01 | 15.855 |
| female | 4451589.0 | 0.156 | 0.507 | 0.69 | 0.63 | 0.06 | 13.447 |
| health | 4456542.5 | 0.181 | 0.507 | 0.65 | 0.47 | 0.18 | 15.608 |
| friend | 4467566.0 | 0.220 | 0.506 | 0.16 | 0.17 | -0.01 | 18.889 |
| Parenth | 4464664.0 | 0.212 | 0.506 | 0.56 | 0.52 | 0.04 | 18.200 |
| Exclam | 4460400.0 | 0.098 | 0.506 | 0.04 | 0.05 | -0.01 | 8.416 |
| ipron | 4476870.0 | 0.273 | 0.505 | 4.74 | 4.71 | 0.03 | 23.461 |
| male | 4498125.5 | 0.386 | 0.502 | 1.63 | 1.63 | 0.00 | 33.224 |
| family | 4507666.5 | 0.438 | 0.501 | 0.27 | 0.29 | -0.02 | 37.705 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| you | 4510550.5 | 0.458 | 0.501 | 0.48 | 0.48 | 0.00 | 39.401 |
| insight | 4508382.5 | 0.446 | 0.501 | 1.96 | 1.95 | 0.01 | 38.328 |
| filler | 4509865.0 | 0.376 | 0.501 | 0.01 | 0.01 | 0.00 | 32.359 |
| pronoun | 4507193.5 | 0.439 | 0.501 | 9.27 | 9.08 | 0.19 | 37.729 |
| ppron | 4517453.5 | 0.499 | 0.500 | 4.53 | 4.37 | 0.16 | 42.933 |
| body | 4515469.5 | 0.487 | 0.500 | 0.30 | 0.28 | 0.02 | 41.912 |

# APPENDIX B

# STUDY 2 SUPPLEMENT

This appendix presents supplementary analysis and visualizations for Chapter 3. First, we provide a detailed overview of our sample selection process, as well as the rules and behaviors common in each subreddit. We also detail additional descriptive analyses that examine the concentration and stability within the subreddits we examine. These analyses further confirm our primary results: While we find evidence of drastic concentration, there is little support for the idea that news subreddits are a stable attention system. Finally, we present the results of robustness checks designed to account for potential cross-subreddit behavior differences, and to examine submission behavior around breaking news events.

## B.1 Subreddit Details

This section details the selection process we used to build our sample of news-focused subreddits. It also provides an overview of the governing rules and types of content that characterize each community.

### B.1.1 Selection Process

Our analyses focus on r/news, r/worldnews, and r/politics as the largest news-focused communities on Reddit. Broadly, we focused on large communities as a selection criterion for a couple reasons. First, because each community acts as a self-contained attention market, with its own rules, norms, membership, and set of submissions, sampling at the community level provides the closest fit to the structure of the platform. Second, since large communities command the most activity, they have the highest potential to ultimately impact attention to news sources. More specifically, we reach these three communities via a manual filtering procedure. Reddit's API provides an endpoint that returns a list of up to 100 of the most popular subreddits, as sorted by activity. We queried this endpoint in April 2019. We then selected subreddits from that list of 100 with two criteria. First, to

capture content from outside news sources, we included subreddits that only accepted link-based submissions (as opposed to text or multimedia). Second, we removed any subreddits that did not focus on news content, defined as those that consisted of links to news articles. These steps left us with the three subreddits in our final sample.

### B.1.2 Community Characteristics

Each of the subreddits in our sample focuses on external news content, in the form of news article hyperlinks submitted by Reddit users. Within that broad commonality, each community has a specific focus and set of rules that distinguishes it from the others. Here we detail the high-level characteristics of each subreddit. We also account for the potential impacts of these community-specific rules and behaviors with additional analyses, the details of which can be found in the Robustness Checks section of this supplement.

R/worldnews, as the name implies, prioritizes a global perspective in its submissions. This means the subreddit does not allow stories that only concern the U.S. All topics—such as politics, business, and social issues—are allowed, if the news story being shared is either international in scope or primarily concerns another country. R/worldnews is the largest subreddit in our sample, with over 25 million members. Its most frequent sources for submissions within our sample were Reuters, YouTube, and The Guardian.

R/news does not have a strict geographic focus. It allows stories concerning both the U.S. and other countries, across a broad range of topics. The subreddit discourages "soft news", such as celebrity updates or jokes. It is the second largest subreddit in our sample, with over 22 million members. Its most-used sources in our sample were CNBC, iCrowdNewswire, and Twitter.

R/politics solely focuses on U.S. politics. While it is still entirely comprised of articles from external news sources, this subreddit is more permissive of opinion-oriented content than the other

two. However, r/politics is the most restrictive in terms of the news sources it allows, as it maintains a list of approved domains from which users may submit stories . It is also by far the smallest subreddit in our sample, with over 7 million members. Its most-used sources in our sample were the Washington Post, the Hill, and the New York Times.

Finally, there are several important commonalities in the rules these subreddits enforce. All three stress the importance of recency—r/worldnews and r/poltics have rules against posting old articles, while r/news blocks the re-posting of already-submitted stories. All three also, either explicitly in the case of r/worldnews or implicitly, prioritize or require English-language content. R/worldnews and r/news do not allow editorial or opinion pieces, while r/news and r/politics have a rule against posting stories from sites with paywalls. Each subreddit takes a somewhat distinct approach to governing the construction of a shared news information resource. As our results show, though, the mechanisms governing attention activity within those resources remain constant.

## B.2 Additional Descriptive Analyses

Hindman and Rogers (2018) begin with a thorough descriptive examination of their data, demonstrating the stability afforded to top sites before attempting to simulate traffic growth. To better interpret any departures from expectation in our simulation, we replicate or adapt their most consequential descriptive analyses before testing the model.

### B.2.1 Log-log plot

First, we plot rank and mean points per submission on a log-log plot (Fig. 1). In line with the concentration indicated by the Gini coefficient, we expect a relatively straight line along the diagonal. Instead, we see large departures from the diagonal, particularly at top ranks. This indicates that, by rank, top sites deviate from a highly concentrated distribution (e.g., log-normal or power law).

### B.2.2 Leakage

Second, we examine a metric Hindman and Rogers (2018) term "leakage". We calculate the percent of total days (out of 273 possible) that any site appears in each subreddit (Fig. 2). Hindman and Rogers (2018) observe an upward slope in leakage by rank, indicating that higher-ranked sites are less likely to drop out. In contrast, our measure shows that most sites are only present for a handful of days—67% of sites are present in a subreddit five days or fewer, and only 46 sites are present on all days.

### B.2.3 Rank Swaps and Rank Occupants

We next conduct two related analyses intended to demonstrate the rank stability of high-performing sites. We calculate 1) the number of times the site occupying a rank changes and 2) the number of unique sites that occupy each rank. The first measure captures cases in which a small number of sites constantly swap ranks (e.g., if nytimes.com and washingtonpost.com are fighting for the top rank), while the second measure captures cases in which many sites occupy a rank at different points in time. In Hindman and Rogers (2018), both measures increase with rank. However, in our data, we do not observe this dynamic. Instead, we see an extremely high level of swaps (Fig. 3) and occupying sites (Fig. 4) at every rank. Even top ranks change hands every day, and a new site occupies the top spot almost every day. While both measures surge in lower rankings, this is more a function of ties (e.g., many sites getting one or two points per submission) than increased volatility. These measures provide no evidence for stability at top ranks.

### B.2.4 Median Rank Gap

To examine how point allocation is distributed across the full distribution, we calculate the median difference in log-transformed performance between sites at each adjacent rank (e.g., the difference

in performance between ranks 1 and 2, and between 2 and 3). As expected, this gap declines rapidly after the first few ranks, and generally continues declining (Fig. 5). This analysis indicates closer proximity in performance at lower ranks, which could lead to increased volatility.

### B.2.5 Growth Rate and Site Size

We then turn to growth rate as a function of site size. To do so, we examine the correlation between the absolute value of sites' mean daily performance growth and the total number of posts from each site in a subreddit. We expect a negative correlation. This would indicate that larger (more frequent) sites experience smaller daily variation in performance, while smaller (less frequent) sites are volatile. Our correlations follow this expectation: $r_{news} = -0.15$, $r_{politics} = -0.18$, and $r_{worldnews} = -0.13$.

Finally, we examine site performance change from first to last appearance. We expect to see a funnel-shaped distribution—sites that start with lower performance display greater variance later, while those with high performance maintain it throughout. While low-performing sites are volatile, high-performing ones also follow no clear pattern (fig. 6).

### B.2.6 Summary

In summary, we adapted the descriptive analyses utilized by Hindman and Rogers (2018) in examining web traffic and find mixed agreement in news subreddits (Table 1). Some analyses, namely median performance gaps and size/growth volatility correlation, produce expected results. However, those that examine the full distribution of sites and their relative performance do not. Attention allocation to news sources on Reddit appears to be more volatile and less stable at top ranks than that to news websites.

## B.3 Robustness Checks

While our results are extremely similar across subreddits, it is worth examining whether rules or behaviors specific to any one community might affect the outcomes of our analyses. To do so, we conducted a series of additional checks. First, we measured submission-level score trends, to check for any unexpected behavior that might influence our simulation. Second, we examined the extent to which subreddit-level domain restrictions were enforced. To check for differences in user base preferences, we looked for sizable differences in the performance of domains across all three subreddits. Finally, to account for breaking news situations in which many users might submit similar stories in rapid succession, we looked for signs of bursty submission behavior. None of these analyses produced results that would materially impact our findings.

### B.3.1 Submission Behavior

Because our simulation occurs at the aggregated level of domain performance, it may be influenced by unobserved behavior in individual submissions. We examine submission-level voting trends to determine whether Reddit posts deviate from expectations based on prior work. Across many platforms, user-submitted content displays a largely consistent pattern of bursty attention, in which engagement to individual posts rapidly decays (Ratkiewicz, Fortunato, Flammini, Menczer, & Vespignani, 2010; Szabo & Huberman, 2010).

For this analysis, we collected data for all r/news posts submitted in April 2020. Each submission's score was queried every 10 minutes. We only considered submissions for which we collected at least an hour of observations (giving us enough data to measure their trend over time), and which ultimately reached at least 100 points (filtering out noisy posts with little engagement). This process gave us 96,968 observations of 1,160 unique submissions.

We normalized each post's hourly performance by dividing its score at each interval by its maximum score. We then represented the general performance trend over time for each quintile, with 95% confidence intervals, using a generalized additive model (GAM). A GAM allows us to represent each quintile's growth pattern as a smoothed trend, in a nonlinear fashion, estimated directly from the data. We divided submissions into quintiles by their final score, allowing us to examine differences in how quickly votes accumulate to posts of varying popularity.

Submissions attract the vast majority of attention within 24 hours of posting. Figure 7 shows the GAM-estimated cumulative score trend for each r/news submission quintile over the first 72 hours after posting. The consistency across quintiles is striking. After one day, voting activity drops off regardless of the absolute magnitude of a submission's final score. Such a stable pattern of score accumulation suggests that, regardless of size, news submissions accrue attention following patterns consistent with our expectations. This suggests that the empirical deviations from cumulative advantage that we observe arise not from submission-level voting, but from higher-order aggregate behavior.

### B.3.2 Domain Restrictions

As discussed above, r/news disallows submissions from paywalled news sources (e.g., the Wall Street Journal), and r/politics maintains a whitelist of allowed domains. These policies might influence the concentration of submission across sources, either by disallowing credible news sources and forcing users to find alternatives, or by filtering out any source the moderators do not deem credible.

To measure whitelist compliance in r/politics, we parsed a list of approved domains from an archived version of the whitelist published during our sample's timeframe. We found 7,504 domains in our sample that did not appear in the whitelist. However, these domains did perform worse

than those that were whitelisted—they had a median submission score of 1, while whitelisted domain submissions had a median score of 22. The paywall restriction in r/news appears to have had a similar effect. Since there is not a comprehensive list of restricted news sources with paywalls in the subreddit, we searched for submissions from the New York Times and the Wall Street Journal. Both sources are prominent news outlets with paywalls of varying strictness. Again, we find extensive submissions from both domains, but both have a median score of 1.

It appears in both cases that, while users are not prohibited from submitting stores that go against the rules, those stories are either removed after submission or otherwise depressed. While this dynamic impacts which news sources can perform well within the subreddits, it does not impact our results. If anything, the favoring of certain news sources over others should increase stability, as it provides a more predictable set of domains from which news content gets sourced. However, we see no evidence that these rules impact popularity stability.

### B.3.3 Cross-Subreddit Reception

We also checked for differences in performance for the same site across multiple subreddits. This variance matters for a couple reasons. First, if these communities have similar receptions to content from the same domains, then the consistency of our findings across subreddits becomes less surprising. Variance across subreddits also lends credence to the idea that there is community-level randomness driving popularity dynamics.

In accordance with this view, we find evidence of substantial variance. We first identified domains that appeared in all three subreddits, 2,543 in total. We then measured their mean performance per submission within each subreddit and examined the variance across all three. There was a median difference of 48 points between the highest- and lowest-scoring subreddits. Since the median points per submission for these domains ranges from 1 to 6 across subreddits, this

difference represents a drastic swing for most sites. This difference also displays variance across domains, ranging from no difference between minimum and maximum score, to a 46,223-point difference. Reception therefore appears to vary across subreddits at the domain level, making the consistency of our findings even more surprising.

### B.3.4 Competing Submissions

Finally, we looked for cases in which multiple users rapidly submitted stories on the same news event to the same subreddit. This pile-on dynamic could contribute to the randomness we observe in attention allocation, as users upvote the first story they see pertaining to breaking news events regardless of source. Conversely, if submissions around major events only ever come from authoritative news sources, voting behavior might be more consistent with the concentration story of cumulative advantage. To look for this behavior, we focused on the news subreddit. We calculated the time difference between every submission in our sample in sequence, log transformed it, then calculated a 3-post rolling mean for this delta. We then constructed a set of submissions that appeared in sequence, and for which the time between submissions was less than the 25th percentile delta. This gave us a set of submission sequences that occurred in rapid succession. We then manually examined these sequences, to determine if they referred to the same news event.

Based on a sample of 300 of these headline sequences, bursts of submissions appear extremely rare. Only two occurred—one after the death of the Mormon church president, and one after the banning of a Pennsylvania fraternity. In most cases it appears that, either via submission behavior or moderation decisions, attention focuses on single submissions for major news stories. This dynamic is not inconsistent with our results—attention can still focus on sources at random—but it is surprising that this behavior does not produce more consistent outcomes. Major news outlets tend to have the most resources to cover breaking news, making it seem likely that their stories

would consistently be high performers in response to large news events. We do not see evidence of this dynamic, though.

# APPENDIX C

# STUDY 3 SUPPLEMENT

Here we share all dictionaries used in our feature engineering pipeline.

### C.1 Shareability

will make you this is why can we guess only [0-9]+ in the reason is are freaking out [0-9]+ stunning photos tears of joy is what happens make you cry give you goosebumps talking about it is too cute shocked to see melt your heart [0-9]+ things only can't stop laughing top [0-9]+ songs twitter reacts to what happened next [0-9]+ reasons why [0-9]+ things you this is what this is the this is how [0-9]+ of the [0-9]+ ways to the [0-9]+ best how to make these are the here are the how to get [0-9]+ things that [0-9]+ things to you can now the [0-9]+ most [0-9]+ things only why you should the world [0-9]+ years goes viral to know [0-9]+ days on twitter are you right now can you on instagram first time the internet all time your life is epic [0-9]+ minutes study finds on facebook regrets it your heart

### C.2 A/V - manual

watch listen video audio clip image slideshow gallery interactive graphic nsfw

### C.3 LIWC - see

beaut* black blacke* blackish* blacks blind* blond* blue* bright* brown* candle* circle click* color* colour* column* cream eye* eying gaz* glanc* glow* gray* green* grey* image* lit look looked looker* looking looks orange* picture pink* purpl* rectang* red redde* reddish* redness reds round* saw scan scann* scans screen see seeing seen seer sees shine shini* shiny sight* squar* stare* staring sunli* sunshin* triang* view viewer* viewing* views vivid* watch* white* whitish* yellow*

## C.4 LIWC - perception

acid* acrid* aroma* audibl* audio* beaut* bitter* black blacke* blackish* blacks blind* blond* blue* boom* bright* brown* brush* butter* candle* caramel* caress* chocolate* choir* circle citrus* click* cold* cologne* color* colour* column* concert* cool* cream deaf* delectabl* delicious* deoder* drie* drily drool* dry* ear ears edge edges edging experienc* eye* eying feel feeling* feels felt fetid* finger* fire fizz* flavor* flavour* flexib* fragil* fragran* freez* froze* fruit* fuzz* gaz* glanc* glow* grab* gray* greas* green* grey* grip gripp* grips hair* hand handful* hands hard harde* harmon* hear heard hearing hears heavie* heavy* honey hot hott* hush* image* inaudibl* inhal* leather* lick* light limp* listen listened listener* listening listens lit look looked looker* looking looks loose* loud* mint* musi* nasal noise noises noisy nose* nostril* odor* odour* oil* orange* palatabl* perfum* picture pink* press pressed presser* presses pungen* purpl* quiet* rancid* rang rectang* red redde* reddish* redness reds reek* ring ringing rings rotten rough* round* rub rubbed rubbing rubs saccharine said saliv* salt* sampl* sand sands sandy sang savor* savour* saw say* scan scann* scans scent* scratch* scream* screen scrumptious* see seeing seen seer sees sharp* shine shini* shiny shout* sight* silen* silk* skin skin'* smell* smooth* sniff* snort* soft* song* sound* sour soure* souri* sours soury speak speaker* speaking speaks speech* spice spiced spices spicy spoke* squar* squeez* stank stare* staring stench* stink* stroke* stroki* stunk sugar* sumptuous* sunli* sunshin* sweet sweetness sweets tang tangy tart tast* thick* thin thinn* thunder* tight* tongue* touch* triang* unsavo* view viewer* viewing* views vivid* voic* waft* warm* watch* weight weighted weighting weightless* weightlift* weights wet wetly whiff* whisper* white* whitish* yell yelled yelling yellow* yells yum*

### C.5 Empath - conflict results

confrontation discord squabbling feuding bickering political_struggle rancor confrontations antagonism recriminations skirmish animosities acrimony squabble ideological_differences squabbles animosity bad_blood conflict clash dissension mistrust antagonisms sniping feud rift maneuvering infighting name-calling political_battle bad_feelings tussle disunity divisiveness political_differences political_conflict rivalries power_struggles power_struggle deep_divisions friction political_maneuvering distrust ill_will stalemate tension schism bitter_feelings disagreement rifts polarization resentments bloodletting open_warfare standoff feuds intransigence battle quarrels internal_conflicts political_debate public_debate political_controversy controversies struggle debate frictions hostility bitter_debate growing_tension rivalry jockeying skirmishes quarrel fracas tug-of-war national_debate militancy current_crisis passionate_debate internal_conflict enmity strife public_outrage long-running_debate dispute bloodshed brinkmanship indecision skirmishing continuing_debate abortion_debate sharp_debate political_crisis fierce_debate political_forces new_debate agitation ethnic_tensions misunderstandings tensions public_anger partisanship internal_struggle furious_debate outside_forces racial_divisions controversy internal_divisions disagreements antagonists backbiting recrimination growing_tensions heated_debate adversaries escalation fight belligerence finger-pointing intense_debate showdown battles factionalism personal_ambitions bitterness contentiousness hatreds internal_disputes opposing_sides pitched_battle political_infighting disputes hostilities bitter_conflict rebellion rebellions conflicts deep_differences political_battles emotional_debate political_compromise constitutional_crisis political_problems battle_lines inertia street_protests political_tensions fierce_struggle political_arena political_posturing vitriol ethnic_conflicts political_divisions political_dispute hysteria resentment turf_battles bitter_dispute direct_confrontation lawlessness more_violence political_opposition turf_wars high_stakes unrest tough_talk

bitter_disputes factions fissures disaffection growing_sense unease posturing violence new_tensions ethnic_divisions policy_debate partisans internal_strife Israeli-Palestinian_conflict political_storm loyalties personal_attacks political_firestorm extremism discontent deep_distrust disenchantment common_enemy inevitability indecisiveness fighting national_crisis anti-Americanism intolerance growing_frustration divide hatred latest_crisis internal_debate backlash racial_tensions demagoguery unpleasantness fierce_battle ensued new_war fisticuffs anger fury rage quagmire fault_lines free-for-all full-scale_war media_frenzy current_debate racial_division culture_wars inaction revolt clashes lively_debate all-out_war civil_unrest passions antipathy political_fallout revulsion wrangling serious_debate political_furor xenophobia confusion furor counterattacks fundamental_differences ethnic_conflict Balkan_war chaos conspiracy_theories sparring long_war deep_resentment policy_differences differing_views election-year_politics basic_issues hard-liners rival_factions

### C.6 Empath - conflict seed terms

conflict disaccord discord discordance discordancy disharmony dissension dissent dissidence dissonance disunion disunity division friction infighting inharmony schism strife variance war warfare clash collision competition contention altercation argument bicker brawl debate disagreement dispute divide fissure falling-out fight hassle jar miff mix-up quarrel row run-in scrap spat squabble tiff wrangle incompatibility incongruence incongruity incongruousness inconsistence inconsistency inconsonance inharmoniousness animosity antagonism antipathy cold war enmity hostility ill will rancor battle clash combat contest dustup fight fracas fray hassle scrap scrimmage scrum scuffle skirmish struggle tussle pitched battle rough-and-tumble battle royal brawl broil donnybrook free-for-all melee mix-up ruckus ruction blows fistfight fisticuffs grapple handgrip punch-out slugfest confrontation duel face-off joust altercation argument contretemps controversy cross fire disagreement dispute falling-out kickup misunderstanding quarrel row spat squabble tangle tiff wrangle

catfight ball game battle combat competition confrontation contention contest dogfight duel face-off grapple match rivalry strife struggle sweepstakes tug-of-war war warfare horse race nail-biter showdown clash collision discord friction argument controversy debate disagreement disputation dispute dissension quarrel row wrangle clash collide disaccord discord jar battle combat engage fight chafe gall grate jangle differ disagree dissent

### C.7 Empath - surprise results

glee frisson wonderment gasp yawn bewilderment dread gallows_humor astonishment delight giggles bravado perplexity fright befuddlement exhilaration elation incredulity bafflement ecstasy puzzlement amazement bemusement fury sarcasm mirth giddiness shudder contentment nonchalance awe exultation disgust hilarity foreboding self-pity indignation ennui shiver raw_emotion craziness gasps tinge dejection emptiness delirium merriment awkwardness silliness rage self-loathing wistfulness revulsion bonhomie yawns chuckles self-doubt guffaws hysteria pomposity angst exasperation pity rapture disbelief smirk ugliness weariness gasp sighs self-consciousness groans startle insouciance paroxysms torpor laughter grimace flashes tentativeness levity fatalism curiosity grins exuberance venom groan pang ripple reverie strangeness sigh other_moments delirious restlessness momentarily longing edginess such_moments defensiveness horror trance despondency shyness whiff shriek derision weirdness incongruity anguish tremble histrionics flicker swagger tedium babble heartbreak snickers bombast shrieks scream queasiness undertone disorientation blackness nervous_energy smugness breathless mischief few_moments vague_sense bluster trembling impatience shock trepidation loathing glances incomprehension grunts heartache melancholy whispers ache muddle exaltation coldness petulance moans despair ardor boredom self-congratulation profundity bliss desperation hysterics flash excitement emotion wince preciousness pathos revelation deja_vu grimaces vanity earnestness tenderness sneer moaning mur-

mur irritation politeness cynicism naivete high_spirits subconscious intimation annoyance commotion ferocity nastiness nothingness chuckle senses mortification sighing single_moment conflicting_emotions sinking_feeling black_humor righteous_indignation crankiness helplessness adulation whole_scene lunacy self-importance wisecrack coolness murmurs wisecracks ebullience pique teasing nervous_laughter claustrophobia big_moment foolishness self-deprecation cringing verbiage condescension dullness solemnity self-absorption high_drama uncertainly cuteness stoicism murk sentimentality squeals jealousy giggle melancholia slightest outbursts gloom paranoia bathos self-satisfaction wonder slumber bluntness cockiness certain_sense cleverness frivolity many_moments fickleness instinctively enchantment fleetingly crudeness froth vulgarity frown sensations jubilation meanness cheer tingle wallow doom hissing introspection aimlessness joy snarl abandon stridency alacrity hopefulness banality sense bleakness jolt fierceness calmness howl murmurs quivering cacophony faint exclamations behold feeling torment desolation self-righteousness body_language recoil scorn wince catharsis gall good_cheer furies pose whole_affair romanticism intimations very_presence smiles cliches gaiety playfulness ordinariness twitches monotony passivity fleeting_moment imperfection agony brief_moment revel jarring grandiosity languor longings flinching irreverence weeping pratfalls glow shrugs bile emotional_response murmuring marvel impending_doom swoon wailing adoration irony coarseness stares bad_dream goofiness wildness scowl sparks malevolence sensation seethe spunk whisper stasis madness melancholy flutter whimsy rare_moments stings premonition hubris trembling own_emotions clumsiness glints self-assurance apprehension stare wisp brashness thought antics brood undertow exuding lassitude contorted creeps cliche whining moan teasing reveries envy utter final_image audibly dark_humor naivete hisses vividness shriek facial_expression wallowing split_second blandness emotionalism relentlessness lamentation dizzy foreboding tantrums exclamation wry_humor eccentricities heaviness ridicule silences Inevitably flickers self-mockery glimmer whirl egotism in-

evitability pettiness exult theatrics glint unreality eccentricity caldron stillness smirks pangs groaning menace hubbub cocoon dazzle nakedness hypnotized messiness shocking neuroses radiance mistaking instant punch_lines diffidence curses strong_emotions discouragement cringe extremity nostalgia same_sense twitching aloofness obviousness groan portents banter muttering grief dreariness roughness weep miasma hollowness mere_mention numbed chatter even_a_hint whine slightest_hint disconcerting yearning prurience fascination hyperbole adrenaline_rush thrusts blankness hiss halo clich0̆0e9 nervousness itch smidgen sob vibrates agitation starkness sparkle exhilarated grin audacity vortex discomfort uneasy_feeling pretentiousness willfulness hauteur goose_bumps rawness consternation ambivalence frenzy real_drama human_dimension tempest poignancy

### C.8 Empath - surprise seed terms

surprise bombshell jar jaw-dropper jolt stunner shock thunderclap eye-opener revelation shocker amazement marvel wonder fillip kick kicker twist wrinkle amazement astonishment shock startlement stupefaction awe wonder wonderment startle bewilderment confusion consternation discomfiture dismay amaze astonish astound bowl over dumbfound flabbergast floor rock shock startle stun stupefy thunderstrike befuddle bewilder blindside blow away confound confuse daze discomfit disconcert dismay jar muddle nonplus perplex shake up