Mapping U.S. Talk Radio: A Textual Survey at Scale

by

William Brannon

B.S., College of William & Mary (2011)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2020

Author .....
Program in Media Arts and Sciences
August 17, 2020

Certified by .....
Deb Roy
Professor of Media Arts and Sciences
Thesis Supervisor

Accepted by .....
Tod Machover
Academic Head, Program in Media Arts and Sciences

Mapping U.S. Talk Radio: A Textual Survey at Scale

by

William Brannon

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 17, 2020, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

This thesis attempts the first large-scale mapping of American talk radio, leveraging recently developed datasets of transcribed radio programs. We set out to explore the internal structure of this influential medium along three axes, reflecting a typology of the main social contexts in which it is embedded: its corporate ownership, its geographical location in the country and perhaps most importantly its relationship to the broader media ecosystem, operationalized through Twitter.

The results depict a radio ecosystem operating in a mostly centralized way. In talk radio, station ownership, usually by large publicly traded companies, is the strongest correlate of cosine similarity in the stations' distributions of airtime to shows; in public radio, the greater similarity of these airtime distributions medium-wide than in talk radio reflects the influence of centralized, nationwide syndication networks like NPR. The distribution of important topics of political discussion is also surprisingly similar across stations. Geography plays relatively little role, with programming and topics varying little along geographic lines. Despite these centralizing tendencies, local radio is not extinct even on large corporate stations, and is meaningfully local in content as well as production. Local programs have lower average cosine similarities between their topic mixtures than syndicated ones do, demonstrating a greater diversity of discussion topics and perhaps perspectives. These shows are also more in touch with local opinion, in that models trained on their text better predict the partisan lean of their listeners than is true of syndicated radio. But syndication makes up the large majority of stations' airtime.

Moving to the third of our three axes, radio reflects the same underlying social structure as Twitter, and this structure is reflected in the content of broadcasts. We examined the relationship by comparing radio to a set of highly followed and influential journalists and politicians. The influential Twitter users manifest a similar social structure to radio: graph communities and a measure of latent ideology extracted from the follow graph fit radio's offline structure well; more directly, the follow graph itself and the "co-airing graph" (between shows, in which two shows are connected if they air on the same station) are quite similar. Moreover, this common social structure is predictable from the text of radio shows. The contents of Twitter and radio are also closely linked, with hosts discussing a similar mix of topics on both platforms. A case study of President Trump's tweets reveals their probable causal effects on

radio discussion, though other evidence of direct influence from Twitter is much more limited.

The American talk radio ecosystem, as revealed here, while still partly local in character, is an integral part of the national media ecosystem and best understood in that context.

Thesis Supervisor: Deb Roy
Title: Professor of Media Arts and Sciences

Mapping U.S. Talk Radio: A Textual Survey at Scale

by

William Brannon

This thesis has been reviewed and approved by the following committee members:

Deb Roy .....

Professor of Media Arts and Sciences
Massachusetts Institute of Technology

Jacob Andreas .....

Assistant Professor of Electrical Engineering and Computer Science
Massachusetts Institute of Technology

Yochai Benkler .....

Professor of Entrepreneurial Legal Studies
Harvard University

Acknowledgments

I'm grateful to many people for their help and support in writing this thesis. It's a cliche to say that I couldn't have done it alone, but it's true. My advisor Deb Roy has been consistently generous and supportive throughout, as have the readers, Jacob Andreas and Yochai Benkler. It's a much better work for their comments and ideas.

I appreciate the hard work put in by the engineering teams at the Lab for Social Machines and Cortico, whose development of the infrastructure behind radio and Twitter decahose data enabled this project. Alex Siegenfeld and Prashanth Vijayaraghavan also helped considerably by preparing the election-returns data used in parts of this work. I've also benefited greatly from discussing the topic with the capable people at LSM – particularly in-depth conversations with Doug Beeferman and Brandon Roy, but also many incisive comments from fellow students and staff at presentations and seminars. The atmosphere of interdisciplinary curiosity and collaboration fostered around the lab by Deb and its other research affiliates, especially in this case Kathy Cramer, has also been invaluable.

Last but of course not least, none of this would have happened without the support of my family and my fiancee Laura, who have been with me from the beginning.

Contents

List of Figures

List of Tables

Part I

Overview

Chapter 1

Introduction

If you want to understand function, study structure.

Francis Crick

The media, broadly defined, have a tremendous impact on public life. TV, radio, newspapers, social media, and others are both channels for discourse and shape the discourse that takes place on them. Indeed, different media have different effects on public life and political discourse: some directly, some by their influence on other media, and all to some degree by shaping public opinion. We focus here on the influential media of radio and Twitter.

Radio, especially right-wing talk radio, has been influential in politics as far back as the 1980s. Starting just before 1990, changes in the technology of syndication and the repeal of the Fairness Doctrine allowed the development of a new style of radio [1]. Rush Limbaugh and other early pioneers of this combative, political new format won themselves millions of listeners and great influence in politics.

More recently, Twitter has become journalism's water cooler, information clearinghouse and setting place for conventional wisdom, with a substantial impact on other media downstream of it. A large literature in several disciplines has explored Twitter's relationship to the news media and a great many other things, including electoral politics (e.g., [2] [3] [4] [5]). The size of this literature has a lot to do with data availability, of course: Twitter data, which is easily available for research, is widely used for that purpose.

Radio, unlike Twitter, has never been looked at in a systematic quantitative way, through large text corpora (which have been unavailable until very recently [6]). The intersection of the two media, and the influence of each on the other, is similarly lacking a large-scale examination.

Accordingly, this work is fundamentally a mapping exercise, examining hypotheses about how radio relates to other entities on the media landscape. Some of those factors are offline and some are online, but all have plausible means of influencing radio. Ultimately, the motivating and most interesting question here is the causal one: what influences the content of radio broadcasts? What are the downstream effects of radio, on elite discussion, other media channels and public opinion? But causal

questions are hard to answer in this setting, and so this thesis, as a first approach to the topic, is a mostly descriptive study. (See Chapter 9 for the exception that proves the rule.)

This question is in turn motivated by the importance of the media in influencing and indeed constituting public opinion: as explored in a long literature in political science, which we discuss below, media coverage influences both issue salience and the public's opinions on issues. Radio has remarkably wide reach in the U.S. [7], which makes its limited place in media ecosystems research all the more in need of a rethink.

1.1 Research Questions

The analysis here is threefold: internal radio factors (Part II), the intersection of Twitter and radio (Part III), and a pair of case studies (Part IV). The "internal radio factors" section of the work focuses on station ownership and geography. This program of work in turn reflects a typology of the most important social contexts in which radio is embedded:

The media ecosystem Both public and talk radio participate in the broader world of media, operationalized here as Twitter and especially influential journalists.

The corporate world Radio stations have owners, who make syndication decisions and hire / fire local hosts.

The physical world Radio stations are real-world entities with geographic locations, and their employees live in physical communities.

Ultimately, we want to understand how characteristics of radio content vary along these three dimensions, individually and as a joint 3d space, and to summarize key patterns of variation along one or more dimensions as defining characteristics of the overall "radioscape."

1.1.1 Radio Alone

In this part of the work (explored in Part II), we focus on station ownership and geography, and their relationship to the content of broadcasts. The overall question of interest here is whether stations with the same owner, or in the same geographic region, are more similar to each other than to stations with different owners or in different regions, and if so, how.

We get at this question in two ways:

Schedule similarity We estimate (Chapter 3) the degree of similarity between stations by comparing their schedules of programming. We can calculate the fraction of each station's airtime given over to each show, and compute the cosine similarities of the resulting vectors. Stations can then be compared along dimensions of ownership or geography.

Topic similarity We can estimate (Chapter 4) the similarity of both stations and shows by comparing how much they discuss certain topics. Given a set of modeled topics, we can again estimate how much of each station or show's airtime is given over to discussing it and compute cosine similarities between stations and shows. Stations and, this time, shows, can then be compared according to their ownership or geography.

1.1.2 Radio-Twitter Interface

Next we attempt to relate aspects of Twitter to the organization and broadcast content of radio (explored in Part III). It's important to note that there are several theoretical reasons to want to do this:

- Twitter is a site of elite discourse, with many influential journalists, politicians, academics and others among its regular users. In this regard, it offers an opportunity unavailable before social media to fit radio into the dynamics of elite opinion that in turn influence public opinion [8].
- Because journalists in particular are such heavy Twitter users[1], Twitter can also be thought of as a proxy for discourse in other media like newspapers and television.[2]
- Finally, though the population of Twitter users is not representative of the American public [11] [12], discussion among non-elite users offers an opportunity to compare radio discourse with the opinions of a broad if perhaps unrepresentative set of potential media consumers.

Concretely, we'll examine a variety of Twitter-side entities and compare them to radio-side entities and text: the text of tweets, communities in the follow graph, estimates of latent ideology from network structure, and others. The goal is to map the relationships between the two media, especially through the linkage provided by hosts who can be matched to their Twitter handles. This part of the work breaks down naturally into more specific questions, which we briefly explore.

Social Structure

Given Twitter's role as a site of discourse among elites, we should expect much of the underlying social structure of those elites to be visible there. It would be a powerful demonstration of radio's integration into the national media ecosystem if it encoded a similar social structure to media elites on Twitter.

[1] The academic literature has looked at how individual journalists use Twitter [9], and found extensive use. Estimates of the rate of Twitter usage among journalists, however, mostly come from commercial market research. One industry report found that fully 83% of journalists are on Twitter as of 2019 [10].

[2] Though our results in Section 7.2 don't directly address any medium but radio, this reading is more plausible in light of the similarity between Twitter and radio found there.

Note that by "integration" here we are not implying similarity of content, which is discussed below and examined later (Chapter 6). Radio might be integrated into the social structures of the media ecosystem without having similar content to other media. (For example, 20th Century Fox and Fox News were for a long time integrated into one company. Presumably their staffs had more social connections to each other than two divisions of different companies would have had, but as texts, the Marvel movies and Tucker Carlson Tonight are quite different.)

Operationally, we explore the question of common social structure (Chapter 5) in the following ways:

Latent ideology There are several methods for estimating a latent left-right ideological dimension from Twitter's follow graph (see Section 5.1). If we make such estimates for radio hosts and other personnel, are the estimates plausible? Do they rank hosts sensibly and line up with properties of stations and shows?

Graph communities Do communities in the follow, mention and retweet graphs correspond to meaningful groupings of shows on the radio side?

Graph similarity How structurally similar is the Twitter follow graph to radio's "co-airing graph," in which two shows are connected if they air on the same station? Both should be expected to reflect homophily, in one case driven by the decisions of journalists and hosts themselves, and in the other by radio programmers' opinions about those hosts.

We also attempt to relate the ideology and graph-community aspects of Twitter to the geographic and ownership dimensions of radio. An especially important question is whether some large owners' lineups are noticeably more liberal or conservative than others.[3]

Social Structure vs Radio Text

Unless it influences what's said on radio and Twitter, a common social structure between them won't matter much for public opinion. Thus in this part of the thesis (Chapter 6) we ask whether the social structure visible in Twitter is predictable from radio text. This question is operationalized in two ways:

Graph communities Do Twitter communities correspond to identifiable linguistic differences in shows, and which ones?

Latent ideology Does ideology as estimated from Twitter correspond to real differences in the ways hosts present their programs, and which ones?

Ideology vs election returns Finally, as a concrete test of Twitter's relationship to geography, we ask whether the text of a show is more closely related to, on the one hand, the hosts' ideology as estimated from Twitter or, on the other, the 2016 election results of the areas the show is broadcast to.

[3]Of course, finding that they are does not prove the ideological skew reflects the beliefs of the owners, but one could argue that on other grounds.

We adopt a predictive methodology for exploring these questions, because establishing that graph communities or ideology estimates are predictable from the text of a show establishes a relationship between them. The details of the relationship are mostly deferred to future work.

Comparison of Text

Because the effects of media on public opinion reflect what is said as well as who speaks, we consider how the text of Twitter discussion compares with the text of radio discussion.

First, we can ask the question in a static way, leaving out any consideration of time. How do the topics discussed vary between Twitter and radio, and how do the results change if we break both into subgroups? Do hosts themselves engage in different ways on the air and on Twitter?

Taking time into account, even more questions present themselves. What temporal relationships exist between discussion of topics on Twitter and radio? Does discussion of topics in one medium consistently lag or lead discussion in the other, and are causal effects plausible?

We're interested in questions like:

Statics Do broadcast media (radio) and social media (Twitter) reflect similar social structures? Are relationships evident in one also reflected in the other?

Dynamics Are there clear temporal relationships between them? In particular, does Twitter discussion predict later radio discussion? Can we make any convincing causal claims?

In view of Twitter's well-explored relationship to the traditional media, what do these results say about radio in relation to the broader media ecosystem?

1.2 Related Work

1.2.1 Cross-Medium Studies

First, we present some attempts to identify cross-medium relationships involving Twitter, especially those which also involve radio. Both media are discussed in more detail in their own sections below.

Part of the inspiration for this thesis was Benkler et al 2018 [13]. Their book provides a comprehensive and cross-medium look at the structure of the media ecosystem with a focus on right-wing misinformation. Though much of the book considers the impact of social media, they call out the long-standing importance of TV and talk radio in the right-wing media ecosystem. Radio itself, however, makes an appearance only through historical and legal analysis, rather than also through data. As the authors observe:

> We are only at the very beginning of the ability to create the capacity to engage in such broad, cross-platform research. Television archival data is becoming available; talk radio is still observable only through sporadic transcripts. (p. 384)

The development of the radio dataset described in Section 2.1 is one step toward greater observability for radio, and enables the sort of large-scale textual work conducted in this thesis.

Despite these data issues, a small group of studies have explored the intersection of radio and Twitter. Bonini and Sellas [14] examine the use of Twitter by European public radio broadcasters, and find their use of it wanting. The broadcasters tended to use Twitter, an inherently participatory medium, as an additional channel for broadcasting their content, and failed to engage users or listeners much. Herrera-Damas and Hermida [15] carry out a detailed case study of Twitter use by three radio stations, and also find stations using Twitter mostly as a one-way broadcast medium. They find radio stations as institutions and journalists individually using Twitter in very different ways, with the individuals treating Twitter much more interactively. Ferguson and Greer [16] examine a larger set of 111 stations and their Twitter use, finding similarly that stations have consistent brand identities on-air and online. News stations post mainly news items, music stations promote their content and live events, and in general online activity is closely related to offline content. On the other side of the radio transmitter, Chiumbu and Ligaga [17] did a detailed study of social media's effect on audience engagement with another three radio stations in South Africa. They find that Twitter, Facebook and other social media help knit audiences together and create more community around radio programs. They also mention that official and individual engagement on these platforms differs, with individual radio personalities engaging in a more authentic, interactive and varied way. All of these results help explain our choice to ignore official station accounts in analysis in favor of hosts and production staff.

A number of other literatures have explored Twitter's relationship to traditional media. Many studies (e.g., [18]) examine the diverse ways journalists interact with their audiences on Twitter, while others (e.g., [19] [20]) describe the ways Twitter changes journalists' coverage. The service is frequently used for news gathering and appears in published or broadcast stories. There are also studies of the reverse direction of Twitter activity being influenced by offline media. Larsson [21], for example, finds a clear pattern of tweets about a news TV show reflecting audience engagement with airings of the show.

Finally, while the subject matter is further afield from our aim here, a very large body of work has examined the spread of misinformation in social and traditional media. This literature has picked up steam especially since the start of the 2016 U.S. election campaign, though it began much earlier (e.g., [22]). Different streams of work have examined health-related misinformation [23] [24] [25]; fake or false news [26] [27] [28] [29]; and the determinants of people's uptake of such misinformation [30].

1.2.2 Public Opinion

A long literature in political science explores the determinants, correlates and nature of public opinion. Though consideration of the popular will and the people's role in government goes back to ancient philosophy, the modern literature began with Lippmann [31] in 1922. Because, as he put it, "the real environment is altogether too big, too complex, and too fleeting for direct acquaintance," people apply a range of heuristics to make sense of a complex world. As one of these heuristics is considering the opinions of others, and journalists are susceptible to the same biases as anyone else, mass media may distort people's perception of the world.

Later work in this paradigm explored in more detail how the public forms its opinions and acts on them (see [32] for a review of this latter side), with much of the foundational theoretical work occurring several decades ago. Converse in 1964 explored "belief systems" [33], or what are now better known as ideologies, finding that few members of the general public have a coherent ideological identity. The public's opinions on issues, in other words, do not line up well with the clear ideologies espoused by elites. Iyengar and Kinder in 1987 [34] drew a natural extension of these ideas to the mass media. They argued that media coverage, by determining which issues are salient and which issues are linked with which others in voters' minds, can influence expressed public opinion without having to actually persuade anyone on any issue.

This line of reasoning culminated, for our purposes, in Zaller's 1992 book [8] and a very similar article the same year by Zaller and Feldman [35]. This model of public opinion disputed the commonsense view that people have stable, consistent opinions about issues. Rather, individuals may have multiple conflicting opinions, and construct ideas about new situations on the fly; how these opinions shake out in practice is determined by which issues are salient. This in turn is a function of elite discussion, to which voters are exposed through the media. In this model, the "Receive-accept-sample" model of opinion [8], voters have

- Preexisting levels of exposure to and interest in politics, which influences how likely they are to receive new elite opinions;
- Preexisting beliefs, which influence how likely they are to accept any such new opinions;
- A sampling strategy for forming an opinion about a concrete political issue: voters combine the most salient ideas sampled from their recent experience to form opinions. More salient ideas have more influence on opinions as actually expressed.

In this framework, it's clear why we might be interested in Twitter's influence on other media like radio. Depending on which accounts one looks at, Twitter both exposes the general public (at least that portion of it with accounts) to elite discourse, and hosts that same elite discourse. It provides a powerful opportunity to examine part of this model in detail: to investigate whether the structures and flows of elite discourse do in fact correspond to what the broader public hears on the radio.

Alongside this theoretical work, a number of empirical studies have examined the impact of media consumption on people's behavior or opinions. There are far too many of these studies to list them all, so we'll highlight a few recent and relevant examples here. A number of specifically radio-related works are instead discussed in the appropriate subsection below.

King et al [3] confirm a core reason we find radio coverage and on-air discussion interesting: news media's impact on public discussion and opinion. A randomized experiment they ran with 48 media outlets, randomizing the dates outlets wrote stories about certain topics, finds that these outlets' coverage drives public discussion. Topics which were the subject of an article saw a 63% increase in Twitter discussion. Similarly, Martin and Yurukoglu [36] found a persuasive effect of cable news, using an instrumental-variables design and instrumenting by random assignment of channel number. Fox News watching in particular was found to increase Republican vote share in general elections by a substantively meaningful amount (up to 2.5 percentage points).

Since the COVID-19 pandemic began, several studies have found that media consumption can drive not just ideology and voting behavior but even opinions and information about public health. Jamieson and Albarracin [37] identified a relationship between exposure to mainstream or conservative media and (respectively) accurate or inaccurate beliefs about COVID-19. Simonson et al [38] found that Fox News exposure in particular increased noncompliance with social distancing guidelines. Media exposure, in other words, can influence not just changes in people's voting behavior but life-or-death decisions.

1.2.3 Radio

A number of literatures have considered the structure and impact of radio. Radio as a commercial broadcast medium is about 100 years old, and we won't attempt to look back that far. Many of the relevant features of radio today – concentrated ownership, widespread syndication, and the talk format – are only a few decades old.

Rosenwald [1] lays out the history of talk radio as a format, and its effects on politics, in a thorough and valuable recent book. He identifies the current dominance of conservative talk as having two causes: a wave of media deregulation in the 1980s, especially the FCC's repeal of the Fairness Doctrine, and technological changes that made syndication affordable. Several other authors have similar diagnoses: Berry and Sobieraj [39], for example, also trace talk radio's rise to deregulation and technological change. They do, however, stress the importance of online music streaming and piracy as a form of technological change, because these effects of the Internet eroded the economics of music radio and especially FM stations. As music became less profitable and syndication was cheap, stations shifted to talk. The consolidated ownership that's such a notable feature of the contemporary radio market came later than syndication but had similar origins: Drushel, as far back as 1998 [40], traced it to deregulation adopted in the Telecommunications Act of 1996.

Rather than trying to understand talk radio's economic or legal origins, other work has tried to locate it within broader social structures. Hendy [41] places radio in a

global context and tries to provide a worldwide survey. For our purposes, his emphasis on the importance of rigid formats (talk, public, country music, etc) and their origins provides valuable context for understanding talk radio. Berry and Sobieraj, in a 2016 book [42] and an earlier article [43], identify talk radio as part of a broader trend toward "outrage" in media. They describe outrage-based media as increasingly stable, institutionalized and permanent, and locate its origins in economics. Put simply, outrage doesn't require much in the way of real journalism, research or production values, which makes it cheap; distribution and production costs (especially in radio) are low; and a combination of long-standing deregulation and more recent digital technology has made segmenting a specific outrage genre to the right audience practical in most media. Jamieson [44], writing several years earlier, made a related but distinct effort to place talk radio within the conservative media ecosystem. She focused in particular on the epistemic closure developing in conservative media, a particularly concerning problem if conservative outlets like much of radio have large impacts on their listeners.

Exactly that impact, and its connection to models of public opinion like that presented above, is demonstrated in a line of empirical research on radio. Hofstetter et al [45] [46] examine the roles that talk radio plays for its listeners. The most important are seeking information, contextualizing and interpreting the information viewers have about the world, and "parasocial interaction" with the hosts and guests. They find that these programs do indeed communicate information to their listeners, and also that listeners more exposed to right-wing radio are more misinformed about certain verifiable facts. Barker et al, in three articles from the late 1990s [47] [48] [49], test for the effect of exposure to right-wing radio on several political outcomes – opposition to the Clinton health care plan, voting for Republican candidates, and views on specific issues – and find that it has meaningful impact on its listeners. All of these, and especially [49], can be viewed as a test of the model of public opinion presented above. In that study, Rush Limbaugh listeners were moved to express more agreement with his views on issues he discussed on-air, but not on other issues – just as a model of public opinion in terms of issue salience and elite cues would predict.

Indeed, much of this work [1] [44] [47] [48] [49] stresses the role of Rush Limbaugh in particular in defining the talk radio format. His show, which found national success at the very moment deregulation made it possible for such a show to be widely syndicated, pioneered conservative talk and set the standard for later hosts. His use of his newfound platform to push for conservative policy changes in some instances, while remaining firmly an entertainer with an eye on his brand [1], set a very influential example.

1.2.4 Twitter

While Twitter came on the scene more recently than radio, it's been the subject of a remarkable amount of research since. The company was founded in 2006, and the research on its impact in the succeeding 14 years has reached almost as far as the impact itself. Here we discuss research on those aspects relevant to radio and media agenda-setting; some cross-medium studies of radio and Twitter together were

discussed above.

One important line of research focuses on the structure of Twitter in general, asking what role it plays for those who use it. In 2010, Kwak et al [50] conducted an influential examination of the graph structure of Twitter's entire network. They described several "deviation[s] from known characteristics of human social networks," indicating that the service included and was used for both social structure and non-social channels for information spread. They found that while communication on Twitter may be mediated by social aspects of the service, its content was heavily skewed toward information gathering. Fully 85% of top trending topics were news articles, headlines, or other such content. Myers et al [51], writing in 2014, confirmed this view with similar findings and added a focus on the experience of Twitter for an individual user. They hypothesized that smaller and larger users by follower count experienced the social and informational aspects of the service differently: smaller users found Twitter useful for following news and discovering information, while larger and more central users were more tightly connected to Twitter's social structure.

Other work has taken a different angle, focusing not on how users use Twitter but on (so to speak) how Twitter uses its users. Through its structure, the platform provides incentives for certain kinds of behavior, and much work has tried to understand what these incentives are. As far back as 2010 [52], Yardi and Boyd found that Twitter is groupish. While it exposes people to diverse points of view, social status incentives drive interaction with in-group and out-group members in different directions. Outgroup interactions are frequently hostile and are used to reinforce group cohesion, while ingroup interactions are affirming and strengthen one's identification with the ingroup. Conover et al [53] made the closely related finding that retweet and mention behavior on Twitter are strikingly different from each other. The retweet graph decomposes into ideological groups, while mentions form a more connected, politically heterogeneous whole. They take this finding as evidence for trolling behavior: users mention each other as a way to inject political content into their adversaries' communities and provoke confrontation.

Much research has made the point that Twitter is not representative of public opinion. We will cite only certain examples here. Most recently, the Pew Research Center confirmed again [12] the widely investigated fact that most tweets about any given topic come from a tiny minority of tweeters. In this case, tweets about national politics (which made up 13% of tweets in total) came overwhelmingly from only 10% of users who produced 97% of national politics-related tweets. Given that Twitter has less than 10% of the US population as daily active users [54], a remarkably small slice of the public is determining the content of discussion on what has become the digital public square. Moreover, neither the Twitter user population in general nor those highly engaged tweeters are representative of the general public. Barberá and Rivero [5] found tweets about elections (the 2012 US presidential election in their case) come disproportionately from a highly demographically unrepresentative set of users. These users are more likely to be male than the US population, have higher incomes, and are more likely to live in urban areas. Most of all (cf the public opinion model above), political tweeters are highly unusual in having stable ideological perspectives, and notably extreme ones at that.

A considerable amount of work has looked at Twitter's involvement with journalism and journalists. Journalists themselves discuss this topic regularly and at great length (e.g., [55] [56] [57]), and the academic literature has not ignored it either. Wu et al [58] made the early observation that attention on Twitter is concentrated heavily onto a small number of the most followed users, and that these users are disproportionately journalists as well as various kinds of celebrities. The journalists in this elite group, however, spread the most information and account for the larger share of the content provided by the most followed users. (This view of Twitter's structure inspired our decision to focus on "elite Twitter" of journalists and politicians for most analysis in later chapters.)

Later work has drilled down on the ways journalists use social media, and what it can reveal about their professional practices. Willnat and Weaver [59] find that nearly all journalists are on social media, and that the most common motivation for using it is to keep up with one's peers and follow the work of others. Usher and Ng [4] take a close look at the social structure of a group of Washington political journalists, and find that they form cohesive subcommunities. Their interpretation, perhaps unfortunately in light of Willnat and Weaver's results, is that journalists may operate in and be constrained by even smaller professional bubbles than previously thought.

1.2.5 Open Questions

Taken together, these literatures suggest a close link between social and traditional media, and particularly a link between Twitter and radio. They don't, however, address whether such a link exists on a larger scale than single shows. Because of the social-network aspect of Twitter, considering multiple shows and their hosts' Twitter presences at once may reveal important social structure. Nor does the existing literature examine on a large scale how any link between radio and Twitter is reflected in their content. With large radio corpora like [6] now available, we can begin to provide this missing piece and locate radio in the context of the broader media ecosystem.

Chapter 2

Data Sources

Because this thesis focuses on drawing social-science conclusions from data, this chapter discusses the data sources which go into our analysis in some detail. The most important is a large and novel dataset of terrestrial radio broadcasts, sourced from a diverse set of stations and enriched with various metadata. Other important datasets include a random sample of all tweets from Twitter's firehose, the tweets and follow relationships of a set of 2,836 influential Twitter users in media and politics, and a mapping of 198 of these users to the radio shows they host or produce. For certain analyses, we used standard population data from the 2016 American Community Survey [60], or precinct-level election returns (discussed below).

2.1 Radio Data

The main dataset we rely on consists of talk radio transcripts, collected at the Lab for Social Machines since early 2018. This is the first large-scale dataset of radio transcripts[1], and makes the analysis here possible. A sample of this data was published at Interspeech 2019 [6], though it covers an earlier period than this thesis. We focus here on two periods of time: September – October 2019, and March – April 2020. The transcripts were produced by automatic speech recognition, discussed below. They are segmented into estimated speaker turns, which we'll sometimes refer to as "snippets."

The stations included cover a wide range of geographic areas and original research purposes, and have changed since the first collection of radio content in early 2018. The first 50 stations were a random sample from Radio-Locator, a database of all U.S. radio stations [61], with a number of other stations subsequently added from the same database for other research projects. Because of technical issues and stations added for other projects, not all stations included in our corpus were observed for all four months. We work around this in later analysis in several ways: aggregating all talk or public radio stations together, choosing the best airing (on any station) of each show on a given day, and others as appropriate.

[1] Besides bespoke corpora constructed for specific research projects, certain shows provide transcripts, including several NPR shows and Rush Limbaugh.

There are several kinds of metadata appended to the corpus that we make use of here:

Station-level attributes As provided in [61]. For our purposes, the most important ones are geographic location, format (public radio, talk, news/talk, etc), owner's name and AM/FM frequency band. Most of these attributes are ultimately from regulatory filings with the FCC. We also manually standardized the owner names, so that, for example, several of iHeartMedia's holding companies and subsidiaries were combined into one owner.

Coverage maps Also as provided in [61]. These were computed using the location, power and frequency of the station's transmitter, as well as the topography of surrounding areas. Radio-Locator provided three sets of "local," "distant" and "fringe" maps, corresponding to respectively "predicted 60, 50, and 40 dBμ signal strength contours" for FM stations and to "predicted 2.0, 0.5, and 0.15 mV/m contours" for AM stations. Of these, we generally used only the "distant" contour. Within this region, a station should have adequate reception on "a good car radio or a good stereo with a good antenna."

Schedule data That is, which shows were airing at which times, collected generally by scraping station websites. Not all recorded content could be assigned a show. We kept the content without recorded show names, and used it in some analysis. We did, however, exclude from the corpus certain shows: Those which were either a) all or mostly music, b) sports-focused or c) the BBC World Service, which is not focused on American politics or culture. These shows were not the sort of "talk" radio we were interested in, and music in particular is not transcribed well. After these exclusions, the remaining 847 shows account for 83.3% of the whole corpus, with the remaining 16.7% having no schedule information. There are also show-confidence scores, expressing how confident the scraping team was in the correctness of the data.

Recognition outputs Each speaker turn or snippet includes the average recognition confidence on a 0 to 1 scale, as well as the imputed male/female sex of the speaker and whether the speaker was recorded by a telephone (i.e., a call-in) or professional audio equipment (i.e., present in a studio). Diarization also attempted to disambiguate speakers, but the generated speaker IDs were not unique across the corpus and we did not use them.

Twitter mapping We were able to match 67 shows to the Twitter accounts of their hosts and sometimes production staff. To select the shows, we first listed all 847 shows in descending order by amount of content collected and searched[2] for relevant Twitter accounts. After finding approximately 45 shows, we switched to searching for local shows, considering the sample complete after finding approximately half as many. The final sample has 44 syndicated shows and 23 local shows, with a total of 198 Twitter users.

[2]Manually. The most useful tools were Google, Twitter search, LinkedIn, etc.

Syndication status We experimented with several ways of determining whether a piece of content was from a syndicated show. None were perfect: some content lacked schedule data, content-based detection suffered from different recognition errors in the same show recorded on multiple stations, and detection based on the underlying audio was too involved. In the end we considered "syndicated" any content assigned to a show that appeared on multiple stations in the corpus, and "local" any content assigned to a show appearing on only one station. Content without a show name is excluded from analysis that relies on a syndicated-vs-local split.

An example snippet is shown in 2.2, and an example coverage map in 2-1. Note that the set of shows linked to Twitter accounts is skewed toward large syndicated shows. This is both because such shows' hosts are more likely to be on Twitter, and because accounts for small-time and local hosts are harder to find even when they do exist. The five largest and five smallest such shows are shown in table 2.1, and the full list is presented in Appendix C.

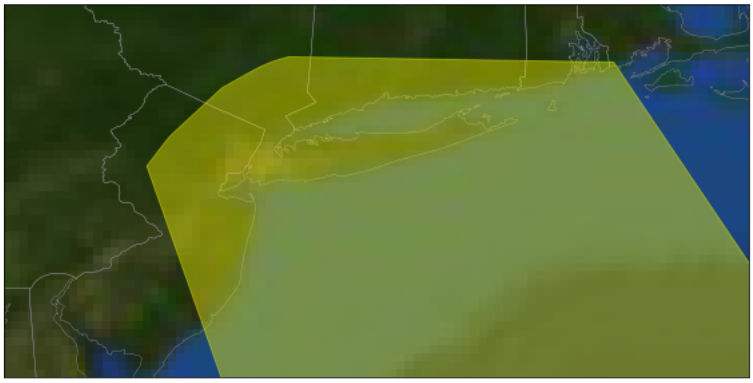


Figure 2-1: An example coverage map, with the coverage area in yellow. The station depicted here is WNYC-AM in New York City during the day. (AM radio stations can be received from much greater distances at night.) Only Radio-Locator's "distant coverage" area, the 50 decibel-microvolt signal strength contour, is shown.

In total, the final corpus includes:

- 212 stations;
- 847 shows distinct shows recorded;
- 410,154 hours of audio / 3.96 billion words of text;
- 67 shows matched to Twitter accounts of hosts and other personnel, totaling 198 accounts and comprising 47.2% of the corpus by word count.

2.1.1 Speech recognition

The ASR system used for this corpus is the same one used by Beeferman et al [6]. As they described it:

Show Name	Words Recorded	Twitter accounts
Coast to Coast AM with George Noory	305,068,674	JChurchRadio, g_knapp, GeorgeNooryC2C, coasttocoastam
The Sean Hannity Show	221,607,196	seanhannity, PrdcrJG, LyndaMick, f_treacher, benbrownmiller
Rush Limbaugh	205,276,329	limbaugh, rushlimbaugh, BoSnerdley, yesnicksearcy, toddeherman, RogerHedgecock, MarkSteynOnline, KenMatthews
Glenn Beck	120,943,995	glennbeck, marissananos, DomTheodore
Morning Edition	98,067,537	ReenaAdvani, rachelnpr, PhilHarrellNPR, nprkyoung, NPRinskeep, nprgreene, NoelKing, Nancy_Pearl, MorningEdition, mirandatk, bgordemer
Talk With The Green Guy	252,228	greenguymedia
Watch Dog on Wall Street with Chris Markowski	199,027	chrismarko
City Visions	173,698	cityvisionsKALW
RadioLab Invisibilia	98,889	HannaRosin, aspiegelnpr
Tech It Out	61,421	marc_saltzman

Table 2.1: The five shows matched to Twitter accounts which have the most recorded audio, and the five which have the least. The corresponding Twitter accounts are shown as well.

> Our speech-to-text model is based on an entry by Peddinti et al. [62] in the IARPA ASpIRE challenge. Its acoustic model has a time-delay neural network (TDNN) architecture geared for speech in reverberant environments, and offered an appropriate trade-off of accuracy on radio and decoding efficiency for our needs. It is trained on the English portion of the Fisher corpus. To reduce word error rates, we replaced the lexicon and language model, retraining them on several corpora of human-transcribed radio: several years each of broadcasts from a conservative talk show [Rush Limbaugh] and two National Public Radio news/talk shows. [Talk of the Nation, Morning Edition] Keeping current with these sources gives our system better coverage of proper names in the news. The final speech-to-text model is implemented with the commonly used Kaldi toolkit [63]. We observed a word error rate of approximately 13.1% with this system, as

Data Point	Value
Callsign	KNST
Format	News/Talk
Frequency Band	AM
Standardized Owner	iHeart / Clear Channel
Is Public Radio?	false
State	AZ
Census Region 5-way	West
Show Name	Garret Lewis
Show Confidence	0.78
Content	going to floods facebook with undeclared coronavirus propaganda ads blaming trump so china is buying a whole bunch of facebook ads
Diarization: Studio or Telephone?	S
Diarization: Gender	M
Mean Word Confidence	0.8857
Word Count	21
Duration	9.74 s
Start Time	2020-04-07 10:50:53.36-04
End Time	2020-04-07 10:51:03.1-04
Approx Syndicated	false

Table 2.2: An example snippet from the radio corpus employed in this thesis. Not all fields are shown, and in particular coverage maps are omitted. This snippet is from a local show in Arizona. One can see a common sort of error the ASR system makes: "flood facebook" becomes "floods facebook", which is incorrect but still close to the correct word.

> measured on a set of human-transcribed talk radio content that aired after the time period of the system's training data. (On the same basis, the Google Cloud Speech-to-Text API gave a 7.1% word error rate, but its cost was prohibitive for the scale of our project, more than 40 times the cost per hour of our Kaldi-based solution.)

We refer the reader to [6] for further information.

Anecdotally, the transcripts are readily comprehensible but contain noticeable errors. The use of up-to-date language model training data should help with recognition of proper names, but we haven't undertaken a systematic effort to assess relative misrecognition rates for important categories of words.

2.2 Elite Twitter Data

We identified an "elite" Twitter universe, consisting of journalists, pundits, politicians and others who are plugged into or are prominent subjects of media discussion. The final universe contained 2,836 Twitter users, including 198 who are hosts or production staff on radio shows. It is not intended to be exhaustive of the sets of journalists, pundits or prominent politicians, but to provide a sufficiently representative convenience sample of them. Note that this universe should not be taken to be a distinct community from the broader set of Twitter users; indeed, the users in it are widely followed and influential.

The universe was constructed as follows, relying partly on pre-built Twitter lists from media outlets:

- To cover politicians, we started with C-SPAN's Twitter list "@cspan/members-of-congress".[3]
- For journalists, we wanted to capture a range of perspectives from traditional newspapers, TV and online outlets. This goal led to including the following Twitter lists: "@cspan/political-reporters", "@nytimes/nyt-journalists", "@cspan/congressional-media", "@washingtonpost/washington-post-people", "@slate/left-leaning-tweets", "@msnbc/msnbc-hosts", "@slate/right-leaning-tweets", and "@foxnews/shows-hosts". These lists provide some explicitly left- and right-leaning accounts (as selected by Slate) and a number of staffers at traditionally objective mainstream media sources.
- We also included 197 additional manually selected users, of whom 49 were associated with Fox News; 24 were liberal politicians or commentators, 107 were conservative commentators or political figures, and 17 were Trump-administration or Trump-campaign officials. In particular, President Trump's account @realDonaldTrump was included.
- Finally, we included the 198 users identified in 2.1 as hosts or staff of shows.

Given this list of accounts, we pulled the following information about the users from the Twitter API:

- Their user profiles, including self-reported location, short biographies, verified status and whether the users are verified or have protected tweets.
- The lists of all users they follow or are followed by.
- Their tweets, as far back as available. (The Twitter API we had access to generally returns only the most recent 3200 tweets for users who have posted more than this many, and of course protected users had no publicly available tweets.) The initial pull of tweets occurred in early November 2019, and new tweets were pulled daily thereafter.

[3]Lists are formatted here in the same way Twitter's API returns them, @USER/LISTNAME. To see the users in a list specified this way, one can go to https://twitter.com/USER/lists/LISTNAME.

From these primary Twitter data fields we computed several additional datasets, including in particular the mention, reply and retweet graphs connecting the users.

2.3 Twitter Decahose

Twitter provides a variety of options for real-time access to the stream of tweets. The complete stream is called the firehose, and as the name implies is an impractically large amount of data for most purposes. The Lab for Social Machines instead has access to the "decahose" [64], a 10% random sample of all tweets. Four months of even this is a large amount of data for keyword-counting analysis, so we subsampled further, to 0.2% of the decahose (i.e., 0.2% of 10%, or one in every 5,000 tweets in the overall firehose). The sampling was Bernoulli, which is to say every tweet was selected or not independently with the same probability. The final dataset contained 3,018,779 tweets.

The decahose feed provides only limited information about the users posting the tweets, and in particular does not include any form of the follow graph around them.

2.4 2016 Election Results

Certain parts of our analysis, described in Chapter 6, attempt to compare Twitter-based estimates of show ideology to the opinions of the geographic areas shows are broadcast to. This analysis requires a good proxy for or survey of public opinion on a left-right scale, and requires it to be mapped to a granular geographic unit (to cross-reference with the radio broadcast areas discussed above). The natural choice, if available, is precinct-level returns for the most recent presidential election. Unfortunately, there is no standard or official source of such results for U.S. elections. Academic projects like [65] have compiled this data, but without the shapefiles we need to cross-reference with radio broadcast areas.[4]

Accordingly, we took advantage of a recent New York Times feature [66] presenting precinct-level returns with approximate precinct boundaries. Webscraping these yielded a list of precincts with geographic boundaries, though it should be noted the Times did not provide details of how these boundaries were calculated. The final dataset contained 179,325 precincts, covering the 48 continental U.S. states and D.C.

2.5 Text Standardization

Text content is formatted differently between Twitter and our radio transcripts. Especially given the use of keyword-counting methods and n-gram language models in analysis, these differences required a process to standardize the text across media. The details of this process are described in Appendix A; here we provide only a summary.

[4] Precincts in U.S. elections are nebulous things: they don't always correspond to well-defined geographies, they change from election to election, and data about them is hard to find.

First note that the ASR radio transcripts are formatted in a particular way. They are all lowercase ASCII, with no punctuation, and with numbers written out as words (e.g., a person saying the year 2010 would result in "twenty ten" or "two thousand ten"). Our goal was to enrich this data slightly and get Twitter content to match it. The process consisted of the following steps:

Phrase detection We used automated collocation detection and a list of Wikipedia article titles to combine commonly used phrases into single tokens. This part of the process only was also applied to the radio data.

Simple preprocessing Punctuation and whitespace removal, lowercase, transliteration of non-ASCII characters.

Twitter-only features We discarded URLs, retweet markers, usernames, etc.

Autocomplete A technical issue meant that some of the decahose data was truncated at 140 characters, sometimes resulting in final partial words. We used a hidden Markov model to infer these partially missing words.

Hashtag segmentation We segmented hashtags into words according to a simple algorithm, dropping hashtags whose text did not decompose entirely into known words.

Number formatting We processed numbers in tweets written with numerals to be in the same word-based format as in the radio data.

Part II

Radio Alone

Chapter 3

Similarity of Programming

This chapter begins our exploration of radio's internal structure. This chapter and the following one will consider radio in isolation, without regard to its connections to Twitter and the rest of the media ecosystem. We focus here on examining the distribution of shows on radio - that is, comparing stations, owners and geographic regions according to which programs air.

This chapter and the analysis of topics in Chapter 4 can both be viewed as examining the content of radio broadcasts, but from different angles. Chapter 4 examines this content directly, comparing the topic mixtures present in stations and shows. One can, however, discuss the same topic in a variety of ways and from a variety of perspectives. Quantifying such perspectives and the differences between them is difficult, which is why we've elected to look here only at the distribution of shows. Such analysis of programming bears on content, but only indirectly, insofar as different shows and hosts do express different perspectives.[1]

3.1 Methodology

There are a number of commonly used approaches for comparing the similarity of data points, depending on the type of data. Here, our task is relatively simple: we have stations, owners or other real-world entities, and a number of shows they may choose to air. If we take each show to be a dimension of a vector space, the natural choice is to represent a station or owner as a vector of the amounts of time devoted to each show, and compare stations by their cosine distance.[2] To deal with the fact that not every station was collected for the same length of time, we worked with each show's fraction of airtime rather than raw durations. Because ownership and geography are properties of stations, we kept these analyses at the station level.

Accordingly, we calculated each station's "schedule distribution," the fraction of

[1] For a very topical and very high-impact recent example of hosts doing so, see [67].

[2] This vector space representation is only approximate, because the amounts of airtime stations devote to different shows are dependent. They can't air more than 24 hours of programming a day, so more of one show means less of another. But no single show constitutes a large enough fraction of airtime for this to be a real concern.

airtime devoted to each show, and calculated the matrix of cosine similarities between all station pairs.

In this chapter and Chapter 4, however, we make little use of hypothesis tests or any sort of inferential statistics. This is for two reasons:

- Taking the population to be the set of stations in our corpus, there's no reason to work with a sample rather than the entire population. So most figures below are, in this sense, population figures.
- Taking the population to be the entire set of US talk radio stations, the sampling strategy that produced our corpus is quite complicated. Rather than being a simple random sample, it reflects both a) a random sample of 50 stations at the beginning of 2018, not all of which have remained in the corpus since, and b) convenience samples added at various points for various research projects. Finding the right weighting approach to make population inferences was a complex enough topic we've elected to present these analyses without making claims about the entire radio station population.

3.2 Results

We examined the relationship between station similarity and various properties of the stations: their ownership, geographic location and frequency band (AM or FM). By far the clearest effect was the difference in these patterns between public and talk radio. Not only do these two types of radio air different shows, but the factors related to which shows they air are different.

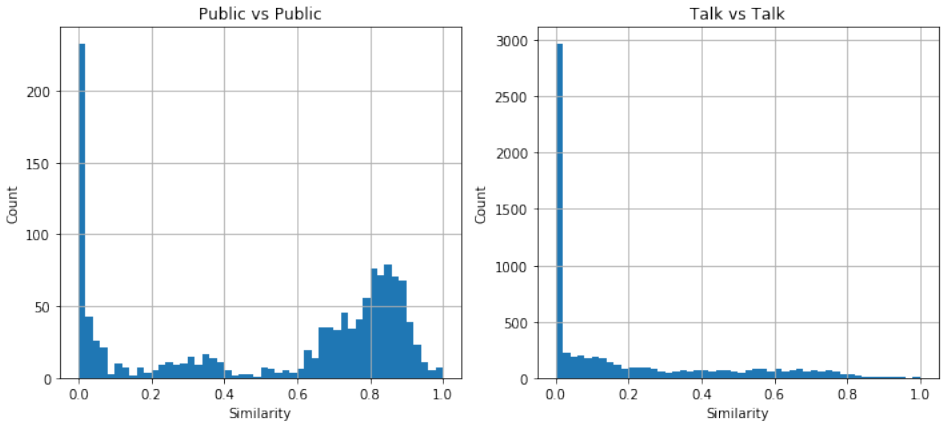
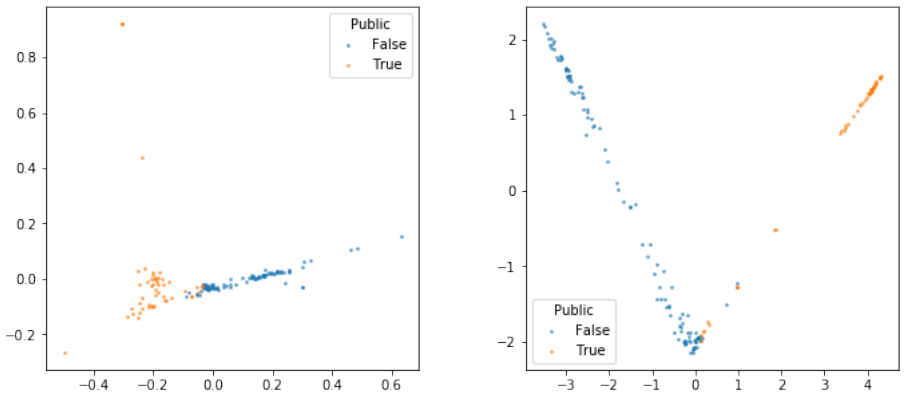


Figure 3-1: The distribution of all-pairs cosine similarities within public and talk radio. The greater homogeneity of public radio is apparent.

3.2.1 Public vs Talk

Public radio is more homogeneous than talk overall: on the 0 to 1 cosine similarity scale[3], the average schedule similarity of two public radio stations is 0.536, while for two talk stations it is only 0.186. (Because there are few shows airing on both public and talk radio, talk and public show distributions are almost perfectly orthogonal with an average talk-vs-public similarity of 0.004.) The increased homogeneity of public radio is due in large part to the large reach and long runtimes of popular syndicated NPR programs, especially All Things Considered and Morning Edition. No syndication network has an at all comparable reach on the talk radio side.



(a) 2d PCA of schedule distributions (b) 2d PCA of similarity matrix

Figure 3-2: The schedule distributions of all radio stations in the corpus and their cosine similarity matrix, both projected onto their first two principal components. The stations are color-coded by whether they are talk or public radio.

Figure 3-1 depicts histograms of these distributions and highlights public radio's greater homogeneity. We've also shown the first two principal components of the schedule distributions and their similarity matrix in Figure 3-2, with the near-orthogonality of talk and public radio clearly visible in Figure 3-2b. Finally, Figure 3-3 displays the first two principal components of talk and public radio stations considered separately. The results help illustrate the nature of the many station pairs in Figure 3-1 with near-zero similarity: for public radio, they are mostly driven by a few outliers, while talk radio also has more subgroup structure.

Though we haven't done this analysis rigorously, weighting airtime by audience data, the trend is even clearer when considering the time slots these shows occupy. Morning Edition airs during morning drive time, when listenership is high, and Coast to Coast after most people have gone to bed. Further down the top-5 list in 3.1, Rush Limbaugh is on in the afternoons on (five) weekdays, and Weekend Edition airs only on (two) weekend days, despite Rush having only 57% greater a share of airtime.

Finally, we note that while public radio relies more heavily on a single centralized syndication network than talk radio, its ownership is more dispersed. There are 115

[3]All entries in the vectors are nonnegative.

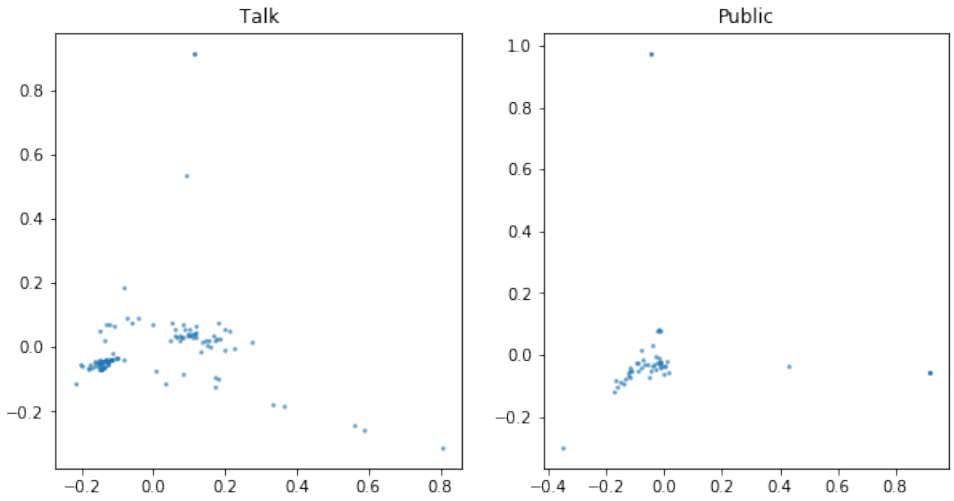


Figure 3-3: The schedule distributions of all radio stations in the corpus, broken out by talk-or-public status and projected onto their first two principal components.

talk stations with show data in our corpus, vs 51 public stations. Only 13 of these 51 are owned by owners who own more than one of them, with no owner having more than three stations. For talk radio, by contrast, only 31 of 115 stations are owned by single-station owners, and 37 are owned by the top owner, iHeartMedia. In fact, as shown in 3.2, all five of the top owners of talk stations in our sample are publicly traded companies.

3.2.2 Correlates of Similarity

Within each kind of radio, the correlates of similarity as summarized in 3.3 are also different. In brief: Syndicated content on talk radio depends most strongly on the ownership of the station, without much relationship to geography, and with a fair bit of diversity among owners. On public radio, despite the high degree of syndication overall, common owners seek out more diversity between their stations rather than synchronizing on the same choices of syndicated content. Geography at the level of Census region (West, South, Midwest, or Northeast) has little relationship to similarity of programming.

Unfortunately, we lacked a large enough set of stations to make reliable within-state comparisons. In the same vein, we note that because, as outlined above, public radio ownership is more diversified than for talk radio, the same-owner comparisons for public radio are not based on as many stations.

Talk Radio

Within talk radio, we can see that the greatest degree of within-group similarity is for stations with a common owner. Two talk stations with a common owner have an average cosine similarity of 0.43, compared to 0.15 for stations with different owners. This large gap reflects common syndication decisions made by station owners (see

Position	Talk	Public
1	Coast to Coast AM with George Noory (11.0%)	Morning Edition (20.5%)
2	The Sean Hannity Show (7.8%)	All Things Considered (13.1%)
3	The Rush Limbaugh Show (7.7%)	Weekend Edition (4.9%)
4	The Glenn Beck Program (4.3%)	Fresh Air (4.8%)
5	This Morning with Gordon Deal (3.4%)	Here and Now (4.8%)
Total	34.2%	48.1%

Table 3.1: Top five shows by duration for public and talk radio. The percentage of, respectively, all public-radio airtime and all talk-radio airtime these shows account for is shown in parentheses.

Position	Talk	Public
1	iHeartMedia (37)	Arizona Western College (3)
2	Cumulus Media (8)	Northern Arizona University (3)
3	Salem Media (6)	St. Paul Bible College (3)
4	Entercom (5)	Georgia Public Broadcasting (2)
5	Townsquare Media (4)	Washington State University (2)
Total	34.2%	48.1%

Table 3.2: Top five owners of stations in our corpus for public and talk radio. The number of stations owned by each is shown in parentheses. For public radio, no other owner owns more than one station in the corpus.

[1] chapter 1 for a discussion of the origins of heavy syndication). Other effects were comparatively modest: pairs which were both AM stations or both FM stations had average similarity 0.19, vs 0.17 for different frequency bands, and even pairs in the same Census region had an average similarity of 0.21, vs 0.18 for different regions. While talk radio is not as homogeneous as public radio, the centralizing and homogenizing influences on it are related to ownership rather than geography.

Public Radio

Public radio displays quite different patterns. Notably, while Census region continues to have little relationship to similarity at 0.54 for pairs with a common region, vs 0.53 without, other effects were in different directions. Stations with a common owner were less similar than stations without one, at 0.40 average similarity with vs 0.54 without. Being located on the same AM/FM frequency band corresponded to a notable increase in similarity here, at 0.56 for station pairs with the same band vs 0.43 without.

Examining the stations and their schedules individually suggests that the common

Variable	All	Public	Talk
Same Owner	0.432	0.395	0.432
Different Owners	0.123	0.537	0.154
Same Band	0.204	0.561	0.191
Different Bands	0.070	0.443	0.173
Same Census Region	0.156	0.542	0.213
Different Census Regions	0.135	0.534	0.176

Table 3.3: The average cosine similarity of schedules within and between station groups formed by several cross-cutting variables. These are population figures, so no standard errors are shown.

ownership effect is due to multi-station owners pursuing more diverse formats across stations. Recall that ownership is more diversified in public radio, and the largest common owners are not as large. An owner of three stations may pursue a different programming strategy than an owner of 37.

The most likely explanation for the AM-vs-FM difference is that, because public radio airs more music than talk radio does, and FM is higher-fidelity, music and non-music shows polarize by frequency band. While, as noted in Chapter 2, we excluded a number of all-music shows, public radio shows often incorporate some musical content while still being primarily talk. It's plausible, though we have not checked in detail, that such shows would prefer FM.

Chapter 4

Similarity of Content

This chapter continues the exploration of internal radio structure begun in Chapter 3. Here our focus is on the content of radio broadcasts directly, rather than simply which shows are airing on which stations.

The central question of interest here, building on Chapter 3, is what correlates with[1] similarity of content between stations and shows. Of course, text is a much higher-dimensional object than a vector of shares of airtime. If we're interested in correlates of content similarity between stations and shows, how to measure content similarity is an important question. We adopt an approach focused on topic modeling, described below.

Though we do not attempt to quantify different viewpoints, sentiments or perspectives on these topics, it's worth noting that different shows will in general have different perspectives. In that regard, these two directions of analysis – topics and programming – are complementary.

4.1 Methodology

As before, we represent stations as vectors and compare them via cosine similarity. This time, however, we can also compare shows. The vector space in question has one dimension for each of several topics, chosen as described below.

4.1.1 Topic selection

Extracting topics from a corpus of text is usually a complex undertaking. Text corpora rarely decompose cleanly into mutually exclusive categories, leading to a number of methods to extract topic structure in different ways and under different assumptions [68] [69]. Perhaps the most popular technique is latent Dirichlet allocation (LDA) [69], a general-purpose unsupervised algorithm which models documents as combinations of latent topics.

[1] "Causes" is a fraught word and in general we won't be able to do more than speculate about causality in this setting.

A frequent problem with LDA, however, is interpreting the resulting topics. The number of topics is a tunable hyperparameter, and overly large values of it may produce topics which overfit the data and fail to encode meaningful conceptual differences. Conversely, with too few topics, important but rare subjects of discussion may not be identified.

For these and other reasons[2] we elected to avoid bottom-up unsupervised topic detection in favor of a top-down approach based on word vectors. We used the approach and code in [70], which is different in important ways from other uses in the literature of word vectors for topic modeling. Most applications have tried to augment LDA with information contained in a word vector model [71] [72] or to develop another bottom-up model [73], whereas we took a top-down approach focused on indexing mentions of prespecified topics. (See the discussion in Section 4.1.1 about the particular topics chosen.)

For each topic, we expanded the set of terms using a word2vec model [74] trained on all four months of radio data. A configurable number k of additional terms per topic were added, but only if they were sufficiently close to all seed terms in word vector space. The parameter k was 200 for most topics, but more for conceptually "large" topics like politics and sports. When too many words were close enough, we sorted by each term's average distance to the seed terms and took the top k. Requiring closeness to all seed terms, rather than any seed term, helps preserve the conceptual coherence of the resulting term list. We refer to the final set of terms for each topic as its query words or query terms. The topics and complete lists of terms are presented in Appendix B.

From these lists of terms, we generated the final vectors as follows: given n topics, each of the n elements of a station or show's topic distribution is the fraction of all its words consisting of that topic's query words, rather than the raw count. This normalization adjusts for many possible sources of bias, including some hosts speaking faster than others. It also obviates the need to de-duplicate syndicated content: if, for example, we have 10 stations airing the same Rush Limbaugh episode on any given date, both the numerator and denominator will become 10 times larger.

This approach to topic modeling has benefits (clearly interpretable topics) over the usual unsupervised methods. It also has drawbacks, especially a non-exhaustive list of topics discussed. On balance we find it a good match for this application and have also used it in Chapter 7.

Word2vec model

Here we briefly describe the word2vec model trained for topic detection. Training a set of embeddings on this corpus rather than using pretrained ones was intended to a) ensure the model was in-distribution for the radio data, as one trained on web text would not have been, and b) allow it to take advantage of the phrase detection described in Section 2.5. The trained embeddings were 300-dimensional and used a skipgram objective with negative sampling and a context window of 8 words. Words

[2]Namely: The prevalence of unmarked advertising in the corpus, and the relatively small amount of airtime devoted to politically important topics like abortion relative to, e.g., the weather.

with fewer than 50 occurrences were ignored. Training was on the entire corpus, without a holdout set. We used the popular Gensim package [75] to train the model.

We evaluated the trained word2vec model not with an eye toward state-of-the-art performance, but with the goal of ensuring enough information had been learned to be useful in this application. To that end, the model's average performance of 38.7% on the Google analogies test set [76] and 39.0% on the MSR analogies test set [77] are reassuring.[3] The same results provide further confirmation that the corpus ingestion and transcription processes are working well.

Choice of topics

We hand-selected several topics of political interest (politics in general, the economy, healthcare, drugs, immigration, etc), each with an initial list of a few hand-selected seed terms. For example, the "climate" topic included seed terms "climate", "climate change", "sustainability", "global warming" and "environment." These topics were intended to cover a wide range of radio discussion, including topics for weather and sports, while preserving a focus on national political issues. The full lists of keywords associated with each topic are listed in Appendix B. The topics themselves with some metadata are given in Table 4.1.

In specifying a set of topics, we had to contend with a (loosely binding) trade-off. More general topics make for broader coverage of the universe of radio discussion, while more specific topics are more likely to uncover small-scale structure. The approach taken in this chapter (and in the analysis of Chapter 7 using the same topics) is to err on the side of breadth. For a first look at the question, we wanted to capture generally what subjects are discussed on the radio; this does, however, mean that we may miss differentiation, diversity or even polarization at smaller scales.

For example, it's conceivable that some groups of stations or shows approach their discussion of political and social issues quite differently from others. One might focus heavily on politics in conjunction with abortion, guns and other social issues, while another stresses politics and economic issues. Our choice of topics is suited to examine this sort of difference between stations and shows, while more fine-grained breakdowns would get at other ways content might differ.

Other projects focused on manual coding have taken similar approaches. The Comparative Agendas Project's codebook [78] for manual content labeling of US political discourse contains 20 top-level topics, though each one does have an average of 11 subtopics. Research using these codes does sometimes use only the top-level categories (e.g., [79]).

4.2 Results

The most striking fact about these results is radio's high degree of homogeneity in topic distributions, especially compared to the large differences observed for schedule

[3] Compared to a model trained on web text, lower performance should be expected here. Despite the large size of the corpus, radio discourse covers fewer topics than internet content.

Topic	No. of Seed Terms	No. of Expanded Terms
Climate	5	96
COVID-19	7	52
Crime	5	105
Drugs (incl. opioids)	5	62
Economy	4	48
Education	7	124
Guns	4	27
Healthcare	7	35
Immigration	3	49
Inclusivity	6	43
Other social issues	9	57
Politics	9	259
Sports	6	472
Trump	7	7
Weather	6	106

Table 4.1: The full list of topics used in our topic similarity analysis, together with the number of seed and expanded terms. The Trump topic was not expanded as the others were, because of our interest in tracking mentions of Trump in particular (rather than closely related words like "president").

similarity. The largest differences observed here, whether between groups of stations or of shows, are smaller, and the baseline level of similarity is notably higher.

4.2.1 Show-level

Radio shows display a surprisingly high degree of homogeneity in their topic distributions. Over all 847 shows in our corpus, the average pair of shows have a cosine similarity of fully 0.767, with a standard deviation of 0.173. But most shows are even more similar than this: the median is 0.813. The full distribution is shown in Figure 4-1.

This pattern stands in contrast to the low degree of similarity between stations when comparing their schedules. While stations display some diversity in their choice of shows, those shows' choice of topics has a clearly identifiable mainstream.

Moreover, and in further contrast to schedule similarity, these shows' topic similarities do not display much between-group variation across a number of important real-world groups. Except for being linked to a Twitter handle, which reflects our sampling choices, the largest difference in Table 4.2 across one of these variables is about 0.070. The average pair of public radio shows are an average of 0.065 less similar in their topic mixes than the average pair of talk radio shows. A similar gap is observed for syndicated shows vs local shows within public radio (0.069), and the syndicated-vs-local difference is smaller.

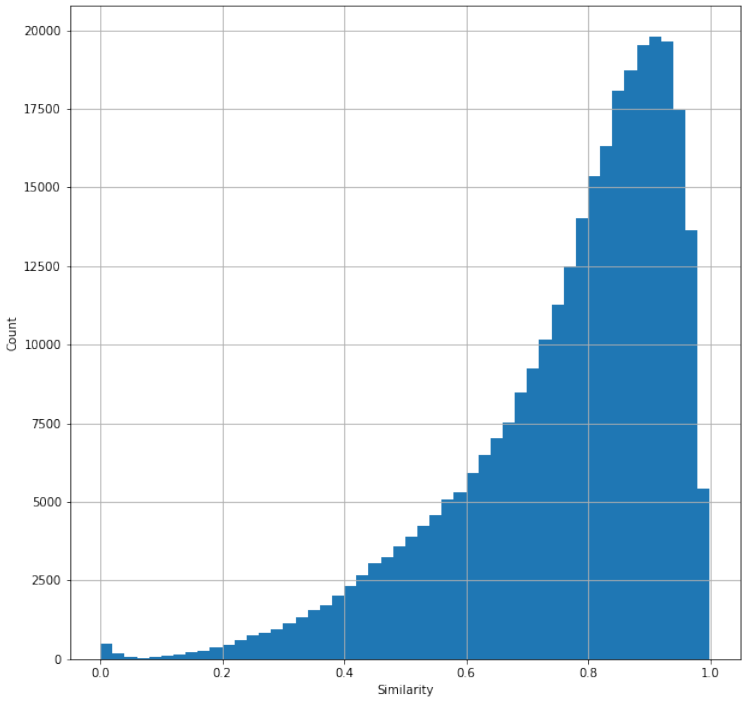


Figure 4-1: The distribution of topic similarities between all pairs of shows in the corpus is shown.

Strikingly, however, our sample of Twitter-matched shows[4] does appear to be more homogeneous in its topic mixtures than radio overall. The average two shows with Twitter handles have a similarity of 0.86, vs 0.76 for two shows without.

This effect becomes much larger among only public radio shows, where shows with Twitter handles express a near unanimity on topic mixtures at 0.935, vs 0.721 for shows without. The most likely cause here, though we can't say for sure, is NPR and PRI's role in production and distribution. Of 14 public radio shows in our Twitter-matched sample[5] most of them and all of the largest ones are produced by NPR, PRI or their flagship affiliate stations in the Acela corridor. For example, while Open Source is not an NPR program, it's produced by WBUR, the same station that produces NPR's Here and Now. Smaller public radio shows (like First Coast Connect, a local public show in Florida) are more likely to be locally staffed and produced.

For a slightly more intuitive sense of how large these gaps are, we'll note that the similarity scale here, with all nonnegative entries, only ranges from 0 to 1. The standard deviation (σ) of the distribution of similarities is 0.173, so all but the last difference mentioned (Twitter vs non-Twitter within public radio shows) are less than

[4]We can't say exactly how likely shows without Twitter handles recorded are to actually have any, not having checked for them. From searching for handles, however, it's clear that a) many but not all local shows have a Twitter presence, while nearly all syndicated shows do and b) local shows' handles are harder to find.

[5]Morning Edition, All Things Considered, Weekend Edition, Fresh Air, Here and Now, 1A, The World, On Point, Planet Money, It's Been a Minute, Radio Lab Invisibilia, City Visions, Open Source, and First Coast Connect.

1σ. The next largest, at 0.10, is only about 0.6σ.

Variable	Both In Group	Both Out of Group
Public-radio show	0.730	0.795
Syndicated show	0.796	0.747
Has Twitter handle	0.861	0.762
Syndicated show (public radio only)	0.762	0.693
Syndicated show (talk radio only)	0.829	0.776
Has Twitter handle (public radio only)	0.935	0.721
Has Twitter handle (talk radio only)	0.851	0.793

Table 4.2: The average cosine similarity of radio shows' topic mixtures across several cross-cutting variables. These are population figures, so no standard errors are shown. The greater similarity between stations with different owners shown for public radio (i.e., greater diversity among an owner's stations than for all stations) may reflect a desire by station owners to diversify their holdings. See further discussion in the text.

4.2.2 Station-level

Stations display an even greater degree of similarity in topic distributions than the shows they air. The average pair of stations in our corpus have a similarity of 0.857, with a standard deviation of 0.14, and once again the median is higher at 0.893. The full distribution is shown in Figure 4-2.

This result provides an even more direct contrast with the schedule similarity of the stations. While they air widely differing sets of shows, the subject matter of those stations at the level of broad topics is well aligned. The perspectives of the shows they air may be less so, of course, but we are not set up to carry out this more fine-grained analysis.[6]

As in the case of shows, the most striking fact here is topic similarity's lack of relationship to station attributes. With one exception, the largest between-group difference Table 4.3 recorded on this 0 to 1 scale is 0.04, a far cry from the gap of nearly 0.30 in schedule similarity by owner.

The exception, for ownership differences in public radio, may reflect the same apparent tendency for public station owners[7] to seek diversity in their stations' programming we saw in Chapter 3. It is, however, based on fairly few stations: as noted there, only five owners in our sample operate more than one public radio station,

[6] As above, though, note that some differences in perspective could translate into differences along these topics. One hypothetical example: a socially conservative host, who discusses guns, abortion and other social issues more than the economy, vs a business-focused conservative with the reverse set of interests.

[7] Public radio stations, despite the name, are not all operated by a central government entity. They have owners with FCC licenses just as talk stations do, though those owners usually syndicate a great deal of programming from NPR or PRI.

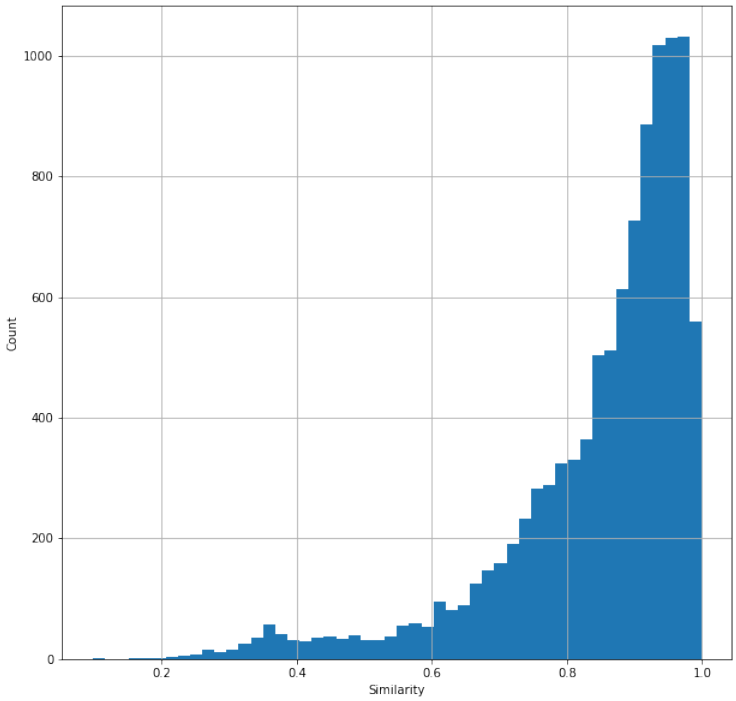


Figure 4-2: The distribution of topic similarities between all pairs of stations in the corpus is shown.

and none of them has more than three stations. Moreover, even if it reflects a trend in the population of public radio stations broadly, we can't say without further research what causes it. (Station owners might seek out stations which already have a greater diversity of topics discussed, or push their stations to differentiate themselves from each other.)

4.2.3 Discussion

Recall that our topics are focused on politics, and are not exhaustive of what's on the radio. Stations and shows may still differ systematically in attention paid to non-political content, or in their perspectives on these political topics.

Nevertheless, the clear degree of homogeneity here, across geographies, frequency bands, syndication status and even station format is strong evidence of a national mainstream of radio discussion. On public radio, this mainstream is expressed most strongly in centrally syndicated content (which our Twitter-matched public shows skew toward). The much greater homogeneity of these large NPR/PRI shows provides evidence that the public syndication networks not only homogenize stations' programming but also their content. Talk radio, by contrast, has much less of a topic-diversity gap between syndicated and local programming.

Whether and how the national mainstreams of public and talk radio tie into the rest of the media is the subject of Part III.

Variable	All	Public	Talk
Same Owner	0.881	0.700	0.882
Different Owners	0.844	0.830	0.860
Same Band	0.858	0.825	0.871
Different Bands	0.834	0.842	0.837
Same Census Region	0.854	0.818	0.876
Different Census Regions	0.845	0.833	0.859

Table 4.3: The average cosine similarity of topic distributions within and between station groups formed by several cross-cutting variables. These are population figures, so no standard errors are shown.

Part III

Radio-Twitter Interface

Chapter 5

Social Structures of Twitter and Radio

This chapter presents our investigation of the common social structures present in Twitter and radio. The focus is on demonstrating that both manifest social or information-spreading networks among media elites. While these networks are clearly present on Twitter, demonstrating that they exist and have similar structure on radio helps demonstrate that medium's integration into the broader media ecosystem.

5.1 Latent Ideology

Obviously one of the primary dimensions of variation in radio is ideological. Very conservative hosts, like Rush Limbaugh, are clearly different from more liberal hosts, like many of those on NPR. We want some Twitter-based way to infer the ideological leanings of hosts and shows, or at least those hosts and shows we've matched to Twitter accounts. Such estimates can then be compared to radio-side and common-sense indicators of ideology.

5.1.1 Methodology

A number of methods exist for inferring ideological or political orientation from Twitter data. A natural first approach to the problem might attempt to use the text of tweets to train classifiers, which can yield good performance [11] (though it can also be unexpectedly difficult [80]). We, however, have no labeled training data and need an unsupervised approach.

Several such approaches have been developed which rely on the follow graph. The idea is that, because users choose whom to follow partly based on ideology, the homophily in their following behavior encodes a latent ideological dimension into the graph. The most popular method for extracting it is based on Bayesian ideal point estimation [2], and unfortunately is lacking an off-the-shelf implementation.

We've instead adopted a different and simpler approach [81] [82] which gets at the same goal. We base our ideology measure on all Twitter users who follow two or more

radio personnel:

1. Form the bipartite graph between these users, on the one hand, and the radio staff, on the other;

2. Using the adjacency matrix of this graph, with each following user as a row and each radio staffer as a column, compute the cosine similarity matrix between the radio users.

3. Reduce the similarity matrix down to two dimensions via classical multidimensional scaling. One of the resulting dimensions should be interpretable as ideology.

We also converted the resulting scores back to the show level, using the mapping between Twitter users and shows, for use with radio data.

5.1.2 Results

The measure of latent ideology extracted according to this method lines up well with real-world notions of the ideology of radio hosts. Talk radio hosts have higher values than public hosts (i.e., higher values are more conservative), and within both segments of radio, ideology scores are well aligned with hosts' reputation for being conservative or progressive.

We've shown radio users' accounts, plotted along both dimensions of the final scores, in Figure 5-1. The users with the highest (most conservative) and lowest (most liberal) scores are shown in Table 5.1, and fit an intuitive definition of ideological extremes in radio.

User	Ideology Score
@MarkLevinShow	1.680
@DLoesch	1.636
@SeanHannity	1.551
@BuckSexton	1.547
@LarryElder	1.543
@bridgetkelley	-1.251
@nprkyoung	-1.263
@WatsonCarline	1.273
@dyerworld	-1.283
@andrea_c_hsu	-1.285

Table 5.1: The five highest and five lowest Twitter accounts according to our ideology scores. The top five are all very conservative talk hosts, and the bottom five are NPR hosts or production staff.

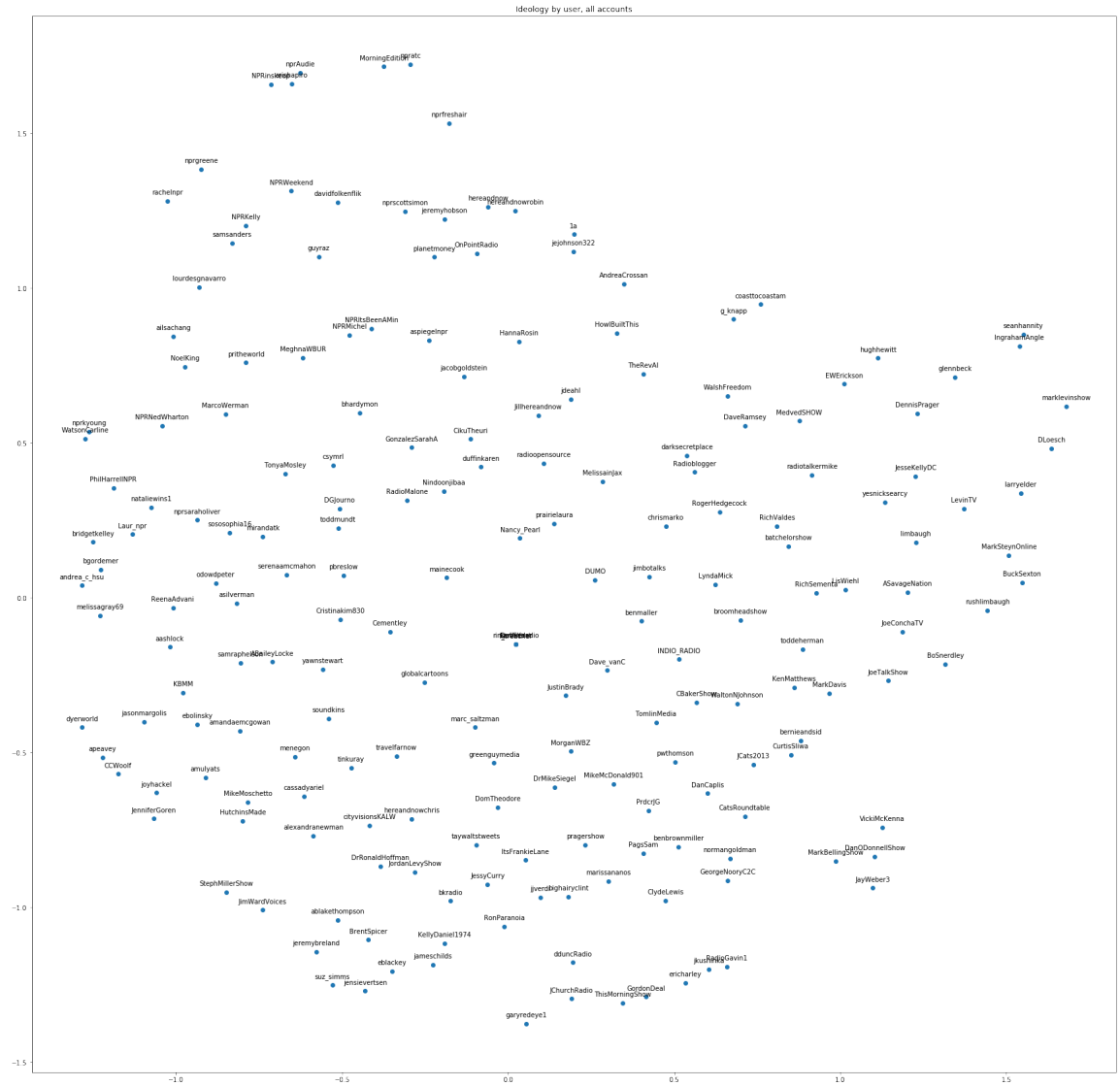


Figure 5-1: Twitter accounts of all radio hosts and staff, plotted along two latent dimensions in the follow graph. The dimension on the horizontal axis is interpretable as ideology, while the dimension on the vertical axis doesn't have a clear interpretation.

5.2 Follow-Graph Communities

Related to but distinct from ideology in radio is the notion of a community. We avoid relying on too precise a definition of community, using the term instead to mean any socially cohesive group of similar shows. Indeed, exploring which community structures exist and what they correspond to is the point of this section.

5.2.1 Methodology

If there is similar social structure between radio and Twitter, the community structure between hosts and shows should be most readily observable on Twitter. This is because of the more comprehensive social graph we've recorded among the "elite Twitter" segment: including other users should make the graph more connected and communities easier to discover than on radio. Accordingly, we focus on community detection in the Twitter follow, mention and retweet graphs.[1] The reply graph among our elite Twitter segment is not included because it was too disconnected, containing too many singleton nodes, to yield useful communities.[2] Each of the three graphs used contains a directed edge from user A to user B if A respectively follows, mentions, or retweets B.

As for how to detect the communities themselves, there is a rich literature on this topic. Papadopoulos et al [83] provide a useful if by now slightly dated survey. We elected to use the popular Louvain algorithm [84] for its good general-purpose performance. An important consideration in doing so was the lack of strong priors about the nature of the communities, posing difficulties for model-based community detection methods like stochastic block models. Because the Louvain algorithm operates on undirected graphs, we discarded edge direction information in detecting communities.

5.2.2 Results

This process yields four communities, which are readily interpretable in radio terms: most conservative users are in the same community, with identifiable smaller groups of non-conservative users in other communities.[3] Here and in later chapters, however, we've elected to refer to the communities by arbitrary numbers rather than names, in order to avoid over-identifying them with our interpretations. We summarize the communities' ideological leanings in Table 5.3, and present a random sample of users from each community in Table 5.2. The entire graph is shown with color coding by community in Figure 5-2. In brief, based on these data, we might interpret the communities as follows:

[1] More specifically, the quotient graphs of all three after collapsing together Twitter accounts associated with the same radio show. This choice was motivated by a small number of users who shared a radio show but ended up in different communities in the raw Twitter graphs.

[2] This may be an indication that the elite Twitter universe should have been larger to capture more social context.

[3] For instance, public radio hosts and the small number of liberal talk hosts are in separate communities.

Community 0 A relatively ideologically balanced set of users, including just a few radio hosts (cf Figure 5-2). The mean ideology for radio staff is in the middle of the distribution at -0.09.

Community 1 A liberal community incorporating many NPR staff, on the radio side, as well as a number of liberal journalists and especially New York Times people. The New York skew is visible in Table 5.2: all three of the non-radio users listed are New York Times employees. The radio staffers, intriguingly, are less easily geographically located: though one clearly lives in Washington, DC, the other two have work histories as foreign correspondents. This community has a quite liberal average ideology score at -0.529.

Community 2 A liberal community similar to community 1, but with a skew toward DC and the Washington Post. One can also see this skew in Table 5.2: listed are two members of Congress, a Washington Post employee, two NPR staffers who live in DC, and the official account of Morning Edition, produced by a DC-based NPR station. The rare non-conservative radio hosts who are not on NPR also end up here; while it's not visible in the table, Stephanie Miller (@StephMillerShow) is in community 2. This community's average ideology score is even further left at -0.607.

Community 3 The main conservative community, with the highest ideology score (higher, again, means more conservative) at 0.605. There's no apparent geographical skew to this community.

The geographic structure visible in these communities most likely has something to do with our choice of users for the "elite Twitter" universe. Recall that these users included (among others) a number of New York Times and Washington Post reporters, who can be expected to have more connections to fellow Times or Post employees than to those at the other paper.

Radio	0	1	2	3
True	jejohnson322 benmaller 1a	DGJourno guyraz lourdesgnavarro	nprkyoung MorningEdition ailsachang	INDIO_RADIO larryelder bkradio
False	AnnysWaPo chloe_meister AdamSerwer	meslackman declanwalsh FloFab	RepJoeCourtney pastpunditry RepRubenGallego	FreedomWorks JimInhofe RepAlexMooney

Table 5.2: A random sample of six elite Twitter users from each of our four follow graph communities. We've oversampled radio users, with three radio users and three other users. Users' Twitter bios are omitted for space reasons, but users in communities 0, 1 and 2 are mostly NPR accounts, and community 3 users are conservatives.

Interestingly, without the broader context of elite Twitter, community detection is much less useful. The graph of show hosts and personnel only is not even connected,

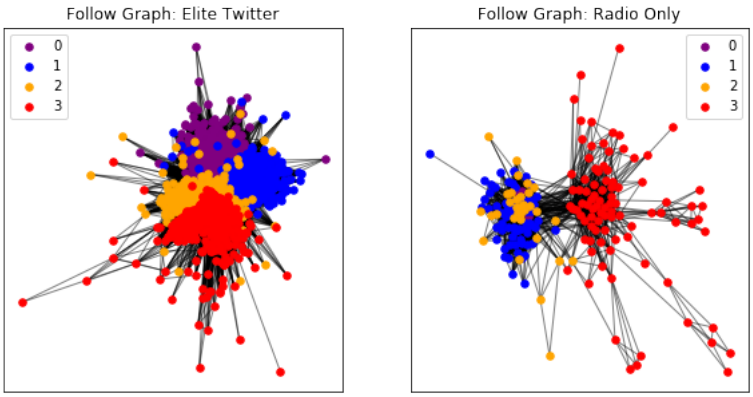


Figure 5-2: The elite Twitter follow graph, color coded by Louvain community, on both the entire elite Twitter segment (left-hand panel) and restricted to accounts of radio staff only (right-hand panel). These plots are spring layouts of the graph, ignoring edge directions. The color codes are consistent between the two subplots. We've dropped 11 isolates with no follow connections to other radio users from the right-hand panel to make it easier to visualize. Community 3 is the main conservative community and is appropriately colored red; communities 1 and 2 are respectively New York- and DC-centered liberal communities; and community 0 is both more ideologically balanced and poorly represented on the radio. Finally, note the right-hand panel's resemblance to the coairing graph shown in Figure 5-3.

with several users left as isolates, and communities based on the largest component are much less informative. We find this fact to be a convincing example and demonstration of our broad thesis: radio is embedded in the broader media ecosystem, and more easily understood with that context.

Community	Avg. Ideology
0	-0.09
1	-0.529
2	-0.607
3	0.605

Table 5.3: The average ideology score by follow-graph community for radio staff. Higher scores are more conservative.

Next, and perhaps most interesting of all, the communities detected are approximately consistent across follow, mention and retweet graphs, with in particular a core group of highly central conservative users in the intersection of a specific community from each graph. We haven't depicted the mention and retweet graphs in as much detail as the follow graph, but this situation is summarized in Table 5.4 and Table 5.5. The fact that conservative discourse patterns appear to line up with conservative social structure as reflected in the follow graph is suggestive. It is likely a very literal demonstration of the widely discussed epistemic closure of right-wing media [13].

Mention Community Follow Community	0	1	2	3	4	5	All
0	1	0	2	0	0	0	3
1	0	0	2	7	0	0	9
2	0	2	6	1	0	0	9
3	0	2	41	0	1	0	46
All	1	5	51	8	1	1	69

Table 5.4: A crosstab of follow community membership against mention community membership at the radio show level. Note in particular that most conservatives (in follow community 3) are also in mention community 2.

Retweet Community Follow Community	0	1	2	3	4	6	8	9	All
0	1	0	0	1	0	0	0	0	3
1	0	1	4	4	0	0	0	0	9
2	0	2	0	6	0	0	0	0	9
3	0	0	0	3	35	1	1	1	46
All	1	3	4	14	35	1	1	1	69

Table 5.5: A crosstab of follow community membership against retweet community membership at the radio show level. Note in particular that most conservatives (in follow community 3) are also in retweet community 3.

5.3 Follow Graph vs Co-Airing Graph

To provide a more direct test of the similarity of social and information structures between radio and Twitter, we compared Twitter's follow graph to the most closely analogous network that can be observed in radio: the co-airing graph between shows.

5.3.1 Defining the Co-Airing Graph

What we call the co-airing graph is an undirected graph formed over the set of shows, where two shows are connected if there is any station which airs both of them within a week of each other. Because most stations schedule their programs on a weekly basis, the requirement to air within a week of each other should capture real co-airing while minimizing the number of spurious edges introduced by data problems. We consider this graph to be the closest observable equivalent to Twitter's follow graph for two reasons, paralleling the twin social and informational purposes of Twitter [50]:

- Hosts of shows which air on the same station, particularly if they are both local shows, are more likely to be professionally connected to each other than randomly selected pairs of hosts;

- Stations which choose to air both of a pair of shows clearly think both shows will be of interest to their audience, implying commonality of subject matter or viewpoint.

This graph has 847 nodes, one for each show in the corpus, and 24,338 edges, with only about 6.8% of all possible edges existing. There are two connected components, one containing 835 shows and the other containing 12.[4] The larger connected component is shown in Figure 5-3.

In the results subsection, we focus on the subgraph of shows matched to Twitter handles in order to make a more direct comparison to Twitter. These two graphs are both on the show level (each node is a show); for Twitter data, it's converted to the show level by connecting two shows if any pair of users, one from each show, are connected.[5] We omit nine shows which are isolates in the Twitter graph (their associated Twitter users have no follow edges to or from any other shows), leaving 58 shows.

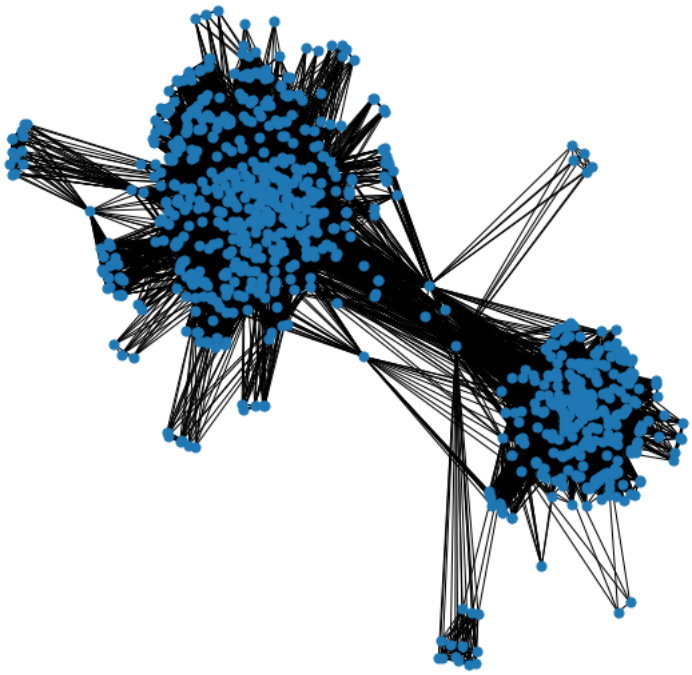


Figure 5-3: The giant component of the co-airing graph of radio shows, comprising 835 of 847 shows. Note the resemblance to the structure of the radio-only follow graph in Figure 5-2.

[4]The 12 shows not connected to any other are all from WWJ-AM in Detroit, which airs only content presented by local staff. Coincidentally, WWJ happens to be the world's oldest commercial radio station.

[5]That is, the quotient graph under the equivalence relation of being on the same show.

5.3.2 Results

Basically, the co-airing graph of shows and the Twitter follow graph display surprisingly similar structure. They have comparable values for many graph properties: degree distributions, connectivity, clustering coefficients, and others. The degree distributions are shown in Figure 5-4, while several other properties are summarized in Table 5.6. These indicators do reveal some differences between the two graphs – essentially, the coairing graph is more densely connected – but not radical ones.

A striking illustration of the point is that, while Figure 5-3 does include shows not matched to Twitter, the two graph structures as shown there and in the right panel of Figure 5-2 look quite similar. Meanwhile, the fact that the follow graph on all of elite Twitter (in the left panel of Figure 5-2) does not resemble the radio coairing graph reinforces the point we made in Section 5.2: radio has a distinct social structure, on the air and online, but this structure is embedded in a broader context.

Statistic	Follow	Coairing
Order (# of nodes)	58	58
Size (# of edges)	381	449
Average degree	13.1	15.5
Transitivity	0.524	0.659
Average clustering coefficient	0.595	0.768
Size of largest component	57	58

Table 5.6: Selected summary statistics of the follow and co-airing graphs are shown for the set of Twitter-matched shows.

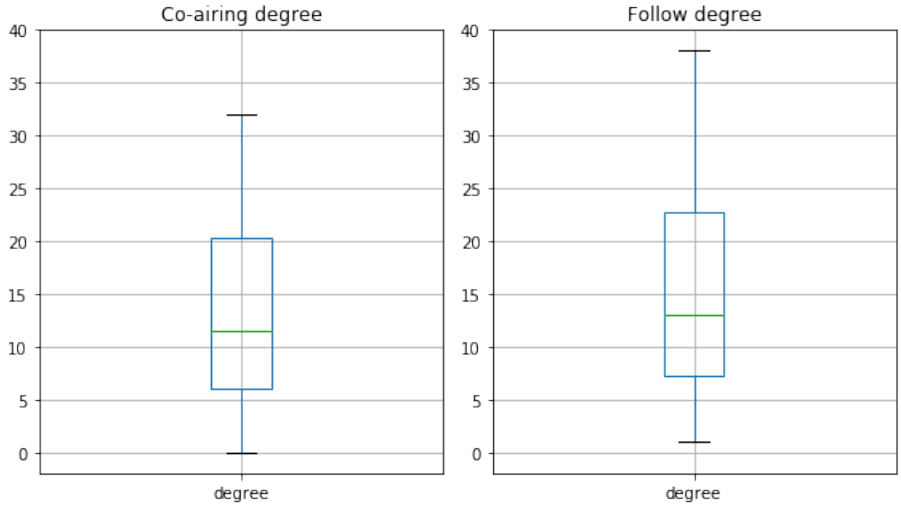


Figure 5-4: The degree distributions of the show-level follow and coairing graphs are depicted. On only 58 data points, a boxplot gives a better sense of the distribution than a histogram.

More readily interpretable and socially meaningful properties also fit the pattern. The Louvain communities detected in the two graphs are quite similar, and most shows are put into clearly corresponding communities in the two graphs. This situation is depicted visually in Figure 5-5.

Now, Figure 5-5 collapses all but the largest community in both graphs together. Because these largest communities are also the most conservative communities by ideology score, the plots also show us that conservative shows are the most insular: they overlap the least with other communities of shows in both station schedules and Twitter follow relationships. Paralleling Kwak's delineation of social and information-spreading roles for Twitter [50], this insularity has both social dimensions (the apparent social structure of conservative hosts on Twitter includes fewer connections to non-conservatives) and information-spreading ones (listeners to conservative stations are less exposed to other perspectives).

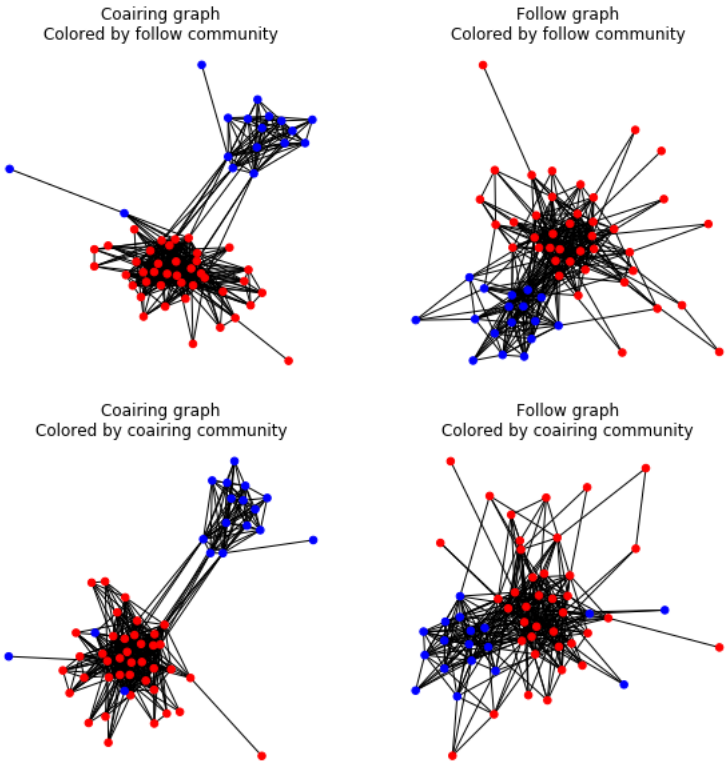


Figure 5-5: Spring layouts of the show co-airing graph and the Twitter follow graph, both on the show level and each colored by communities in both graphs. In both follow and coairing graphs, only shows matched to Twitter are shown. For comparability with the coairing graph, which is undirected, the follow graph is laid out ignoring edge directions. In all four subplots, the red nodes are the largest follow-graph or coairing-graph community, and the blue nodes are nodes in all other communities. We've adopted this color scheme because in both the follow and co-airing graphs the largest community is the conservative community.

The results suggest quite clearly that radio manifests the same underlying social

structure both on the air and in its presence on Twitter.

5.4 Twitter vs Ownership

We compared station ownership to Twitter-side variables, namely follow-graph communities and ideology. (Note that this means we are only considering the set of shows matched to Twitter accounts.) To reduce sensitivity to station-specific effects, we included only the largest five owners[6] which each own many stations.

The picture that emerges is one of owners whose programming closely replicates a national mix of discourse communities, but have more freedom to choose within them. Comparing the fraction of each owner's airtime that goes to each follow community, owners are surprisingly similar and close to the overall average. With one exception (Entercom airing the Al Sharpton Show rather than another conservative talk show), the largest owners' mix of shows closely matches the overall distribution of follow communities. This mix of shows is compared to the overall distribution in Table 5.7.

Owner	Follow Community	Community % of Owner	Community % of Talk Radio	Diff.
Cumulus Media	3	100.0	98.8	1.2
Entercom	3	85.8	98.8	-13.0
Entercom	2	14.2	0.8	13.3
Salem Media	3	100.0	98.8	1.2
Townsquare Media	3	100.0	98.8	1.2
iHeart	3	98.6	98.8	-0.2
iHeart	2	0.9	0.8	0.0
iHeart	0	0.5	0.4	0.2

Table 5.7: The breakdown by Twitter follow community of airtime for the Twitter-matched shows aired by five large owners. The follow community numbering is arbitrary; #3 is the largest and contains most conservative hosts. Percentages have been truncated to the nearest 0.1%.

But their choice of shows within each community shows considerable variation, especially by ideology (Table 5.8). The average ideology of an owner's programming within each follow community can vary notably from overall averages, in either direction. For example, Entercom's conservatives are less conservative than average, while iHeart's non-conservatives are significantly less liberal than average. These differences also lead to large variation in average ideology between owners, as can be seen in Figure 5-6.

Though community #3 accounts for most of the airtime, the stability of community fractions across owners helps confirm that our Twitter communities do correspond to a real social entity. Meanwhile, within communities, the large differences in ideology

[6]Cumulus Media, iHeartMedia, Townsquare Media, Salem Media, and Entercom.

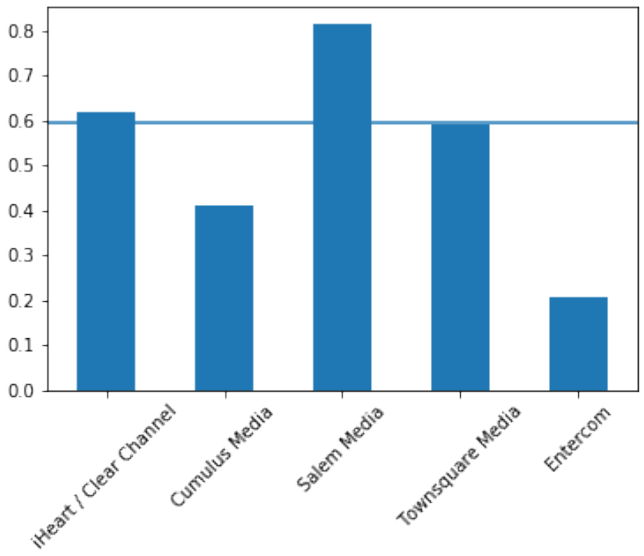


Figure 5-6: The average ideology of airtime devoted to Twitter-matched shows (i.e., weighted average by length of program) for five large station owners. The horizontal line indicates the average value for all non-public talk radio.

by owner suggest the possibility that different owners' syndication choices reflect considered ideological opinions of the owners (cf the clear station-level perspectives in Section 8.3).

5.5 Twitter vs Geography

We compared station geography to the same Twitter-side variables as for ownership. This time, however, we included not only all owners but both public and talk radio. The conclusions are quite different from Section 5.4: geography at the level of Census region has little relationship to either follow community or ideology.

As with owners, each region's mix of follow communities is similar to overall averages, as can be seen in Table 5.9. The magnitude of the differences is larger than for owners, but this shouldn't be too surprising. We've included public radio (and all owners) in these comparisons, so there will be some effect of station selection.

Regional differences in ideology within those communities, shown in Table 5.10 are an order of magnitude smaller than the largest ones for ownership in Section 5.4. Note however that these Twitter metrics represent a universe of mostly syndicated hosts, and so it's difficult to generalize from them to local content. (One should also keep in mind that syndicated content is the large majority of airtime, though.)

To help the reader visualize the distribution of ideology, we've mapped station broadcast areas with an ideology color-coding in Figure 5-7.

On balance, the story these charts and tables tell is clear, and it's the same story as in Part II. The most influential entities in the radio ecosystem are centralized ones like corporations, with local influence not entirely absent but distinctly secondary.

Owner	Follow Community	Ideology	Community Avg.	Diff.
Cumulus Media	3	0.832	1.066	-0.234
Entercom	3	0.851	1.066	-0.214
Townsquare Media	3	0.868	1.066	-0.197
Entercom	1	-0.409	-0.407	-0.001
Cumulus Media	0	0.197	0.193	0.004
Entercom	0	0.197	0.193	0.004
iHeart	3	1.095	1.066	0.029
Salem Media	3	1.158	1.066	0.092
Cumulus Media	1	-0.226	-0.407	0.181
iHeart	0	0.401	0.193	0.207
iHeart	2	-0.132	-0.683	0.551
Entercom	2	0.407	-0.683	1.090

Table 5.8: The average ideology of Twitter-matched content, by follow community by owner for five large owners. Higher ideology scores are more conservative. Ideology values have been truncated to the nearest 0.001.

Census Region	Follow Community	Community % of Region	Community % of All	Diff.
Midwest	0	1.1	1.5	-0.4
Northeast	0	2.0	1.5	0.5
South	0	1.5	1.5	0.0
West	0	1.3	1.5	-0.2
Midwest	1	3.5	5.1	-1.6
Northeast	1	7.7	5.1	2.6
South	1	4.0	5.1	-1.0
West	1	5.8	5.1	0.8
Midwest	2	4.4	9.5	-5.1
Northeast	2	12.2	9.5	2.7
South	2	8.3	9.5	-1.2
West	2	12.3	9.5	2.8
Midwest	3	91.0	83.9	7.1
Northeast	3	78.1	83.9	-5.8
South	3	86.1	83.9	2.2
West	3	80.5	83.9	-3.4

Table 5.9: The breakdown by Twitter follow community of airtime for the Twitter-matched shows aired by five large owners. The follow community numbering is arbitrary; #3 is the largest and contains most conservative hosts. Percentages have been truncated to the nearest 0.1%.

Owner	Follow Community	Ideology	Community Avg.	Diff.
Midwest	0	0.121	0.193	-0.072
Northeast	0	0.178	0.193	-0.016
South	0	0.240	0.193	0.047
West	0	0.137	0.193	-0.056
Midwest	1	-0.438	-0.407	-0.031
Northeast	1	-0.402	-0.407	0.005
South	1	-0.393	-0.407	0.014
West	1	-0.423	-0.407	-0.016
Midwest	2	-0.732	-0.683	-0.048
Northeast	2	-0.748	-0.683	-0.065
South	2	-0.667	-0.683	0.017
West	2	-0.646	-0.683	0.037
Midwest	3	0.990	1.066	-0.076
Northeast	3	0.992	1.066	-0.073
South	3	1.099	1.066	0.033
West	3	1.081	1.066	0.015

Table 5.10: The average ideology of Twitter-matched content, by follow community by Census region for the continental United States. Higher ideology scores are more conservative. Ideology values have been truncated to the nearest 0.001.

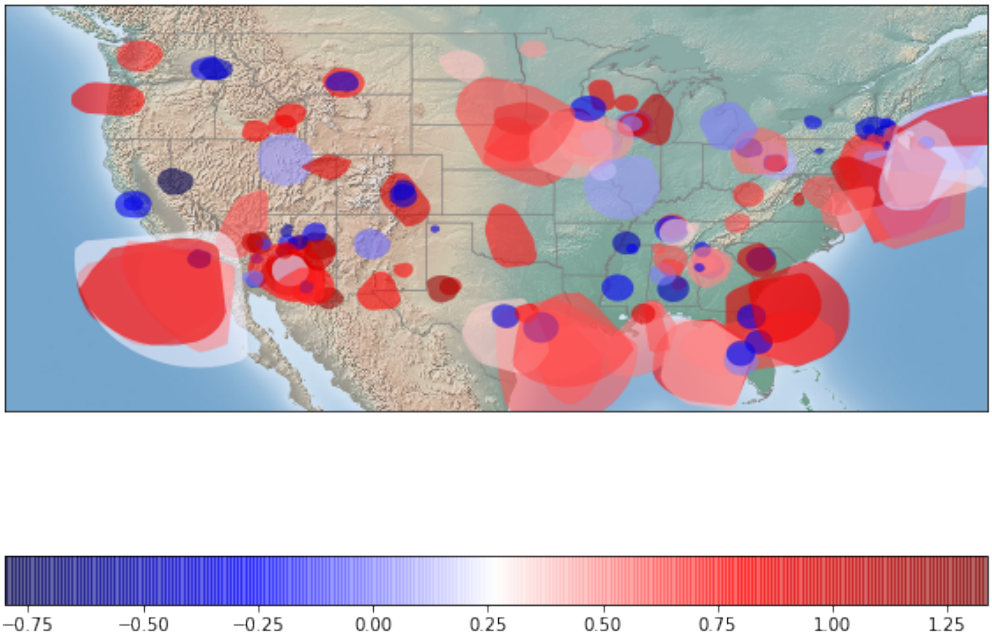


Figure 5-7: Station broadcast areas for all stations in the corpus, color-coded by the average ideology of their Twitter-matched programs. Higher values of ideology are more conservative.

Chapter 6

Social Structure and Radio Text

In this chapter, we examine whether the social structure mapped in Chapter 5 is related to the text of radio broadcasts. This is a central question from the public-opinion perspective discussed in Section 1.2. If elite discourse influences public opinion through media coverage, ultimately that influence must bottom out in actual coverage – in this case, discussion on air.

6.1 Methodology

Billions of words of radio text together are a very high-dimensional dataset. There are many ways to relate ideology, graph communities or any other piece of social structure to that text corpus, each of which would reveal different aspects of the relationship. We've elected to take perhaps the simplest approach, in attempting to predict the Twitter-side social structures from the radio text. Specifically, this analysis focuses on the estimated Twitter-based ideology and community of radio hosts. While it reveals less about discursive differences by community or along ideological gradients, if the Twitter-side variables are predictable it establishes a relationship quite well.

There is one snag, however: we want to avoid simply memorizing the link between host or show names and Twitter variables. (Learning that, e.g., "Rush Limbaugh" predicts a particular Twitter community is not very interesting.) To deal with this problem, given the n-gram based modeling approach discussed below, we simply built a list of excluded terms. The names of shows, hosts and stations were compiled, and fine-tuned over several fits of the models. The final list was kept out of the set of allowed predictors. In all, 245 terms were excluded - see Section B.4 for the details.

The various models we fit all had the same quite simple form: linear or logistic regression on n-gram features. We experimented with a number of other models, and went with this simple approach for certain benefits: easy interpretability, speed of experimentation, the ability to exclude the n-grams mentioned above, and last but not least to make the point that the even a simple model can detect these relationships.

The details are as follows:

Define the universe We began by considering every episode (i.e., a show-date-station combination) in the corpus for those shows linked to Twitter accounts.

We de-duplicated these by choosing the best episode for each show and date, while also imposing some quality filters. Specifically, the ASR and schedule-data confidence scores had to be above certain thresholds, and in a further effort to avoid schedule-data mistakes, the episode couldn't be too long or too short (at least 45 minutes and no more than five hours). The specific confidence thresholds were chosen to cut off the duration and confidence tails of the data, where most mistakes and misrecognized shows were, without sacrificing too much content. The final dataset included 3,850 episodes, all matched to at least one Twitter account.[1] We joined to the text data the show-level community and ideology estimates produced in Chapter 5.

Train-test split Split this universe into approximately 75% train and 25% test sets, with the randomization into train or test clustered at the show level (i.e., every episode of a given show ends up in the same fold).

Features We selected the top 20,000 uni- and bi-gram features across the dataset[2], ordered by term frequency, after excluding the list of n-grams described above, and applied a tf-idf transformation.

Model We fit a generalized linear model to these features, though exactly which kind of model depended on the kind of data. Ideology estimates were fit with linear regression, and communities with logistic or softmax regression.

Both feature processing and modeling used the scikit-learn package for Python [85].

6.2 Results

6.2.1 Ideology

Ideology estimates from Twitter are strongly predictable from the text of shows. We examined two specifications, one predicting continuous ideology scores and the other predicting a dichotomized version (in which the dependent variable was 1 if an episode's ideology score exceeded the sample mean, and 0 otherwise). The theory behind including a dichotomized version was that, while the ideology scores might not perfectly capture hosts' beliefs, they should at least be on the correct side of the left-right divide. If so, dichotomized scores should be as predictable as the original ones, and perhaps more so.

Both are, in fact, quite predictable. Modeling the continuous score yields out-of-sample $R\hat{2}$ of 0.664, while the binary classification model fits comparably well with an

[1] Without applying confidence and length thresholds, we would have had 5,537 episodes. A demonstration of the need for these quality filters is that the 3,850 episodes we include are 69.5% of all episodes, but account for only 40.4% of the total duration of those episodes. A few outliers reflecting bad webscrapes are responsible for much of the difference.

[2] Excepting those features which appeared only in 2020 data. This restriction was intended to focus the prediction task on persistent features of radio, rather than on details of early COVID-19 coverage.

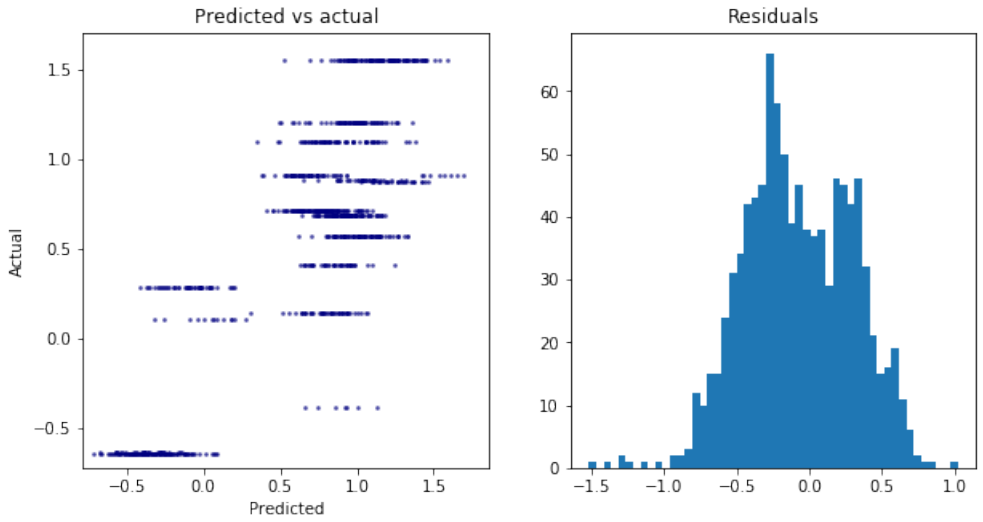


Figure 6-1: Diagnostics of a model predicting a show's ideology score from the show's text. Predicted values are plotted against actual, and a histogram of the residuals is shown. Actual values cluster along horizontal lines because the show ideology score assigned at the show level, and the unit of the data was a single episode.

AUC of 0.953. Predicted values versus actuals, as well as the residual distribution, for the first model are shown in Figure 6-1, while the ROC curve of the second model is presented in Figure 6-2. The confusion matrix for the binary model, in Table 6.2, does however suggest that the current approach could benefit from better calibration. Lacking an immediate need for well-calibrated scores, we decided not to fix these calibration problems. Some examples of the most predictive n-grams are shown in Table 6.1.

6.2.2 Graph Communities

Under several specifications, a show's community membership is quite predictable from its text. We've depicted ROC curves for one-vs-rest binary classification models, one for each of the four follow communities, in Figure 6-3. Some are more predictable than others[3], but all are at least somewhat predictable. The simple average AUC in this case is 0.865. Some examples of the most predictive n-grams are shown in Table 6.3.

6.3 Predicting Election Results

Communities and ideology are not the only things we've explored predicting from radio text. The election data from Section 2.4 provides a compelling prediction target

[3]It should also be noted that some, especially community 0, are much smaller than others and have less training data.

Highest	Lowest
comments on	noise support
extinction	m proud
our american	the_environment
threatening	make their
stand for	t sound

Table 6.1: Five n-grams randomly sampled from the most predictive of the continuous ideology scores. Because the models are linear and we centered and scaled the inputs, "most predictive" refers simply to having the most extreme coefficient estimates. Here we've selected five each from the 300 terms with the highest and the 300 terms with the lowest coefficient estimates. Terms with single letters as words generally reflect splitting on apostrophes - like "m proud" from the original text "i'm proud" - which an improved version of these models would avoid.

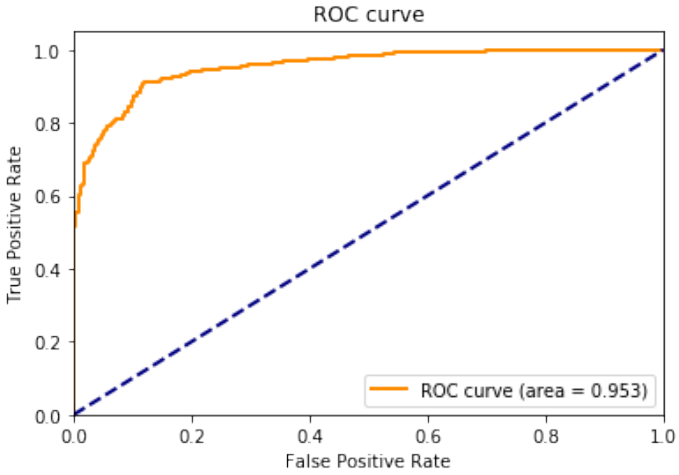


Figure 6-2: The ROC curve for a model predicting dichotomized ideology from show text.

as well.

In predicting vote counts or shares, the natural question to ask is how in touch radio shows are with their listeners. Is the content of radio closely related to listener partisanship, or is radio too nationalized for that? In particular, we want to compare measured listener partisanship to ideology: which is more closely related to radio? Doing so provides a clear test of Part II's conclusions about the nationalization and centralization of radio.

6.3.1 Methodology

The predictive methodology we use is, broadly, the same as in previous sections. We once again use linear regression on n-gram features, chosen in mostly the same way as

		Prediction outcome: 1	Prediction outcome: 0	Total
Actual value	1	361	194	555
	0	13	520	533
Total		374	714	

Table 6.2: The confusion matrix for a model predicting continuous ideology out of sample from show text. This test set contained 1,088 episodes of 17 shows. 1 corresponds to higher ideology score (i.e., more conservative) than the mean in the entire dataset.

above, including exclusions of certain terms from the set of predictors. In more detail:

Define the universe We considered two universes: a) shows matched to Twitter accounts and b) all local shows, defined as those we recorded appearing on only one station. These were not mutually exclusive, and as described in Section 2.2, some of the Twitter-matched shows were also local shows. The universe of eligible episodes for Twitter-matched shows was exactly the same as used above, while for local shows, we considered all available episodes without regard to recognition quality filters.

Train-test split As above, split the universe into approximately 75% train and 25% test sets, clustering at the show level.

Features Select the top 20,000 unigram and bigram features across the dataset, ordered by term frequency, after excluding a list of problematic n-grams, and apply a tf-idf transformation. The list of n-grams used here is a superset of the one used above, because more features might predict election returns without being informative about differences in discussion (for example, state and city names in addition to show names). We added an additional 279 n-grams, with the full list given in Section B.4.

Model As above, fit a linear model to the set of episodes with n-gram features.

But this time we face an additional problem: precincts as recorded in our NYT data are geographies with spatial boundaries, while radio shows are not. How can we convert the election data to be on the level of radio shows? Our answer was as follows:

0	1	2	3
the playoffs	to here	morning to	the_president
this game	e p	jacksonville	and get
we heard	w _n	podcast and	fox_news channel
in denver	geico	but he	eight hundred
is one	the series	for nine	the_left

Table 6.3: Five n-grams randomly sampled from the 300 most predictive for each follow community, which are identified by the same consistent but arbitrary community number we use elsewhere. Because the models are linear and we centered and scaled the inputs, "most predictive" refers simply to having the highest coefficient estimates. Though n-grams with the lowest coefficients are also highly predictive (of not being a community member), we haven't included them to save space.

- Matching precincts to the broadcast areas of radio stations, we calculated the number of Trump and Clinton votes cast within each station's broadcast area.[4] Precincts which crossed a station's coverage boundary had their votes split between stations according to the fraction of land area lying in each station's broadcast area. The result was a list of stations with the number of Trump and Clinton votes cast in each one's broadcast area.

- We converted these station-level numbers to shows according to the fraction of each show's airtime that came from a particular station. If, for example, show A had 33% of its airtime on station X, 50% on station Y and 17% on station Z, show A's vote counts would be the weighted average of the stations' counts, using those airtime fractions as weights.

- The dependent variable for the show-level prediction task described below was then the two-party Democratic vote share: Clinton votes / (Clinton votes + Trump votes).

This methodology has some shortcomings, notably that it doesn't account for the actual distribution of each show's listeners within the station broadcast areas. But because data on listeners' actual locations doesn't exist for many if any shows, and certainly doesn't exist at the precinct level, this method is a reasonable compromise.[5]

[4]Things are actually slightly more complicated than this: Stations, especially AM stations, can have different coverage areas at different times of day. When a station has more than one coverage area recorded, we took the simple average of vote counts for each.

[5]A more sophisticated approach might consider listener demographics, as estimated by Nielsen, in lieu of actual listenership location data when distributing votes.

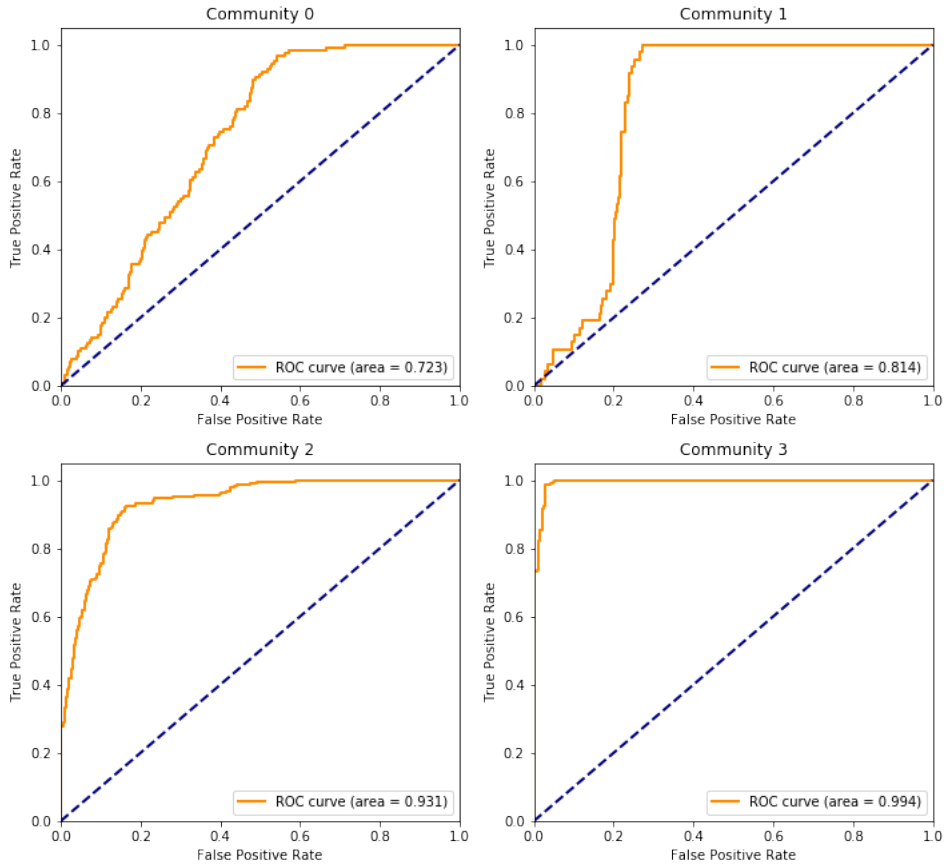


Figure 6-3: ROC curves for models predicting community label from show text. Four curves are shown, one for each community; the corresponding one-vs-rest models predict the given community's label (positive) against all others (negative).

6.3.2 Results

We fit two models: vote share as a function of show text for Twitter-matched shows and all local shows. (Recall as mentioned above that these groups are not mutually exclusive; about one third of the Twitter shows are local.) The results for these two models were quite different, in intriguing ways. Election results were not particularly predictable for Twitter-matched shows, with an R^2 of only about 0.10, vs 0.66 for ideology. As noted above, we are excluding more possibly predictive n-grams here than above: if we use the same list as in Section 6.2, the R^2 value rises to 0.15.

Local shows, however, were much more closely related to the partisanship of surrounding areas, with an R^2 of 0.51. (Using only the terms excluded in Section 6.2, this value is 0.63.) The results are depicted in Figure 6-4 and Figure 6-5. Some of the most predictive n-grams for both tasks are shown in Table 6.4.

The results practically call out for a closer look, but a full analysis of why this happens and how local radio is related to local political opinions will have to wait for future work. We can, however, still look back at some of the topic analysis done in Chapter 4. In that analysis, we compared shows by cosine similarity of vectors of

Twitter +	Twitter -	Local +	Local -
forty live	s news	since then	the lord
l	twitter at	conquered	you_think the
and feel	partner of	minutes we	biden and
in the_city	retirement wealth	these days	dr
christopher	surprisingly	and the	visit our

Table 6.4: N-grams randomly sampled from those most predictive of election returns. Because the models are linear and we centered and scaled the inputs, "most predictive" refers simply to having the most extreme coefficient estimates. Here we've selected five each from the 300 terms with the highest (denoted "+") and the 300 terms with the lowest coefficient estimates (denoted "-"), for both Twitter-matched and all local shows. Note in particular that the term "the lord" is a predictor of lower Democratic vote share.

topic mention rates (see Chapter 4 for more details). The average pair of local shows has a cosine similarity of 0.747, while the average pair of Twitter-matched shows has a similarity of 0.861. In other words, local shows overall display a greater diversity of topic mixtures, while our syndicated-skewed Twitter sample is more homogeneous.

Finally, one clear takeaway is that local radio where it still exists is in fact meaningfully local and at least somewhat in touch with community opinions. Because our sample of Twitter-matched shows skews toward syndicated content, these results provide a solid indicator of the homogenizing effect of syndication.

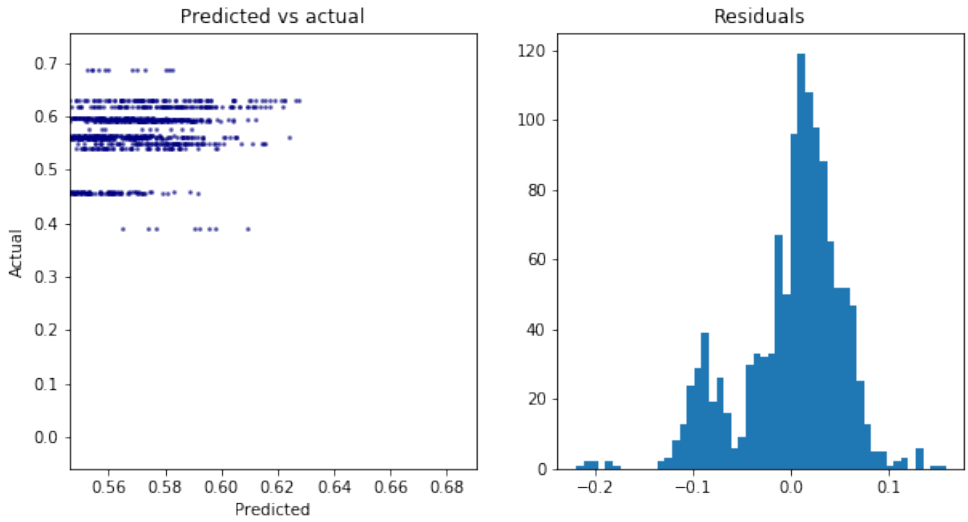


Figure 6-4: Diagnostics of a model predicting listening areas' 2016 election results from the text of Twitter-matched shows. Predicted values are plotted against actual, and a histogram of the residuals is shown.

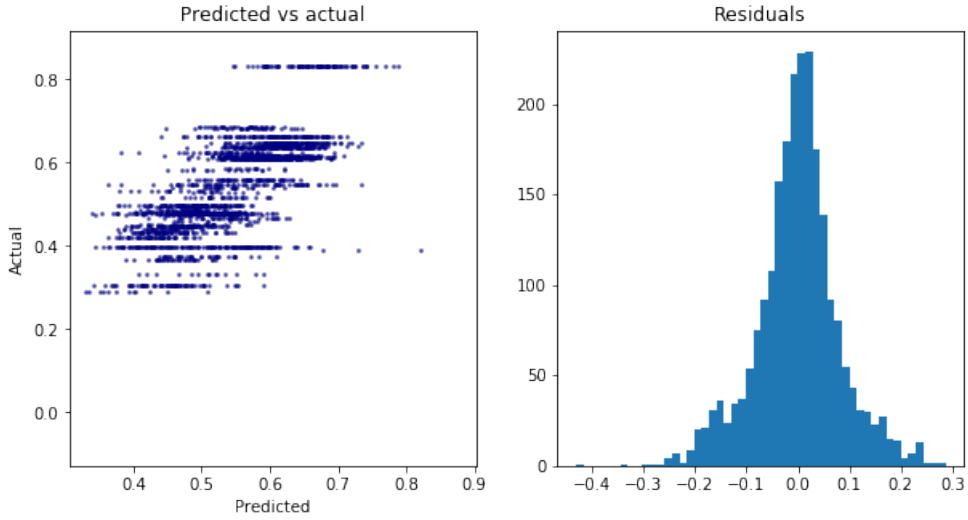


Figure 6-5: Diagnostics of a model predicting listening areas' 2016 election results from the text of local shows. Predicted values are plotted against actual, and a histogram of the residuals is shown.

Chapter 7

Radio Text vs Twitter Text

This chapter expands our analysis of linkages between Twitter and radio to a direct comparison of the contents of the two media. There are many theoretical reasons to be interested in such a comparison. In particular, in terms of the theoretical frame articulated in Section 1.2's discussion of public opinion, both radio and Twitter influence public views only through the information they present to readers or listeners. Any influence of common social structure, then, must bottom out in common patterns of content in order for them to have similar influences on public opinion.

7.1 Methodology

Comparing text corpora by automated means requires reducing their dimensionality. For this analysis, we effect this reduction through the same topic-modeling approach and set of topics referred to in Section 4.1. Much of the analysis in Section 7.2 also uses the same cosine similarity methods as in Chapter 4.

Recall also, from Chapter 4, that the topics we've chosen are quite general. In return for capturing a broader share of what's discussed on radio, we've accepted a less specific set of topics. Some differences in perspective should still be detectable (for example, the difference between social and economic conservatives who respectively stress social issues and the economy), but quite a bit of difference in opinion and even content can hide within each of our topics.

This approach, which prevents us from examining questions of sentiment, perspective or tone, is thus limited in important ways. But before considering the question of how radio and Twitter talk about the world, we should address how much they are in fact talking about the same world.

7.1.1 Modeling

We go beyond Part II, however, in considering questions of time. The goal is to identify relationships over time between radio and Twitter discussion, either overall or by various subgroups. Finding that one medium consistently leads or lags the other, for instance, would be informative and prompt us to consider causal questions.

Results from this analysis are presented in Section 7.3.

We adopt an econometric approach to these questions, using the tools of time series analysis. The workhorses here are vector autoregressive models, which we discuss briefly below. Because this thesis is not primarily about methods, we refer the reader to standard texts like [86] for the full mathematical details. The models were actually fit with the statsmodels package for Python [87].

Before discussing the details, however, we should note that we are not the first to come up with this approach. As we touch on in Section 1.2, a long and diverse set of literatures in political science has explored agenda-setting and the determinants of mass media coverage. For our purposes, what's most relevant is the literature on so-called intermedia agenda-setting, or a "theory [of] how content transfers between news media" [79]. Empirical work in this vein has employed a variety of time series methods to look at the diffusion of content: everything from correlations of series [88] to more sophisticated autoregressive and moving-average models [89]. Some work has also focused on approaches to the spread of content other than time series analysis, like the manually coded story clusters in [79] and the very recent network-science approach of [90]. We are thus working in a broader tradition of data-driven approaches to this question, though it is one with significant methodological diversity. Moreover, even the traditional time-series approach has not previously been applied to large-scale radio datasets like ours.

A vector autoregressive (VAR) model captures linear relationships among variables over time: each series at each timestep is a linear function of its own lagged values and the lagged values of the other series, plus an error term. A variety of methods have been advanced for choosing the lag orders; we use one based on the Akaike information criterion ([86] p. 147). For our purposes, VAR models have the appealing property of requiring few assumptions about the structure of the relationships among the variables (unlike, say, structural equation models). We assume that these relationships are linear and depend on only a certain number of lags[1], but not which lags or the structure of the coefficient matrices. Certain more technical assumptions like stationarity of the input series are still necessary, and we've checked them but avoided including the diagnostic information here.

To estimate dynamic effects of topic mentions in some medium on some other medium (e.g., all of Twitter vs all of radio, Trump's feed vs specific radio shows, etc), we first counted mentions of the keywords selected above per 15-minute bin, yielding one time series per topic with a 15-minute timestep. Each series was normalized to a fraction of the total number of words spoken or tweeted per bin. For each topic, these per-medium series were the inputs to a VAR model of their joint variation over time.

The fitted model allows estimates of the impulse response of discussion in each medium to an exogenous increase or decrease of discussion in another medium – a "shock" in the usual econometric terminology. Absent a more specific shock of interest, a one standard deviation increase is a common choice, which we used below. We might estimate, for example, how much and over what time course radio discussion

[1]This is not always true, and isn't true of moving-average processes in particular. We've checked that (V)AR models fit the autocorrelation structure here sufficiently well, however.

of guns changes for a one standard deviation increase in Twitter discussion of guns. Note that interpreting these impulse responses as causal effects is fraught and requires identification assumptions, which are debatably valid here. We do, however, apply the same methodology to the Trump case study in Chapter 9, where drawing causal conclusions is much more plausible.

We can also use these models to conduct what's called a Granger causality test [86]. Despite the name, it is not a test of true causality. Rather "Granger causality" is the strictly weaker notion of forward predictive value: X Granger-causes Y if future values of Y can be better forecast with past data from both X and Y than with data from Y alone. This might hold because X causes Y, or it might hold for other reasons; interpreting Granger causality as causality requires additional assumptions. Because we run so many hypothesis tests in this chapter, all tests have Bonferroni corrections applied to mitigate the multiple-comparisons problem.

Finally, we also note two important points for reading the impulse response plots in Section 7.3:

Timescale of response The x-axes are in units of 15-minute timesteps, out to 96 timesteps, or 24 hours, after the exogenous shock. The effects we see from IR analysis are generally on the scale of hours, as in, e.g., the peak estimated effect in Figure 7-5 at about 20 timesteps, or 5 hours. (For an example of effects which do not clearly subside within a day, see Figure 9-1 in Chapter 9.)

Magnitude of response The y-axes are in units of mention rates: a value of, for example, 0.08% means an increase of just under one word in 1,000 in the fraction of all words which are topic query words. One of the limitations of our topic modeling approach is that it's not intuitive how much of the total discussion these rates correspond to. For every occurrence of a topic query word, after all, some number of other words (stopwords, names of people, etc.) will be intuitively about the same topic while not being included in the query words.

7.2 Static Results

This section compares the text of radio and Twitter without regard to the time dimension of the data. We made two types of comparisons, both grouping all data from 2019 to 2020 together: within-show, which compared the content of a show to tweets from the show's hosts and staff; and medium-wide, comparing all of radio to all of Twitter. Both sets of comparisons were made by cosine similarity on the vectors of topic mention rates.

The most striking conclusion of this analysis is the degree of similarity in topics between radio and Twitter. What hosts and staff discuss online and what they discuss on-air is remarkably similar, though with considerable individual variation. Moreover, aggregating all content together makes the similarity much more pronounced. This latter fact marks a trend we'll see continued in Section 8.3: large aggregates of the media ecosystem are much more predictable than individual shows.[2]

[2]While we haven't specified any sort of formal model of behavior here, this pattern calls to mind

Within shows, on-air and online content have a fairly high degree of similarity (mean similarity 0.805), but with considerable variation (standard deviation of 0.201 and a minimum of 0.128). The distribution of similarities is depicted in Figure 7-1.

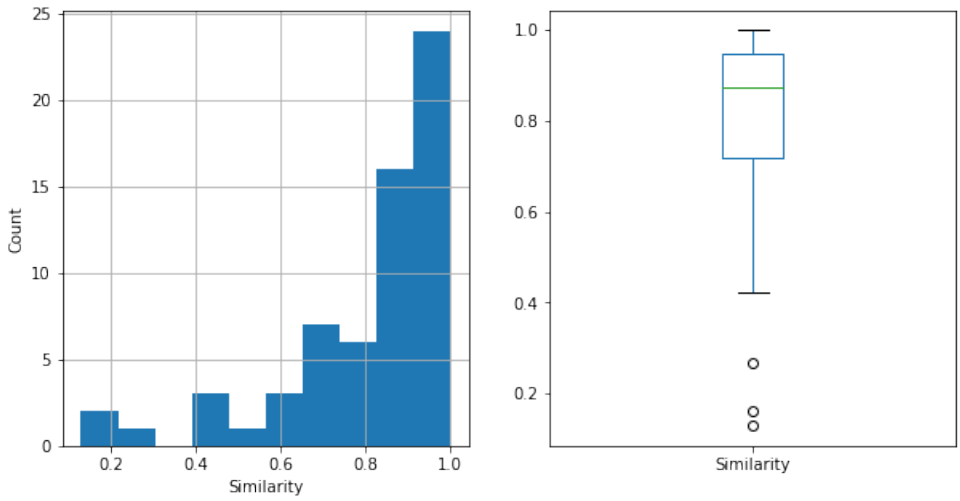


Figure 7-1: A histogram and box plot of the distribution of similarities between Twitter-matched shows' on-air and Twitter content.

Aggregating together all on-air and Twitter discussion for these same shows, however, greatly increases the degree of similarity, to 0.970. This is notably higher than the already substantial degree of similarity between radio and Twitter content overall (not restricted to Twitter-matched show), at 0.882. The difference, given the syndicated skew of our Twitter-matched set, implies more divergent use of the two media by local and smaller syndicated hosts. (Or, equivalently, a greater desire by large syndicated hosts to project a consistent brand identity.) The unnormalized average vectors themselves are shown in Table 7.1.

Now, the results don't support the conclusion that radio and Twitter discuss substantially the same set of topics, or even that hosts do: after all, we're comparing along a prespecified set of topics. Nevertheless, while these two media can serve different roles for any individual user, they are tightly coupled and used fairly similarly by hosts in their role as forums for political discussion.

7.3 Dynamic Results

In this section, we wanted to see whether topic distributions in elite Twitter at a particular time were predictive of topics on the radio in the future, as our public opinion model in Section 1.2 would suggest.

Even if Twitter does influence radio, there are several reasons to expect a lag in response. For one thing, because radio takes more time and more work to produce

limit theorems in probability. If shows behave at least somewhat independently, as they appear to here, large aggregates of them ought to converge to similar average behavior. The old saw about journalists being a "herd of independent minds" may turn out to be true in a formal sense as well.

Topic	Radio	Twitter	Matched Radio	Matched Twitter
Trump	0.172%	0.435%	0.266%	0.389%
Politics	1.498%	2.088%	1.937%	1.861%
Education	0.381%	0.258%	0.342%	0.215%
Crime	0.176%	0.140%	0.164%	0.155%
Healthcare	0.439%	0.272%	0.458%	0.187%
COVID-19	1.347%	1.182%	1.383%	1.110%
Sports	0.987%	0.271%	0.694%	0.259%
Economy	0.243%	0.149%	0.258%	0.119%
Guns	0.053%	0.045%	0.056%	0.051%
Weather	0.754%	0.088%	0.379%	0.092%
Climate	0.078%	0.078%	0.094%	0.095%
Inclusivity	0.026%	0.043%	0.032%	0.040%
Drugs	0.090%	0.056%	0.102%	0.054%
Immigration	0.059%	0.094%	0.074%	0.078%
Other Social Issues	0.040%	0.041%	0.051%	0.042%

Table 7.1: Mention rates of topic query terms for a) radio staff's on-air discussion and their tweets (the "Matched" columns) and b) all radio discussion and all Twitter discussion.

than the immediacy of tweeting, not all content is aired live. Widespread use of syndicated content, which is frequently aired on a time delay, has the same effect. Finally, while production staff may or may not monitor social media or the news, hosts certainly have a hard time doing so themselves while simultaneously broadcasting.

There were three specific analyses, which we discuss in specific subsections. The first (subsection 7.3.1) aggregated all of radio and Twitter together, while the other two (subsection 7.3.2 and subsection 7.3.3) look for effects on particular shows and station owners. We considered only data from September and October 2019, rather than all data, in order to avoid a nonconsecutive break in the middle of the dataset.

7.3.1 Many Topics

Here, we aggregated all radio together, and included two series for each topic, one for radio mentions and one for Twitter mentions. The quantity of interest was, for each topic, the estimated impulse response of radio mentions to an exogenous increase of one standard deviation in Twitter mentions.

Most results are null. We examined 17 topics, and found no significant effect of Twitter on radio for 14 of them. Of a broader set of comparisons, the 144 tests of any Twitter topic's influence on any radio topic, only one additional test was significant, and then only at the (Bonferroni-corrected) $\alpha = 0.05$ level. These results are summarized in Table 7.2, and the distribution of the broader set of p-values is shown in Figure 7-2.

Threshold	Within Topic	Across Topics
$\alpha = 0.05$	3 / 17	3 / 144
$\alpha = 0.01$	3 / 17	3 / 144
$\alpha = 0.01$	3 / 17	4 / 144

Table 7.2: The results of Granger causality tests for the many-topics model. The "Within Topic" column refers to the 17 tests of a Twitter topic Granger-causing the same topic on radio, while the "Across Topics" column examines any Twitter topic Granger-causing any radio topic. The thresholds shown are before applying Bonferroni corrections, so the true threshold used was the depicted α divided by 144. As can be seen, most results are null.

Only three topics – politics, the economy and climate – showed clear estimated response, consisting of a large spike in radio mentions that decayed over the course of about a day. Two of these plots are shown in Figure 7-3. Granger causality tests find the responses to be significant ($p <= 10^{-8}$).

7.3.2 Many Shows

For each of a few large topics, we fit a VAR model with one time series for mentions of that topic in each of several media. These media included in particular elite Twitter, the decahose, five large radio shows and @realDonaldTrump's Twitter feed, which we broke out as a separate medium. The five shows included three leading conservative shows (Rush Limbaugh, Sean Hannity, and Glenn Beck) and the two largest NPR shows (All Things Considered and Morning Edition). The topics were (see Appendix B for the corresponding keyword lists) politics, the economy, guns and climate, which accounted for all significant results in subsection 7.3.1. Our goal was to test for a relationship between either Trump or the rest of elite Twitter, on the one hand, and the specific shows, on the other.

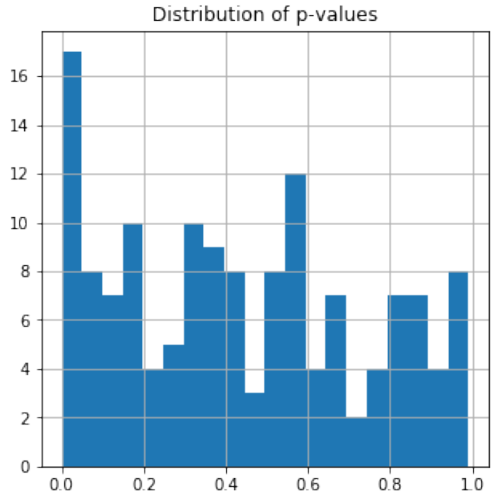


Figure 7-2: The distribution of p-values for Granger causality tests of Twitter influence on radio. The plot summarizes a broader set of tests than the discussion focuses on: there are 144 tests depicted, each of a given Twitter topic's influence on a given radio topic. Even before applying multiple-comparisons corrections, it's clear that most results are not significant.

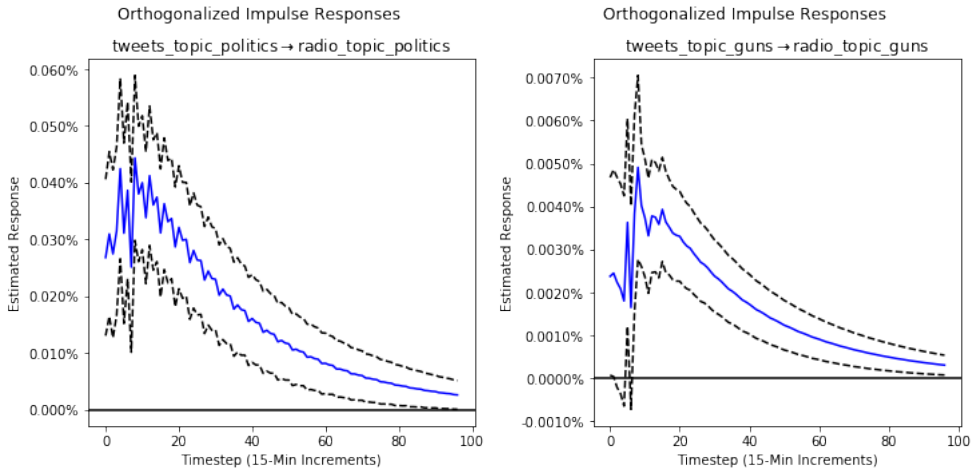
These show-level models also find mixed, and mostly null, results. There are a few highly significant Granger causality estimates (again, Bonferroni-corrected), but the corresponding impulse response functions are noisier, and most cross-medium effects are null. Out of 40 estimates[3], only five are significant at the Bonferroni-corrected $p = 0.001$ level, and an additional three at the $p = 0.01$ level. Four of the first five estimates involve Morning Edition, All Things Considered and Rush Limbaugh lagging Twitter discussion of the economy. We've shown an example of impulse responses corresponding to a significant estimate and to an insignificant estimate in Figure 7-4, and summarized the test results in Table 7.3.

7.3.3 Many Owners

Here, we repeated the per-topic multi-show model approach, but broke out radio by station owner rather than by show. The set of topics was the same as in the previous section. To keep the number of series manageable, we considered only the top three owners of non-public radio stations, by total number of stations ingested during 2019: iHeartMedia (37 stations), Cumulus Media (9 stations) and Townsquare Media (9 stations).

The results here, like the show-level analysis above, are mixed but mostly null. No owner had a significant Granger causality relationship to elite Twitter discussion or Trump tweets for all topics. The closest thing to a pattern was iHeartMedia's discussion of topics being Granger-caused by elite Twitter, which we found at the Bonferroni-corrected $p = 0.01$ level for three of the four topics (politics, climate and

[3] 4 topics x 5 shows x (elite Twitter, Trump).



(a) The politics topic.

(b) The guns topic.

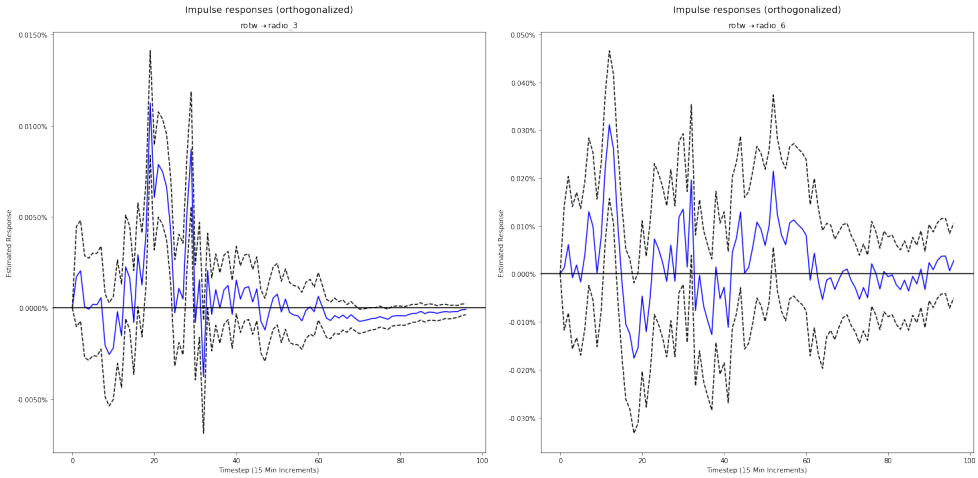
Figure 7-3: Impulse response estimates for radio discussion of a topic after an exogenous increase (or "shock") of one standard deviation in Twitter discussion of the same topic. The politics and guns topics are shown. In both plots, the quantity on the y-axis is the (dimensionless) proportion of words consisting of topic query words. The x-axis is in units of (15-minute) timesteps. The corresponding Granger causality tests are significant at the $p < 10^{-8}$ level.

the economy). The corresponding impulse response function for the politics topic is shown in Figure 7-5, and the test results are summarized in Table 7.4.

7.3.4 Discussion

These results provide, at best, weak evidence of temporal structure in the relationship between Twitter and radio. There's clear evidence of Twitter leading radio discussion during our study period on certain important topics (politics, climate and the economy). But this pattern held only rarely: the other 14 topics, many related to politics, didn't show a relationship, and show or owner breakouts had noisy and still mostly null results.

These findings seem to rule out the simplest form of agenda-setting from Twitter to radio. Topics of radio discussion are not simply an echo of what Twitter was talking about shortly before. Beyond that, how to interpret the results is unclear. It might indeed be that Twitter doesn't systematically lead radio discussion, providing evidence of editorial decoupling despite the similarity of content in Section 7.2. It is also possible that our topics were defined too broadly to capture real but more granular intertemporal relationships, or that those relationships are restricted to certain topics. Further research should pursue these more refined questions.



(a) Economy: Twitter → Limbaugh (b) Politics: Twitter → Glenn Beck

Figure 7-4: Two impulse response estimates for particular shows' discussion of particular topics after an exogenous increase (or "shock") of one standard deviation in elite Twitter discussion of the same topic. The politics and economy topics are shown. In both plots, the quantity on the y-axis is the (dimensionless) proportion of words consisting of topic query words. The x-axis is in units of (15-minute) timesteps. Both estimates are clearly noisier than those from the medium-level aggregate model presented earlier.

Topic	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$
Politics	0 / 10	1 / 10	2 / 10
Economy	4 / 10	4 / 10	4 / 10
Climate	1 / 10	2 / 10	2 / 10
Guns	0 / 10	1 / 10	1 / 10

Table 7.3: The results of Granger causality tests, by topic, for models with disaggregated shows. The significance thresholds shown are before applying Bonferroni corrections, so the true threshold used was the depicted α divided by 64 (the total number of tests between variables in the model). The 10 tests in each cell reflect two potential causing series (Trump and the rest of Twitter) and five potential caused (the five radio shows). As can be seen, most results are null.

Topic	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.05$
Politics	1 / 6	2 / 6	2 / 6
Economy	1 / 6	1 / 6	1 / 6
Climate	1 / 6	2 / 6	2 / 6
Guns	0 / 6	1 / 6	1 / 6

Table 7.4: The results of Granger causality tests, by topic, for models with disaggregated shows. The significance thresholds shown are before applying Bonferroni corrections, so the true threshold used was the depicted α divided by 36 (the total number of tests between variables in the model). The 6 tests in each cell reflect two potential causing series (Trump and the rest of Twitter) and three potential caused (the three station owners). As can be seen, most results are null.

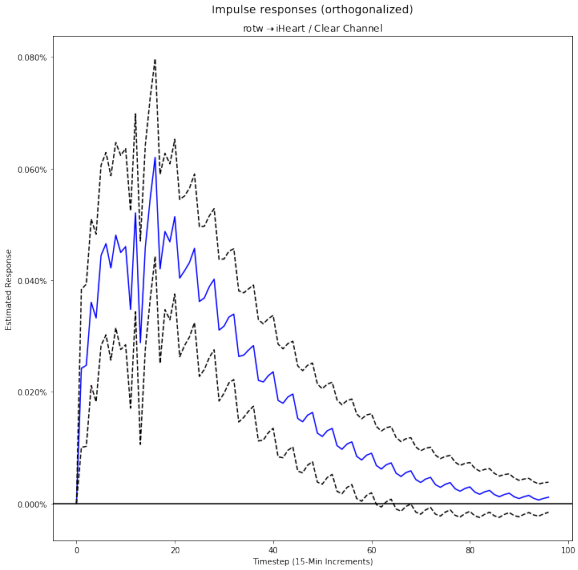


Figure 7-5: The estimated impulse response function of iHeartMedia radio stations, aggregated together, to an exogenous increase (or "shock") of one standard deviation in Twitter discussion of politics. The shocked medium is denoted "rotw," or "rest of Twitter," for elite Twitter without Trump.

Part IV

Case Studies

Chapter 8

Discussion of COVID-19

This chapter presents our first case study of radio and Twitter as an integrated media ecosystem. Both this chapter and the next focus on how these two media interact in specific cases, with an eye toward testing the conclusions drawn in the previous chapters. Here we look in particular at the spread of discussion of COVID-19 and certain COVID-associated memes ("invisible enemy" and "flatten the curve") on radio and Twitter, during the months of March and April 2020.

With the coronavirus spreading rapidly, this period was one of rapid change in media discussion as in every other part of life.[1] The speed and magnitude of the changes may allow us to observe aspects of the media ecosystem that are normally hidden.

8.1 Methodology

As a case study, this period has the additional appealing attribute that its various memes and topics of discussion are clearly lexically marked. Identifying discussion of COVID-19 and the two memes we focus on below, especially early on, is just a matter of counting up mentions of a few keywords. We thus used much the same methodology as in Chapter 7 and Chapter 4: For each meme or topic (here, "COVID-19", "invisible enemy" and "flatten the curve") we listed a set of relevant keywords and searched for mentions of them in both radio and Twitter text. A binary variable for each meme was assigned to each tweet (or speaker turn in radio) indicating whether that tweet or snippet contained any of the meme's keywords. We grouped these up to the day level in most analyses, or by other variables in some cases.

As described in Section 2.5, Twitter text was first preprocessed to make it more similar to radio text. The full keyword lists are in Section B.1; notably, many of the keywords are manually discovered misrecognitions of "COVID-19" in the radio data.[2]

[1] The instinctive thought that the "every [] part of life" here is hyperbole, and then the realization that for once it isn't, makes the point well.

[2] "Covered nineteen" and the like.

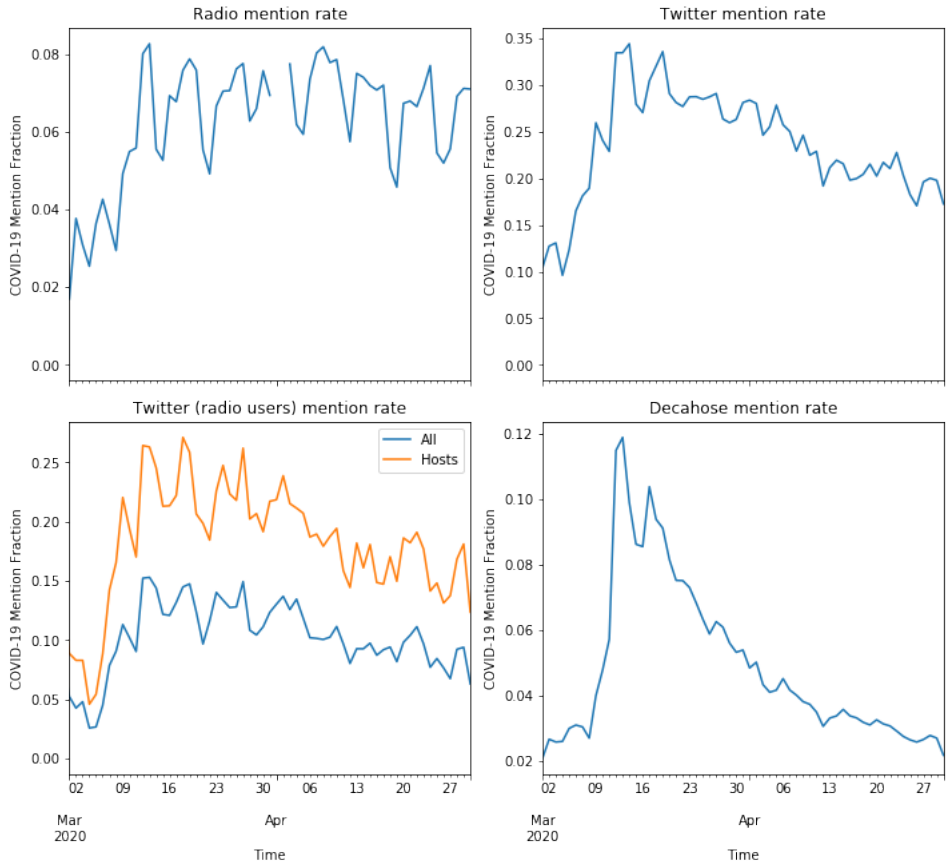


Figure 8-1: The fraction of snippets or tweets mentioning COVID-19 by day during March and April 2020, broken out by medium. "Twitter" here refers to our elite Twitter universe. The interruption in radio data around April 1 reflects a short-lived problem with the audio ingestion system.

8.2 COVID-19

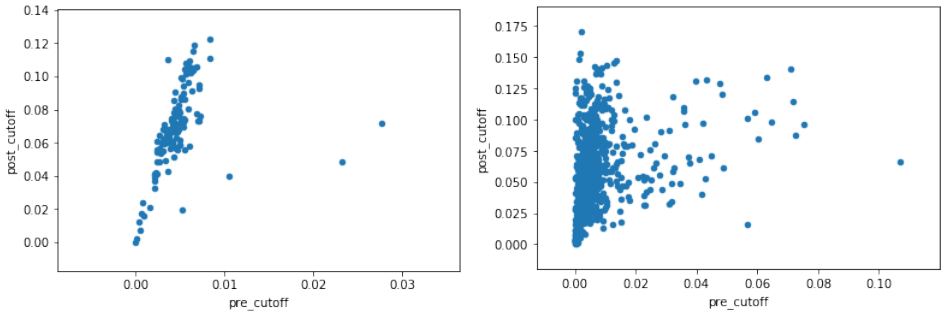
Discussion of COVID spiked similarly in radio, elite Twitter and the decahose during mid-March. COVID-19's rapid rise to the main topic of public discussion (see Figure 8-1) is apparent in all of these media: during the week of March 9th, mention rates doubled on radio, tripled in elite Twitter, and rose by a factor of six in the decahose. After that, however, it evolved quite differently in different media. Coronavirus discussion decayed gradually on elite Twitter, held roughly steady on radio, and decayed rapidly on the decahose, falling back to baseline levels by the end of April. Radio hosts discussed COVID slightly less often than elite Twitter overall, with non-host show staff being even less likely to do so.

Because March 9th and the succeeding 2-3 days marked such an abrupt change in COVID discussion, we conducted some analysis by comparing discussion rates beforehand to discussion rates afterward. Two salient facts are apparent.

First, stations are not passive conduits for information, but exercise editorial judgment just as shows do. As seen in Figure 8-2, there was a remarkably linear

relationship between a station's degree of early coronavirus coverage and its degree of coverage after the 9th. Shows displayed considerably more variance than stations, implying that station operators were choosing their syndicated shows (or influencing local shows) to provide a consistent perspective. Though the absolute amount of discussion changed across the discontinuity the week of March 9th, stations' relative perspectives remained quite stable (with only the four exceptions visible in Figure 8-2a).

Second, though it's hard to say for sure what causes it, more conservative hosts showed a greater increase in discussion of COVID. Nor was this faster growth toward the same level of discussion as liberal shows: before the 9th, as one can see in Figure 8-3, there wasn't much variation in COVID discussion by ideology score. Afterward, conservatives show a distinctly higher discussion rate. The change in COVID discussion on the show level is plotted directly in Figure 8-4, making a similar point.



(a) Change in discussion by station (b) Change in discussion by show

Figure 8-2: Each radio station and show plotted according to the fraction of snippets before March 9th ("pre-cutoff") mentioning COVID-19 vs the fraction mentioning it after March 9th ("post-cutoff").

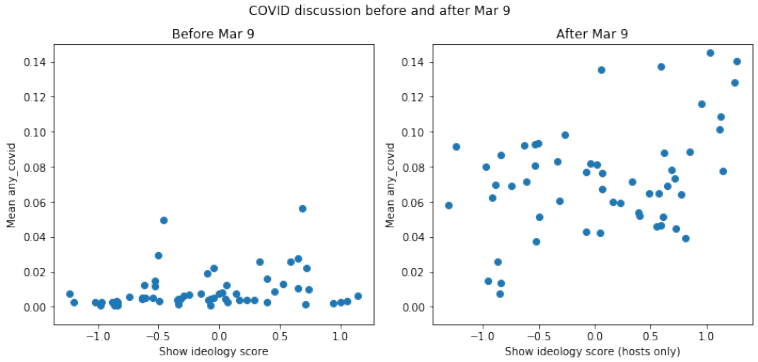


Figure 8-3: COVID-19 discussion before and after the mainstreaming of the pandemic in mid-March (specifically March 9th), by show, vs Twitter-based estimates of the show's ideology. Higher ideology scores correspond to more conservative shows.

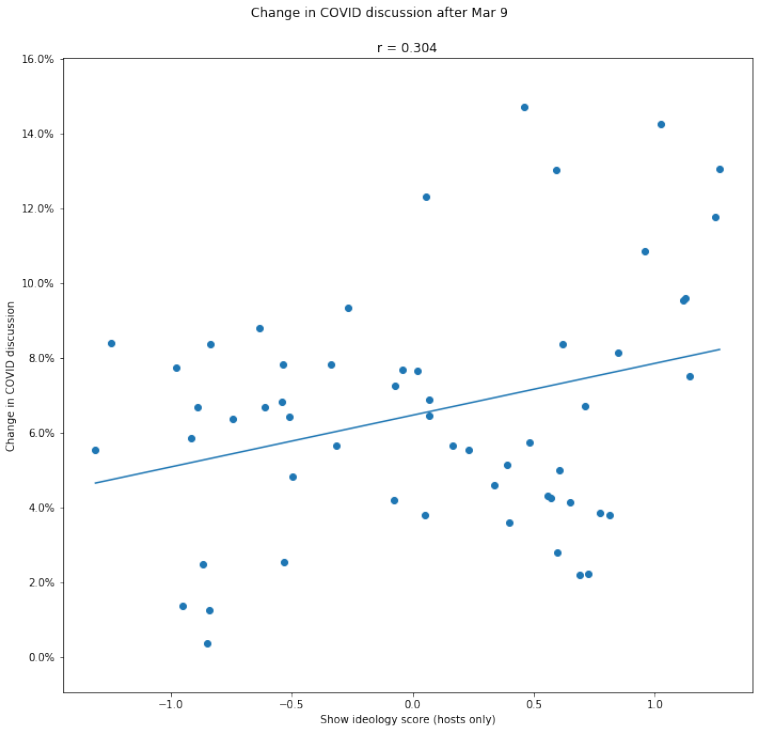


Figure 8-4: The shift in COVID-19 discussion after the mainstreaming of the pandemic in mid-March, by show, vs Twitter-based estimates of the show's ideology. Higher ideology scores correspond to more conservative shows.

8.3 Memes in Detail

"Invisible enemy" (IE), one of President Trump's terms for the coronavirus, was introduced into media discussion when he mentioned it on television at a coronavirus task force briefing on March 16th, 2020. Despite scattered occurrences of the phrase before that, his usage of it seems to represent a new coinage rather than being inspired by earlier usage. He subsequently tweeted about it several times. This phrase thus represents, at least partly, an example of something we haven't examined before: TV's influence on both radio and Twitter. (The tweets themselves presumably had some direct impact as well.)

IE was never a particularly popular meme. It got significantly less uptake than other coronavirus-related memes like "flatten the curve," as can be seen in Figure 8-5. After the briefing on the 16th, it got some initial uptake on the radio, but discussion fell off quickly. It proved more enduring on Twitter, even becoming more popular over time.

"Flatten the curve" (FTC), a meme from the scientific community about controlling the coronavirus, was much more successful than "invisible enemy." It saw notably more uptake on both Twitter and – eventually – radio. (See again Figure 8-5.) It surged to a remarkable 1.2% of all elite Twitter discussion in mid-March, and remained elevated

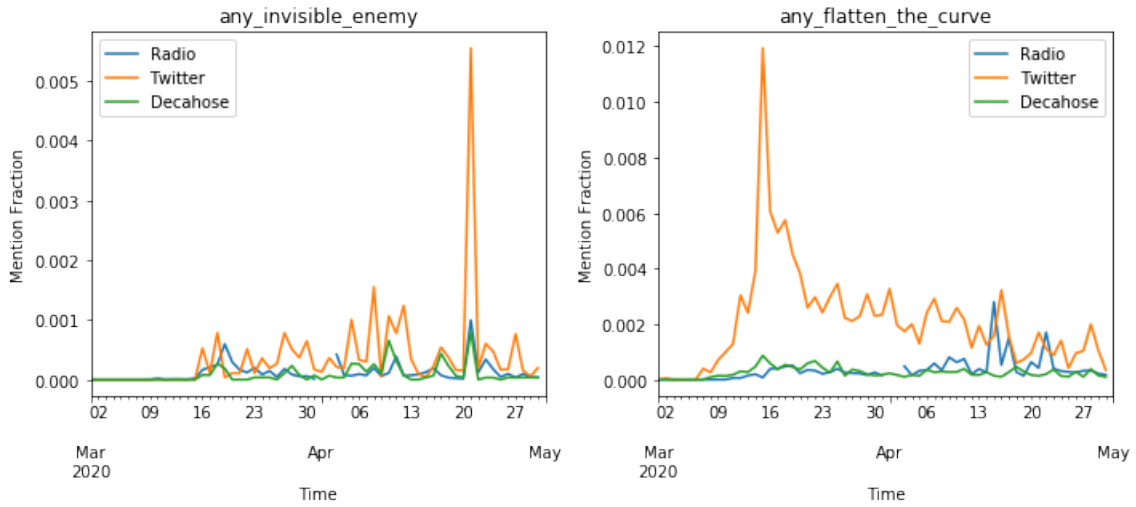


Figure 8-5: Mentions of the "invisible enemy" and "flatten the curve" memes over time in three different media: elite Twitter, decahose Twitter and radio. Note the different y-axis scales on the two subplots.

all through April. The lack of discussion of it on the radio[3] is striking: the first mention of the phrase at more than occasional levels came in mid-April. Interestingly enough, the mid-April surge had a specific cause not tied to a news event: Rush Limbaugh did an entire episode on April 15th about "flattening the curve," arguing that it had been something between a mistake and a broken promise. A spike in elite Twitter discussion followed the next day.

The relative lack of discussion of FTC in the decahose during this entire period is even more striking. Aside from a brief blip in mid-March, during the huge spike in elite discussion, the decahose never talked much about curve-flattening. Moreover, the pattern of greater elite interest is not limited to our universe of the very most influential media users. Figure 8-6 shows mention rates for both memes, and COVID-19 generally, in the Twitter decahose, broken out by deciles of various Twitter metrics. Across several different ways of defining Twitter's most active and elite users – followers, friends[4], and number of posted statuses – the more elite and plugged-in the user, the more likely they were to take up both memes.

8.3.1 Discussion

Unlike discussion of COVID-19 overall, it's fair to say that both of these memes were elite-led efforts. In one case, the President coined and pushed a specific meme, while in the other the originators were public health leaders. Where they got significant uptake, it was mostly among media elites on Twitter or radio.

[3]We included several alternative grammatical forms in the query terms to better match patterns of speech on the radio, including especially "flattening the curve". It's still possible that there is radio discussion we're missing.

[4]The opposite of followers: user A's friends are the users that A follows.

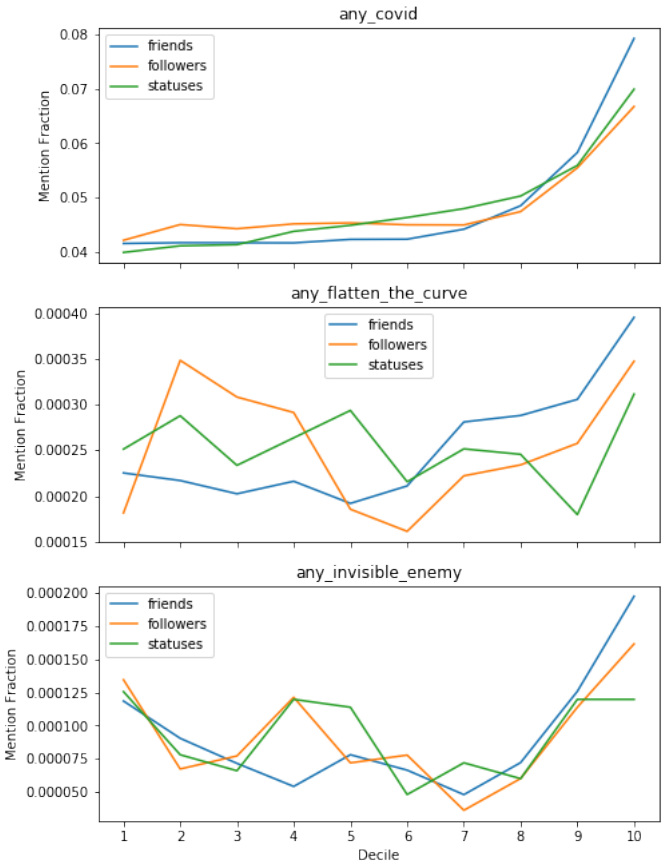


Figure 8-6: Fraction of decahose tweets mentioning each of three topics or memes – COVID-19, "invisible enemy" and "flatten the curve" – broken out by deciles of a user's number of followers, friends and posted statuses.

The general Twitter user population, though it can be seen to respond slightly to elite discussion, was less impressed. In Section 1.2, we motivated questions in this thesis in terms of a model of public opinion, one in which elite discourse reaches the public and influences public opinion through mass media. While we're interested mostly in exploring the elite discourse side of that equation, the example of the decahose here provides our first square look at something approximating public opinion rather than media coverage. Without further work, it's hard to say what the limited response of that proxy for public opinion means. The options range from the technical (decahose users might discuss the same topics as media elites, but without the same keywords) to the more substantial (elite discourse might simply be having limited impact on the public). While the direct findings wouldn't change either way, the interpretation certainly would: are the dueling messaging efforts of this section merely elite pastimes, perhaps aimed at convincing other elites, or effective outreach to the public?

Finally, it should be noted that these elites do not speak with one voice. Just the existence of these two memes, originating from different corners of the media and media-adjacent "elite", provides one example. Rush Limbaugh's curve-flattening radio episode, which argued against a popular meme, provides another, and highlights the

potential for cross-medium influence as discussed in Section 7.3. Elite discourse, as we conceive of it in Section 1.2, proceeds over time and affords opportunities for influence between Twitter and radio without a certainty of influence in either direction.

Chapter 9

Causal Effects of Trump Tweets

This chapter presents our second case study, of a set of President Trump's tweets and their impacts.

9.1 Introduction

Intuitively, it seems clear that these tweets have wide reach and great impact: things Trump tweets often become major subjects of media discussion. We may hear about a Trump tweet for days afterward[1], and even when one fails to get purchase in the media, it may still reach his Twitter followers. If we assume that Trump engages in his characteristic frequent tweeting for some benefit to himself, what is that benefit? Discounting purely psychological explanations, tweets as speech acts must be intended to be heard. One view is that the main audience is the general public: Twitter provides Trump with a way to get his message out independent of media gatekeepers.

We suspect the story is more complicated, and that other audiences – especially those reached by the mass media, in line with Section 1.2's model of public opinion – are more important. Some numbers may help make the point: Most Americans aren't on Twitter, and those who are mostly don't use it daily: as of 2018, there were 27 million active daily US users, out of 66 million monthly active users.[2] Moreover, as mentioned in Chapter 1, only about 10% of Twitter users account for virtually all discussion of national politics [12]; the community of Twitter users who participate actively in the platform's political goings-on is not necessarily large.

In a similar vein, the Trump tweets in our sample period get less direct engagement than their apparent impact would suggest. We can't calculate viewership for these tweets[3], but the average Trump tweet during the same time period as our radio corpus got only 42,892 likes out of all its viewers.

[1] For example, an op-ed in the LA Times [91] was discussing the "LIBERATE MICHIGAN" series of tweets six days later.

[2] Both of these figures come from Twitter's investor disclosures. See the Q4 2018 letter to shareholders [54], which was the last one to report monthly usage figures.

[3] For several related reasons: Twitter doesn't report it; retweets amplify it; and we have no way of knowing which of a user's followers saw but didn't interact with any given tweet.

The mass media, by contrast, reach much larger shares of America. Radio in particular has an enormous audience: according to Nielsen [7], as of 2018 92% of Americans listened to some kind of terrestrial radio during a typical week. Not all of this reach is politically focused radio (music stations are unlikely to discuss Trump tweets), but if even 10% of those listeners listened to some talk radio or NPR news, it's around the same size as Twitter's number of average daily US users during the same period[4], and much larger than the number of frequent politics tweeters. And while we won't explore media other than radio and Twitter here, this estimate doesn't consider TV news, newspapers, or online media.

Much of the value Trump gets from his tweets, then, may be in their influence on the agenda of the traditional press. In keeping with the discussion above, there are two ways we might theorize this influence happens, or of course some mixture of the two:

- Trump is president, and what the president says is news. So while his tweets influence media coverage, they would have the same impact if he said them on TV instead (or, topically, in the traditional weekly radio address).

- Twitter's central role in media discourse means that dominating Twitter discussion allows one to also set the media's agenda. Between the many outlets which may discuss any given Trump tweet, the potential audience is likely larger than the tweet's direct reach.

There is reason to suspect that both of these explanations are true – that, in other words, the presidency gives Trump agenda-setting power, which his use of Twitter enhances. We discuss some support for both views from the existing literature below.

As we've touched on in Section 1.2 and Section 7.1, agenda-setting dynamics and the power different actors have to set the media agenda are long-time research topics in political science. (See McCombs' 2013 book on the subject [93] for a review.) Quantitative analysis of them, frequently using the same time series methods we use here and in Chapter 7, date quite far back (see, e.g., Boyle's work in 2001 [94] and Wanta et al's 1994 [95] time-series analysis.). Much of this work, reflecting its origins in political science, has focused on the agenda-setting power of the US president in particular, just as we do here. For example: Horvit et al [96] attempt to quantify the coverage generated by the president's traditional weekly radio address, and find that it's diminished over time. Peake and Eshbaugh-Soha [97] examine the effect of prime-time TV addresses from the president on the news agenda in the succeeding days and weeks, finding a possiblity but not a guarantee of influence.

The Twitter-centric theory, while obviously more recent, has not gone unexamined either. Indeed, a fair amount of research has looked at the use of Twitter to influence the media, focusing on journalists' own use of Twitter as the distinguishing factor from other means of doing so. Kreiss in 2016 [98] provided a qualitative look at how the 2012 Obama and Romney campaigns had tried to influence journalists' discussion on Twitter.

[4] A ballpark estimate: 27 million according to [54], vs 328 million Americans * 77.7% 18 and older * 92% * 10% = 23.4 million [92].

Both campaigns quite consciously did so, in an attempt to set the agenda. Jungherr, also writing in 2016 [99], surveys research on the role of Twitter in election campaigns; he describes how Twitter is "increasingly incorporated in campaign repertoires of traditional parties and candidates," driven by a "collective negotiation of meaning between political elites, journalists, and other Twitter users during the course of a campaign." Providing support for our focus on elite Twitter, Harder et al [79] conduct a methodologically similar analysis and find that the Twitter accounts of journalists and politicians in particular have meaningful influence on the news agenda. Work from other disciplines has contributed as well: Kwak et al [50] discusses views of Twitter as a news medium and information source with the capacity to influence its most central users (like journalists).

Still, while both explanations are plausible, distinguishing between them might require a different research design. It would certainly require data beyond what we have here. Additional data sources could include transcripts of Trump's TV utterances or comments to the White House press pool, White House press releases, and other types of presidential statements. (From another angle, datasets of news stories or a complete map of radio hosts to their Twitter accounts would allow examining whether journalists' Twitter engagement with Trump predicts their coverage of him.)

Instead, lacking such datasets, we'll explore Trump's agenda-setting power in Twitter and radio without being able to definitively apportion credit for it to Twitter and the traditional bully pulpit. We'll again consider March and April of 2020, as an uninterrupted period with both radio and Twitter data available.

9.2 Methodology

We want to look at the spread of ideas Trump may be tweeting about, and identify their impact on other media. Identifying an idea in a piece of text is difficult in general, but fortunately for us, Trump tends to coin distinctive short phrases and nicknames. The approach here, then, is similar to that used in Chapter 7. Given a set of Trump memes, we identify them by sets of keywords, count up the keywords and generate rates of meme mentions (as a share of all words) on each medium per quarter-hour.

For each meme, we fit a VAR model to series of mentions in five media (Trump's own tweets, elite Twitter without Trump, talk radio, public radio, and the Twitter decahose), each binned to this quarter-hour resolution. These models then allowed us to make inferences about Granger causality and impulse responses to Trump's tweets.

We ran standard goodness-of-fit tests for these models – checking for unit roots in the input series, examining autocorrelation and partial autocorrelation plots for autoregressive signatures, and checking for minimal levels of autocorrelation in the residuals – without finding cause for concern. Each model's lag order was selected by the Akaike information criterion.

9.2.1 Meme Selection

Selecting the relevant memes is the most open-ended aspect of the analysis. We experimented with automated ways of finding phrases[5], but they missed too many rare phrases and included too many common figures of speech. In the end, we simply read every Trump tweet from March and April 2020 and manually made a list of distinctive memes, which appears in Table 9.1.

Meme	# Mentions	Query Phrases
Fake news	78	fake news
Radical left	12	radical left, radical liberals
Do-nothing Democrats	14	do nothing democrats
Enemy of the people	7	enemy of the people
LIBERATE	3	liberate
Mini Mike (Bloomberg)	19	mini mike
Sleepy Joe (Biden)	26	sleepy joe
Crazy Bernie (Sanders)	5	crazy bernie
Pocahontas (Elizabeth Warren)	6	pocahontas
Shifty (Adam Schiff)	2	shifty
Hoax	12	hoax
Invisible enemy	14	invisible enemy
Morning Psycho (Joe Scarborough)	4	morning joe, morning psycho, morning joke
Voter fraud	4	voter fraud
NYT	8	new york times
(Matt) Drudge	2	drudge
Second Amendment	20	second amendment, 2a

Table 9.1: The full list of Trump memes examined for their influence on Twitter and radio. The query terms shown in the third column are listed separated by commas, and each term matched any permutation with underscores replacing spaces. The number of occurrences is in Trump tweets from 2020-03-01 to 2020-04-30.

[5]For example, if we train a logistic regression model on 1-, 2- and 3-gram features to predict whether a Trump tweet was from the desired period, the most informative features should be new coinages.

These memes span several kinds of Trump messaging. Some are catchphrases not related to particular people ("fake news", "radical left"), some are nicknames ("Sleepy Joe" and "Crazy Bernie") and others are political actors or topics of debate (the "failing" New York Times, Matt Drudge, the Second Amendment). Some of them are new ("LIBERATE") and some are classics ("fake news").

9.2.2 Causality

As in Chapter 7, the econometric modeling approach we use allows for estimating impulse responses and Granger causality. We noted there that Granger causality is not causality, but rather predictive value for forecasting. That is, current and past values of X can be useful for predicting future values of Y for reasons other than X causing Y. For example, X and Y might both be lagging indicators of some third variable Z. In general, making a causal claim from a finding of Granger causality requires identification assumptions: ruling out certain potential confounds on grounds external to the analysis itself.

While there weren't grounds to do so in Chapter 7, here we believe causal claims about the effect of Trump's tweets are quite plausible (though more so for some memes than others). The first thing to observe is that the data-generating process is very simple: just one person posting tweets. We don't need to consider multiple stations or shows making up an aggregated text stream, effects of syndication, or many other such difficulties. This one person's media consumption habits, moreover, are fairly well known. He follows few people on Twitter, is not known to be a regular radio listener, and frequently posts spontaneous thoughts.

Trump does regularly watch cable news, which is a potential confounder, and a future version of this analysis should include Fox News data to address this concern. That is, it's possible Trump is tweeting "fake news" shortly before Rush Limbaugh mentions it because they're both taking cues from Fox. The details matter, though: if Trump tweets about things he sees on Fox, but Fox's influence on the subsequent spread of the phrases is only through Trump, a causal effect is still identified. Regardless, as with the "LIBERATE" series of tweets, the number of post-Trump mentions that explicitly reference Trump makes this worry unlikely in most cases.

9.3 Results

Trump's tweets clearly precede certain changes in discussion of memes he coins, and as above it's plausible to claim he causes these changes. These effects are visible on all four other media (elite Twitter, decahose Twitter, talk radio, and public radio) in at least some cases. Not all memes have equally large effects, however. Some induce large and long-lasting shifts in discussion, while others pass by without much notice.

Figure 9-1 and Figure 9-3 illustrate examples of the impactful and ignored kinds of meme, respectively. As in Chapter 7, these plots depict estimates of the orthogonalized impulse response functions. In the interest of saving space, the other 15 sets of impulse

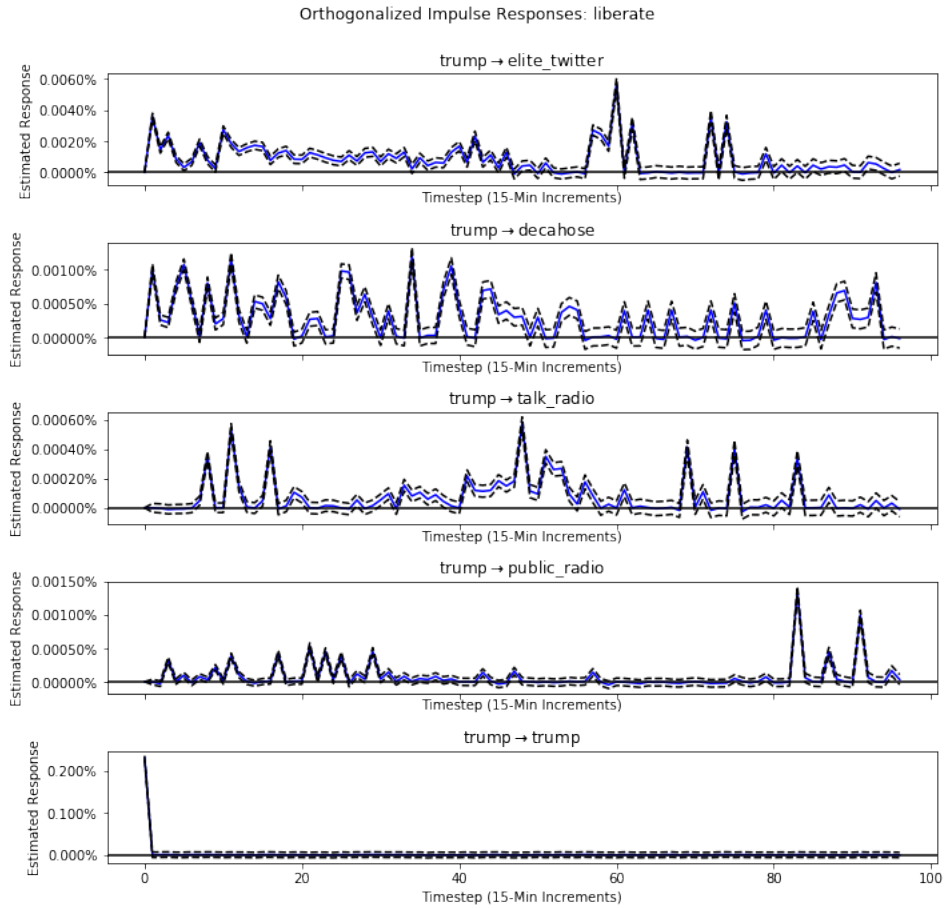


Figure 9-1: The estimated impulse responses to the "LIBERATE" series of tweets over the succeeding 24 hours. The y-axis is in units of mention rates (i.e., meme query words as a share of all words). Note that the different channels have responses on quite different scales: for example, the largest increase in mention rate on elite Twitter (0.006% of all words) was 10 times as high as the largest spike on talk radio.

response estimates are described in the text and sampled in tables but not depicted in plots.

As depicted in Figure 9-2, Table 9.3 and Table 9.2, we tested for Granger causality of Trump tweets on a mentions of a particular meme in a particular medium. We applied Bonferroni corrections to the resulting p-values in order to account for the large number of tests conducted.

9.3.1 Twitter

Notable spikes in discussion on both the Twitter decahose and elite Twitter follow Trump's tweets about most memes, with the increase on elite Twitter often much larger than on Twitter in general. We provide examples of cases where elite Twitter

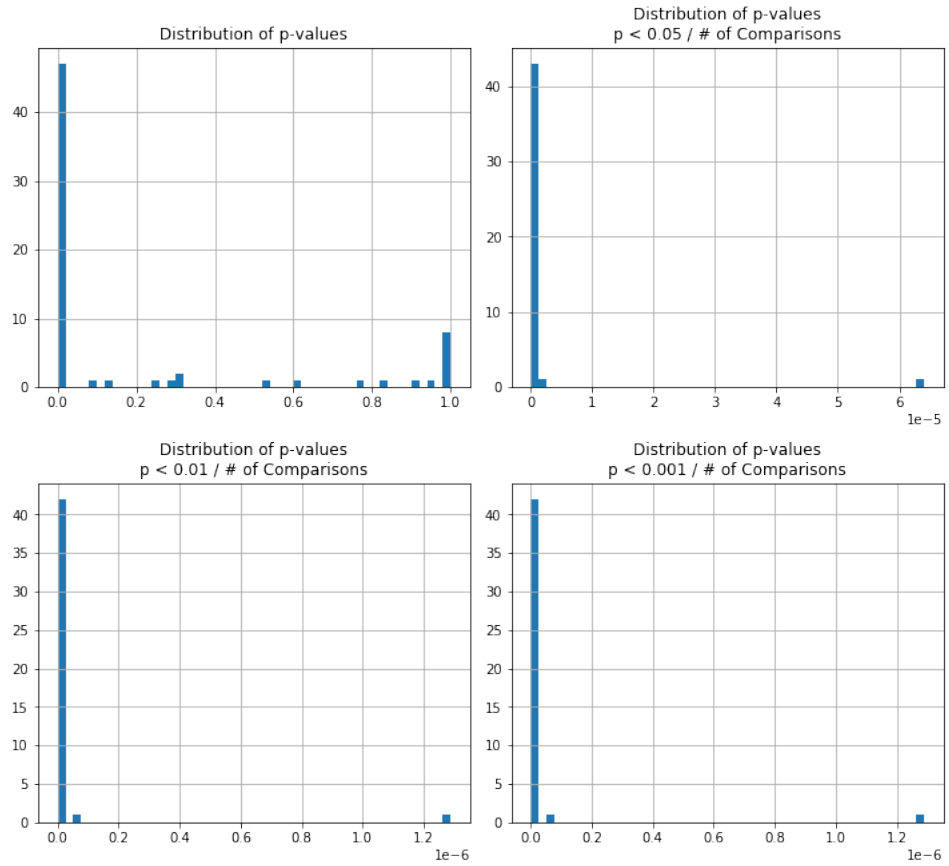


Figure 9-2: The distribution of p-values of Granger causality tests for Trump's influence on Twitter and radio, both overall (at top left) and below three progressively stricter significance thresholds. Note the scientific notation on three of the plots. All three thresholds have been Bonferroni-corrected for the number of comparisons performed. The corrections count all comparisons, not just for Trump's influence, and so are stricter.

shows more interest, as well as one case where it does not, in Table 9.2.[6]

In the first four presented, both the maximum and cumulative effects are larger in elite Twitter, while "Sleepy Joe" seems to elicit a greater response on the decahose. Though these are only a few examples, the elite-decahose disparities are consistent with the agenda-setting view of Trump's tweeting discussed above. (Though, again, we can't distinguish between Twitter and the general bully pulpit as mechanisms: would elite Twitter have been equally interested if he had made these comments on TV?)

The estimates of cumulative effect presented with these tests, in both Table 9.2 and Table 9.3, are derived from numerical integration of the impulse response functions[7], while the maximum estimated effect is the largest for any 15-minute period.

[6]Detailing all 34 estimates is unwieldy.

[7]Specifically the trapezoidal estimate of the area under the curve.

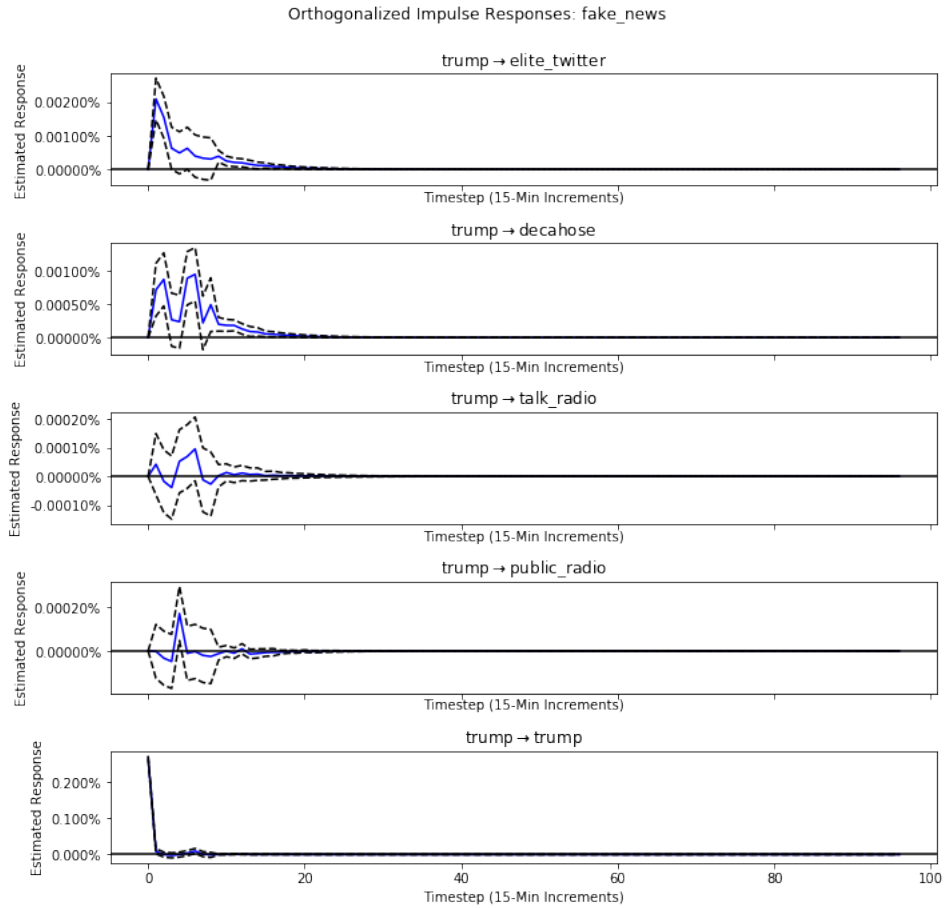


Figure 9-3: The estimated impulse responses to Trump tweeting "fake news" over the succeeding 24 hours. The y-axis is in units of mention rates (i.e., meme query words as a share of all words). Note that the different channels have responses on quite different scales, though the gaps are smaller than for Figure 9-1. For example, the mention rate in elite Twitter (0.002% of all words) is twice as high as in the decahose (0.001% of all words).

9.3.2 Radio

Not all Trump meme tweets produced apparent responses on radio. Large and obvious spikes followed some memes, like "LIBERATE" and "Sleepy Joe", but others fell flat: "fake news" and "shifty Schiff" saw little pickup on radio after a Trump tweet. Some examples of these two different fates a meme may experience are presented in Table 9.3. The "Crazy Bernie" meme is perhaps especially interesting: it saw significant pickup on talk radio, and Trump's tweets about it appear to have driven subsequent discussion there, but public radio saw no increase in discussion as a result of his tweets.

Though we only examined 17 memes, the apparent pattern among these fits that depicted in Table 9.3: newer memes spread more easily. "Fake news," "shifty" and others were old news by spring 2020, possibly contributing to Trump's regular coinage

Meme	Medium	Cume Effect	Max Effect	P-value
LIBERATE	elite	0.35190286	0.02450515	0.00000000
LIBERATE	decahose	0.11697983	0.00519893	0.00000000
Mini Mike	elite	0.06357159	0.00734856	0.00000000
Mini Mike	decahose	0.04637425	0.00489931	0.00000000
Radical left	elite	0.05015744	0.01278717	0.00000000
Radical left	decahose	0.03389270	0.00508587	0.00000000
Pocahontas	elite	0.10284493	0.02662222	0.00000000
Pocahontas	decahose	0.04171395	0.00819199	0.00000000
Sleepy Joe	elite	0.02657666	0.00735311	0.00000000
Sleepy Joe	decahose	0.03712987	0.00439364	0.00000000

Table 9.2: Granger causality and impulse response estimates for several Trump memes on Twitter, both elite and decahose. The "P-value" column is the p-value of the test for Trump Granger-causing the other medium listed; the "Max Effect" column gives the largest estimated response for any 15-minute period (over 24 hours); and the "Cume Effect" column is the estimated cumulative effect (over 24 hours) on meme share in the target medium.

of new phrases.

9.4 Discussion

From one point of view, these results confirm something obvious: Trump's tweets can set the agenda for media discussion. From another, they are surprising: memes can have sharply different effects, effects which appear to depend on the novelty of the memes.

These results, especially the importance of novelty, help contextualize and explain Trump's tweeting behavior. They fail to distinguish, however, between the ways his influence on the news agenda might be operationalized: through the bully pulpit in general or benefits of Twitter in particular. Future work building on this case study should include additional datasets to address this important question.

Meme	Medium	Cume Effect	Max Effect	P-value
LIBERATE	talk	0.03104563	0.00247952	0.00000000
LIBERATE	public	0.03204737	0.00587319	0.00000000
Sleepy Joe	talk	0.00780973	0.00136536	0.00000000
Sleepy Joe	public	0.00099987	0.00030669	0.00000000
Crazy Bernie	talk	0.02076583	0.00685966	0.00000000
Crazy Bernie	public	-0.00004106	0.00000618	0.61087453
Fake news	talk	0.00088124	0.00035847	0.52072995
Fake news	public	-0.00020509	0.00063505	0.31982424
Shifty	talk	0.00013343	0.00017770	0.12013993
Shifty	public	-0.00011999	0.00000505	1.00000000

Table 9.3: Granger causality and impulse response estimates for several Trump memes on radio, both public and talk. The "P-value" column is the p-value of the test for Trump Granger-causing the other medium listed; the "Max Effect" column gives the largest estimated response for any 15-minute period (over 24 hours); and the "Cume Effect" column is the estimated cumulative effect (over 24 hours) on meme share in the target medium.

Part V

Wrap-up

Chapter 10

Future Work and Limitations

While this thesis has broken some ground in large-scale mapping of radio, a great deal more remains to be done. Here we point the way toward some future work, and discuss the limitations of the analysis. Discussion is broken out into sections focusing on changes to respectively data and methodology, with an additional section focusing on future plans for the Trump case study.

10.1 Data Limitations

Limitations on the data we were able to collect impacted the analysis in several ways. Promising future directions of work, both to confirm and to expand on the work done here, suggest themselves if some of these are addressed.

Our analysis is most clearly affected by the quality of the radio data. While they suffice for the work done here, the transcripts are not perfect. Human transcription of so much audio is of course cost-prohibitive, but the word error rate could be brought lower than the current estimate of about 13% [6]. This error rate required, for instance, adding a large number of misrecognitions for "COVID-19" into meme keyword lists (see Appendix B). The schedule data presents another problem: not every snippet has a show assigned, and the data is not updated quickly enough.[1] Finally, our copy of the Radio-Locator data was from mid-2018, and some changes in station broadcast areas or formats may have been made since then. Repeating the analysis with better transcripts, more recent station metadata, and especially better schedule information would add confidence in its conclusions.

In a similar vein, the radio corpus was not collected exclusively for this thesis, and we had too few stations for fine-grained analysis of owners and geographies. Census regions are a very large, very coarse-grained unit of analysis, but there were simply too few stations to work at, say, the state level.

Finally, on the Twitter side, we only collected the follow graph among the elite-Twitter segment once, in early November 2019. This decision was for technical reasons: Many of the users in this segment have very high follower counts, and Twitter's API

[1] For example, a few of the shows scraped from station websites are known to have been off the air permanently at the time of the scrape.

does not allow retrieving follow information only for a subgraph of particular other users. Working with more regular pulls of the follow graph would thus have required too much engineering effort (or, if very frequent, run up against rate limits). In the future, collecting this data would allow very interesting temporal analysis. For example: memes move through the social network expressed by the follow graph, but how do they remodel it? Do follow relationships formed over certain memes cause a community to become more closely connected, or form bridging links to other communities?

10.2 Methodological Improvements

This thesis made certain methodological tradeoffs – usually to reduce engineering effort or work around data issues – which could be avoided with more time or better data.

First, we used only a 1/500th sample of the decahose, which is itself a 10% sample of Twitter. This was for convenience reasons: preprocessing data and counting memes for exploration is easier with smaller data. Using the full decahose would help greatly in detecting smaller memes; at a larger scale, the 100% firehose would provide a full view of Twitter discussion.

Next, the language models we used are very simple: Chapter 5 and Chapter 6 predict various quantities with linear regression on n-gram features. In a way, the use of these models rather than more modern neural ones makes the point more powerfully: if simple approaches uncover a relationship between a text stream and another quantity, it's a strong relationship. (Some testing suggested they also coped better with recognition error in radio transcripts.) On the other hand, better language modeling would likely uncover more subtle relationships.

Finally, we take a simple view of what constitutes discussion in Twitter and radio. Several pieces of analysis (e.g., Chapter 9, Chapter 7) look at distributions of topics defined by fixed keyword lists. From one standpoint, this is good: that the methodology is simple makes the results less open to question on that basis. On another view, it's too simple: we take no account of sentiment, different views of the same topic, subtopics, etc. A more sophisticated approach to characterizing discussion – even just more fine-grained topics – would pay dividends.

10.3 Trump Case Study

The likely causal effects identified in the Trump case study are an especially promising avenue for future research. The analysis here was an early-stage exploration of the spread and influence of memes, but there are a number of promising next steps:

Scope of memes The memes we examined were a convenience sample. A more systematic selection of memes, preferably over a longer period of time, would make the point more convincingly.

Scope of media This analysis examined only Twitter and radio. But Trump's tweets reach more broadly than this, affecting TV news, online news, Reddit, etc. Moreover, these media, influenced by Trump, may in turn influence each other, with Trump's combined effect on each reflecting multiple paths through a causal DAG.[2]

Causal inference, specifically Insofar as the question is not just whether the president has power to set the agenda, but how (i.e., the relative influence of Trump's role as president and the structure of Twitter), other comparisons are essential. We might, for instance, compare Trump's tweets to statements he makes on TV.

Causal inference, generally Attributing causality to Trump tweets here is plausible, because the data-generating process is one fairly predictable person tweeting, but it clearly doesn't scale. Causal attribution methods for more general online settings (without the presumption that a particular user's statements are automatically news) are hard, but essential for tracing the influence of memes.

[2]If, for example, Trump's Twitter audience engages with a Reddit thread sparked by a Fox segment about one of his tweets.

Chapter 11

Conclusion and Summary of Results

This chapter concludes our exploration of radio and its relationship to other media. We provide a short summary of our overall results, followed by more specific recaps of each part of the thesis.

11.1 Summary

This thesis explored the internal structure and relationships to other media of U.S. talk radio. We were motivated by the clear impact this medium has on American politics and culture [13] [1], and the lack of a prior large-scale, textual exploration of it.

The results confirm a nationalized and centralized radio ecosystem, with corporate ownership and national syndication networks playing a large role in the content of broadcasts. While local content still exists, and seems more diverse and more locally grounded than syndicated content, it accounts for a small minority of airtime.

The most surprising finding in this vein is the close connection between radio and Twitter, especially those elite parts of Twitter which host journalistic and political discussion. Twitter-side entities are predictable from the text of radio shows, and the content of radio is similar in broad ways to the content of elite Twitter.

We find less evidence of inter-medium influence over time, however. There are limited, mostly null, results in looking for Granger causality of Twitter on radio. The clear influence of President Trump's tweets on radio provides a possible example of a connection, though his preexisting agenda-setting power as president is a confounder[1].

All in all, we have convincing evidence that radio is not a separate medium, but an integral part of the nationwide media ecosystem. The patterns of influence between it and other media, however, deserve further study.

[1]That is, it's trivially an example of inter-medium influence, but the more interesting question is how much Twitter contributes over what Trump as President could achieve otherwise

11.2 Radio Alone

Our results in Part II suggest that the radio ecosystem operates in a mostly centralized way. Its programming aligns with corporate ownership and syndication networks, more than with other factors, and the topics this programming discusses are fairly homogeneous across the country. Local content and responsiveness to the surrounding geography, while clear and present, are on a smaller scale.

Of course, we're overloading the term "centralized," which means different things for programming and topics:

- The topic mix radio shows discuss is centralized in the sense that it's nationalized: surprisingly similar between stations, regardless of ownership, geography or format. There are some exceptions, discussed below, but the degree of variation is smaller than for programming and only weakly correlated with ownership and geography (Chapter 4).

- Stations' choice of syndicated programming is much more diverse, with more inter-station variation, but is centralized in the sense that (for non-public talk radio) it is unrelated to geography and substantially related to the stations' ownership. The five largest station owners in our sample, which account for a large share of stations, are publicly traded companies accountable to a nationwide set of stockholders. Meanwhile public radio, while operationally controlled by a more diverse set of owners, is also more heavily syndicated thanks to NPR and PRI, and thus more similar between stations (Chapter 3).

Though these forms of centralization are different, they both counsel against a vision of radio as a local media organ rooted in the surrounding community.

It's worth calling out here the differences between public radio and talk formats, as discussed in Chapter 3. Talk radio ownership is concentrated in a few large companies, with syndication and programming correlated with ownership; public radio station ownership is widely dispersed but programming is dominated by national networks. Along with some locally aligned content, both have strong centralizing tendencies, but those tendencies are different.

Finally, the centralized aspects of radio don't imply the absence of any local character. Local shows in general have a greater diversity of topics, and indeed the very largest syndicated shows are more homogeneous than other shows (Chapter 4). The greatest diversity of topics in local programming, in particular, is found in public radio. The partisanship of listening areas for such local shows, as measured by their 2016 election returns, is also far more closely related to their text than it is for syndicated shows (Section 6.3)[2]. Lastly, our analysis has focused on shows and topics, rather than perspectives on those topics – it's quite possible that, as the partisanship analysis suggests, a more sophisticated examination of sentiment would find nationalized topics and locally varying opinions about them.

[2]This piece of analysis is located in Chapter 6 because of its closely similar methodology, but is most conceptually relevant to Part II.

11.3 Twitter-Radio Interface

In a nutshell, our analysis in Part III reveals a strikingly close connection between elite Twitter[3] and radio. A similar social structure is clearly present in both:

- A measure of host ideology from Twitter's follow graph lines up well with common sense and with radio-side variables (Section 5.1);
- Communities in several Twitter graphs form readily interpretable groups of radio shows (Section 5.2);
- Similar network structure is apparent between Twitter's follow graph, the main social and information-spreading graph on that service, and the radio's most closely analogous graph structure: the co-airing graph of shows that appear on the same stations (Section 5.3).

Moreover, this common social structure is clearly closely related to the text of radio shows. A show's ideology and Twitter graph communities can be predicted with a high degree of accuracy from what hosts and guests say on the show (Section 6.2), implying that both the structure of radio as a medium and its content vary along dimensions we can identify from Twitter. Especially given that "elite" Twitter as defined here consists mostly of journalists and politicians, a natural conclusion of all this is that radio is tightly integrated into the national media ecosystem – indeed, more so than was previously known.

Further support for this view comes from the relationship of our Twitter variables to the ownership status of radio stations (Section 5.4). Even within the same conservative talk format, owners differ systematically and significantly in Twitter-side variables like the ideology of their hosts. There is much less variation along lines of geography (Section 5.5), consistent with a radio ecosystem driven by national media networks and national media corporations.

As a further test, when we compare the text of radio shows and the text of tweets, one of the most striking facts is the degree of similarity between the two media in the topics they discuss. The similarity extends even to within-show comparisons of hosts' on-air speech to their own tweets: the average host's tweets are about as similar to their on-air speech as any two shows are to each other. Radio and elite Twitter at the whole-medium level display an even greater degree of similarity. (See Section 7.2 for all of these results.)

We find notably less evidence, however, for temporal structure in the relationship between the two. It is clear that during our study period, elite Twitter led radio by several hours in discussion of certain important topics (politics, the economy and climate). But this pattern was not apparent in a number of other cases: the other 14 topics examined, first of all, many of them related to politics, but also in breakouts of radio by show and owner. While show and owner breakouts produced a few clear examples of Twitter discussion leading a show's or owner's discussion, results in most

[3]As defined in Chapter 2, a universe of journalists, politicians and commentators we use as a proxy for media discussion.

of the cases we looked at were null. What to make of these results is up for debate: it could be that there is not really any influence, that the topics we looked at were defined too broadly to capture it, or just that influence is restricted to certain topics. Further research ought to address this question.

Finally, we note an important caveat: our sample of Twitter-matched shows skews toward syndicated content, both because syndicated hosts are more likely to have Twitter accounts and because those accounts are easier to find. It's plausible, and indeed some of the evidence described here suggests, that local shows are less tightly coupled to Twitter and the national media ecosystem.

11.4 Case Studies

We present two case studies in Part IV, both of which focus on analyzing the cross-medium spread of memes. In one case (Chapter 8), we trace dueling COVID-19 memes from President Trump and the scientific community as they spread through the media ecosystem; in the other (Chapter 9), we take a closer look at the short-term consequences of Trump's tweets on Twitter and radio. Both identify cases of meme spread, and jibe with many of the conclusions drawn above – influential underlying social structures apparent in Twitter and radio, and some connections between their content – while challenging others: despite similarity in their topic mixtures, radio hosts' tweets and on-air speech took up certain memes quite differently.

But they also highlight an important phenomenon less clearly observed in earlier parts of this work: the disconnect between elite Twitter and Twitter at large (our decahose sample), with radio somewhere in between. Elite media Twitter differs in many ways from Twitter users in general. For example:

- Coronavirus discussion in all three media rapidly increased during March, but only in the decahose did it decline back to baseline afterward (Section 8.2);
- All of the memes in Section 8.3 saw sizable uptake in elite Twitter, but the decahose did not see nearly as great an increase in discussion of them;
- Within the decahose, all of those memes saw more uptake among more elite users – those with higher follower counts or more engagement (Section 8.3);
- Perhaps most convincingly, estimating impulse responses to Trump's tweets frequently produces effects several times larger in elite Twitter than in the decahose (Section 9.3).

These results about the decahose challenge the public opinion model from Section 1.2. If our view of the world is that elite discourse leads to changes in public opinion, it's challenging to see that a proxy for public opinion doesn't respond much if at all to a measure of elite discourse. The Twitter user population is not the general public, but the lack of a relationship is still in tension with this model. How to reconcile the two is an open question, with explanations from measurement error to the unrepresentativeness of Twitter to a true lack of relationship in contention.

11.5 Remarks

Radio has been influential in American life and around the world for decades. From Franklin Roosevelt's fireside chats to Rush Limbaugh's more acerbic commentary, the medium has had political impact all along the way. It is gratifying to be able to contribute to understanding it.

It is also worth pointing out the potential policy relevance of these findings. Despite previous waves of deregulation, there are still regulations in the U.S. restricting any one entity's ownership of multiple radio and television stations in the same market [100]. Insofar as these regulations are intended to ensure viewpoint diversity, our results suggest that in a nationalized media environment they do not go far enough. Station owners are not the only central influence on their stations' viewpoints, and requiring minimum amounts of locally produced content would help ensure diverse and locally responsive views receive an airing.

Appendices

Appendix A

Text Standardization

Here we describe the text standardization process applied to align the formats of Twitter and radio data.

A.1 Phrase detection

We used the popular Gensim package [75] to identify commonly used phrases (in the radio data only), producing a list of 8100 common phrases. A "phrase" here was a two-word collocation, or pair of frequently co-occurring words, where words were identified by splitting on whitespace. (Recall from Section 2.5 that the radio transcripts themselves are produced as a sequence of space-separated words, so we can avoid any more sophisticated tokenization strategies.) We required a minimum of 50 occurrences for a phrase to be detected, as well as a minimum NPMI score [101] of 0.75.

These detected phrases, plus a list of 700,000 English Wikipedia article titles, were merged together in the radio data. That is, for example, occurrences of "supreme court" might be merged into occurrences of the single token "supreme_court". Most, though not all, of the article titles and other detected phrases were compound nouns or noun phrases, supporting the decision to merge them together.

Phrases were generally but not always combined in Twitter data as well, depending on the particular analysis.

A.2 Simple Preprocessing

We performed several basic steps to align Twitter data with the radio data's format. Text was converted to lowercase, punctuation[1] was removed, multiple consecutive spaces were replaced by one space, and whitespace at the beginning and end of tweets was stripped.

We handled non-ASCII characters by transliterating them on a best-effort basis to the most similar ASCII character (e.g., ü becomes u), using the unidecode package[2]

[1] Specifically Python's string.punctuation list of characters.

[2] https://pypi.org/project/Unidecode/

for Python. If no character was suitable, the non-ASCII character was dropped.

A.3 Twitter-Only Features

We removed several textual features which are Twitter-specific and do not occur in spoken language. Specifically, we removed

Retweets Twitter's API returns the bodies of retweets with "RT @USERNAME: " prepended. We dropped any text matching the regular expression "RT @[a-zA-Z0-9_]+: " from tweet bodies.

URLs We detected URLs by regular expression matching[3], and dropped them from tweet bodies. Domain names which were not formatted as a full URL with a scheme:// prefix were ignored. Because Twitter replaces the user-inputted text of links with a shortlink through t.co, we searched only for http and https URLs. 70 out of approximately 7 million tweets between the decahose and elite Twitter contained URLs which did not match our regex, and we dropped these tweets.

Usernames We removed Twitter usernames from tweet bodies. Any text matching the regular expression "@[a-zA-Z0-9_]+" was dropped.

Ignoring these features removes a significant amount of signal (URLs perhaps especially) but simplified the process of standardizing with radio content and removed certain researcher degrees of freedom in doing so.

A.4 Hashtag Segmentation

We segmented hashtags into words according to a simple algorithm.

First was the problem of detecting hashtags. Rather than using the full list of hashtag entities returned by Twitter's API (which may contain hashtags with non-ASCII characters, numbers, etc), we detected hashtags with a simple regular expression[4]. Hashtags matching this expression at least in principle may decompose into words without further preprocessing, like the number conversion discussed below.

Each hashtag in this set was broken into words greedily right-to-left, with hashtags whose text did not decompose entirely into words in the vocabulary[5] dropped.

More specifically, a verbal description of the algorithm is as follows:

1. Put the hashtag text (minus the "#") in a buffer, and create an (initially empty) stack of words. Initialize a pointer to the end of the text buffer.

2. While buffer not empty:

[3]The specific regular expression used was "https?:\/\/(www\.)?[-a-zA-Z0-9@:%._\+~#=]{1,256}\.[a-zA-Z0-9()]{1,6}\b([-a-zA-Z0-9()@:%_\+.~#?&//=]*)".

[4]"\#+([a-z]+)", and recall the text has been lowercased.

[5]The vocabulary was the spell-checker vocabulary shipped with Ubuntu at `/usr/share/dict/words`.

(a) If the pointer is at the start of the buffer and the buffer has positive length, we have text that can't be assigned to a vocabulary word. Error out and return an empty stack.

(b) Look at the text to the left of the pointer location.

(c) If that text forms a word in the vocabulary, remove it from the buffer, append it to the words stack, and move the pointer back to the end of the buffer.

(d) Else move the pointer 1 character to the left and repeat.

3. Return the words stack.

This is a simple but robust baseline that handles most hashtags well. Most importantly, it avoids returning words which are parts of other words when the larger word could be identified instead. For example, in "#ImpeachTrumpNow", we won't return 'imp' as a word present in the hashtag. It isn't perfect, of course, and in particular won't identify two words next to each other if together they form a different word. For example, if in context "#SpearMintParty" were about a party with spears and mint, we'd incorrectly say the word that's present is "spearmint". Similarly, "#CompositionAlly" would be returned as "compositionally". As the contrived nature of these examples suggests, this is not a very big disadvantage. It's less common for two words next to each other to coincidentally form a third word than for words to be built compositionally out of parts which may themselves be words.

A.5 Autocomplete

Because of a misconfiguration in the decahose ingest, about 21% of the decahose data was truncated at 140 characters. This truncation sometimes resulted in final partial words within tweet bodies.[6] Tweets at 140 characters are so short that this results in a meaningful loss of signal. In cases of partially present final words, we attempted to recover some of the missing signal by using a hidden-Markov model trained on the data to infer the missing word.

A.6 Number Formatting

We processed numbers written with numerals in tweets to be in the same format as in the radio data. A "number" was anything matching the regular expression "\b([0-9]*\.?[0-9]+)\b", which in particular means that any negative signs which might have been present were ignored.

Numbers were converted to the usual English pronunciation as cardinal numbers via the num2words package for Python[7], except that numbers between 1900 and 2030 were rendered as years are usually pronounced in English. (For example, 1945 becomes

[6] This is so even though all of the data is from after Twitter began allowing 280-character tweets.
[7] https://pypi.org/project/num2words/

"nineteen forty five" rather than "one thousand nine hundred forty five".) Numbers with fractional parts were rendered with "point" and the succeeding string of digits (for example, "392.97" becomes "three hundred and ninety two point nine seven") except that numbers with all zeros after the decimal point were treated as integers, with no "point..." part.

Appendix B

Keyword Lists

Here we present the lists of keywords used in several previous chapters.

B.1 Case Study Memes

Memes and keywords as used in Chapter 8 and Chapter 9 are listed here. Note that when a keyword contains a space, we counted as matching that keyword any string which replaced some number of these spaces with underscores. So, for example, "slow the spread" would match "slow the spread", "slow_the spread", "slow the_spread" and "slow_the_spread".

covid "corona", "virus", "coronavirus", "rona", "covid", "pandemic", "covid nineteen", "covered nineteen", "covert nineteen", "coven nineteen", "coated nineteen", "coping nineteen", "corporate nineteen", "called nineteen", "coleman nineteen", "cove nineteen", "culprit nineteen", "coffee nineteen", "clothing nineteen", "coded nineteen", "kobe nineteen", "komen nineteen", "kovac nineteen", "kobe nineteenth"

fake_news "fake news"

radical_left "radical left", "radical liberals"

do_nothing_democrats "do nothing democrats"

enemy_of_the_people "enemy of the people"

liberate "liberate"

mini_mike "mini mike"

sleepy_joe "sleepy joe"

crazy_bernie "crazy bernie"

pocahontas "pocahontas"

shifty "shifty"

hoax "hoax"

invisible_enemy "invisible enemy"

morning_psycho "morning joe", "morning psycho", "morning joke"

voter_fraud "voter fraud"

nyt "new york times"

drudge "drudge"

2a "second amendment", "2a"

slow_the_spread "slow the spread"

social_distancing "social distancing", "socially distancing", "physical distancing"

flatten_the_curve "flatten the curve", "flattening the curve", "flattened the curve", "flat the curve", "flam the curve", "finding the curve", "flooding the curve", "flattening the curves", "flap the curve", "fly the curve", "fighting the curve", "flattening the curb", "flatly the curve", "fled the curve", "flattening of the curve"

B.2 Topic Keywords Seed Terms

This section presents the seed terms for lists of keywords used in Chapter 4 and Chapter 7.

covid19 "covid", "coronavirus", "respiratory_illness", "reopening", "pandemic", "corona_virus", "reopen"

economy "consumer_price", "interest_rates", "unemployment", "jobs"

trump "trump"s", "trump_administration", "trump", "president_trump", "tromp", "donald_trump", "donald"

inclusivity "equality", "diversity", "accessibility", "inequality", "discrimination", "inclusivity"

guns "guns", "second_amendment", "gun_rights", "gun_control"

healthcare "health_insurance", "mental_health", "health_care", "healthcare", "nurse", "hospital", "doctor"

climate "climate", "climate_change", "sustainability", "global_warming", "environment"

drugs "overdose", "fentanyl", "opioid_crisis", "opioids", "heroin"

other_social_issues "affirmative_action", "transgender_rights", "abortion_rights", "interracial_marriage", "gay_marriage", "abortion", "religious_freedom", "gay_rights", "birth_control"

education "class_size", "school_district", "special_education", "i.e.p.s", "schools", "public_schools", "education"

sports "hockey", "basketball", "soccer", "lacrosse", "tennis", "football"

politics "politics", "president", "republican", "senate", "election", "democratic", "governor", "congress", "democrat"

crime "murder", "robbery", "crime", "assault", "home_invasion"

immigration "immigrant", "immigrants", "immigration"

weather "snow", "weather", "cloudy", "rain", "temperature", "sunny"

B.3 Topic Keywords Expanded Terms

These keyword lists were expanded from the seed terms in the previous section based on a word2vec model; further details can be found in Chapter 4. The actual counting of keywords for each topic, whether in radio or Twitter, included both these expanded terms and the seed terms. Unlike in the case studies, we matched exactly, without considering variants that differed only in substituting underscores for spaces. (The different treatment in the two cases was because of phrase detection. These keywords are all known to be in the vocabulary because they come from a word2vec model trained on the radio data, but the case study keywords were collected in a more manual way and may have left out some variants.)

covid19 "covert", "coven", "virus", "corona", "outbreak", "kobe", "coping", "the_disease", "reopened", "illness", "kovac", "endemic", "shutdown", "spread", "copa", "shutdowns", "public_health", "cupboard", "restarting", "easing", "infection", "the_crisis", "nineteen", "three_phase", "coded", "guidelines", "health_crisis", "coated", "death_toll", "infected", "governors", "lockdown", "the_infection", "infections", "governor", "anthony_fauci", "health_department", "infectious_disease", "fallout", "govan", "economy", "task_force", "loosening", "testing", "from_kobe"

economy "unemployment_benefits", "jobless", "jobless_claims", "unemployment_rate", "unemployed", "unemployment_insurance", "employment", "the_fed", "employment_rate", "labor_market", "interest_rate", "consumer_confidence", "unemployment's", "inflation", "economists", "oil_prices", "economic_growth", "fed's", "furloughed", "central_bank", "retail_sales", "yields", "consumer_spending", "economy", "numbers", "full_employment",

"treasuries", "borrowing", "labor_force", "paychecks", "recession", "benchmark", "treasury", "service_sector", "federal_reserve", "unemployment_benefit", "wages", "workforce_development", "gdp", "tax_revenue", "gdp_growth", "incomes", "earnings", "reserve's"

inclusivity "racism", "racial", "inclusion", "gender_equality", "sexism", "equal_rights", "racial_equality", "oppression", "sexual_orientation", "disparities", "income_inequality", "economic_inequality", "transgender", "sustainability", "semitism", "women's_rights", "intolerance", "segregation", "white_supremacy", "social_justice", "gender", "feminism", "sex_discrimination", "environmental_degradation", "gender_identity", "nondiscrimination", "environmental_justice", "discriminatory", "cultural", "affirmative_action", "identity_politics", "gender_diversity", "civil_rights", "ethnic_diversity", "reproductive_justice", "activism", "fairness"

guns "gun", "firearms", "assault_weapons", "gun_safety", "gun_violence", "gun_laws", "ammunition", "rifles", "amendment", "confiscation", "n._r._a.", "firearm", "shootings", "open_carry", "assault_rifles", "red_flag", "handguns", "ammo", "hand_gun", "semiautomatic", "gun_law", "constitutional_carry", "handgun"

healthcare "nurses", "hospitals", "doctors", "physician", "emergency_room", "medical_center", "healthcare_system", "health_system", "patient", "nurse_practitioner", "primary_care", "patients", "critical_care", "workers", "physicians", "clinicians", "medical_care", "urgent_care", "medical", "the_doctors", "nursing", "the_hospital", "childcare", "child_care", "providers", "emergency_department", "palliative_care", "i._c._u."

climate "environmental", "climate_science", "greenhouse_gas", "the_environment", "fossil_fuel", "deniers", "fossil_fuels", "climate_action", "climate_justice", "emissions", "environmental_degradation", "greta", "deforestation", "pollution", "carbon_emissions", "warming", "greenhouse_gases", "environmentalists", "global_cooling", "renewable_energy", "environmental_justice", "biodiversity", "catastrophe", "oceans", "activists", "environmental_movement", "carbon_dioxide", "environments", "science_policy", "overpopulation", "activism", "inequality", "planet", "ecosystem", "thunberg", "united_nations", "economic_inequality", "diversity", "think_global", "environmental_issues", "youth_activism", "climate_warming", "atmosphere", "ecosystems", "environmentalist", "sea_level", "extinction", "air_pollution", "carbon_footprint", "plastic_pollution", "overfishing", "hurricanes", "polar_bears", "indigenous_rights", "circular_economy", "population_growth", "gun_violence", "carbon_tax", "global_challenge", "earth's_climate", "atmospheric", "urban_sprawl", "droughts", "ecosystem_services", "greenhouse_effect", "green_energy", "ocean_acidification", "shareholder_resolution", "environmental_policy", "carbon_price", "carbon_sequestration", "income_inequality",

"acidification", "earth_day", "new_deal", "green_infrastructure", "managed_retreat", "habitat_destruction", "algal_blooms", "intergovernmental", "paris_agreement", "sustainable", "climate_finance", "food_systems", "environmental_impact", "sundberg", "emission", "earth's", "population_control", "summit", "alarmist"

drugs "opioid", "oxycontin", "opioid_epidemic", "overdoses", "methamphetamine", "narcotics", "cocaine", "fenton", "addiction", "illegal_drugs", "opiates", "painkillers", "purdue_pharma", "painkiller", "carfentanil", "drug", "overdosing", "addicted", "drugs", "crack_cocaine", "oxycodone", "medications", "overdosed", "narcotic", "drug_overdose", "drug_distribution", "mac_miller", "cigarettes", "alcohol", "alcoholism", "suicide", "methamphetamines", "methadone", "marijuana", "drug_abuse", "narcan", "morphine", "vaping", "opioid_overdose", "cough_suppressant", "pain_medication", "meth", "the_addiction", "prescription", "illicit", "opiate", "drug_addiction", "pills", "alcohol_abuse", "purdue", "drug_trafficking", "sackler", "heroine", "prescribing", "percocet", "black_market", "addictive"

other_social_issues "abortions", "women's_rights", "transgender", "reproductive_rights", "religious", "traditional_marriage", "planned_parenthood", "discrimination", "civil_rights", "conversion_therapy", "abortion_law", "marriage_equality", "establishment_clause", "homosexuality", "equal_rights", "reproductive_health", "antidiscrimination", "reproductive_justice", "contraception", "sexual_orientation", "amendment", "sex_discrimination", "pro_choice", "originalists", "homosexual_agenda", "religion", "roe", "lesbian", "free_speech", "compelled_speech", "citizens_united", "infanticide", "death_penalty", "homosexual", "heterosexuality", "marriage", "a._c._l._u.", "gender_equality", "legalization", "bisexual", "school_choice", "first_amendment", "segregation", "sex_education", "nondiscrimination", "discriminatory", "equal_protection", "activists"

education "public_school", "teachers", "students", "colleges", "elementary_school", "childcare", "universities", "superintendent", "middle_school", "teachers'", "elementary_schools", "elementary", "the_districts", "charter_school", "classrooms", "districts", "charter_schools", "the_students", "distance_learning", "teacher", "public_education", "private_school", "school_board", "the_district", "u._s._d.", "high_school", "campuses", "school", "independent_school", "child_care", "state_school", "online_learning", "school's", "superintendents", "curriculum", "daycare", "school_year", "daycares", "preschools", "kindergarten", "teacher's", "ethnic_studies", "instruction", "preschool", "community_college", "dual_language", "student", "i._s._d.", "highschool", "early_childhood", "after_school", "standardized_testing", "school_choice", "tutoring", "student's", "general_education", "semester", "achievement_gap", "standardized_tests", "elementary_education", "day_school", "educational", "lightfoot", "tuition", "libraries", "digital_learning", "administrators", "higher_education", "district's", "pub-

lic_universities", "student_activities", "playgrounds", "twelfth_grade", "undergraduate", "career_development", "the_academic", "affordable_housing", "first_grade", "sharkey", "catholic_school", "schoolers", "learning", "learners", "driver's_education", "public_library", "standardized_test", "college", "classes", "youth_programs", "underprivileged", "unified", "superintendent's", "liberal_arts", "truancy", "the_student", "sex_education", "magnet_school", "physical_education", "coursework", "diocese", "educators", "school_reform", "public", "i._e._p.s", "community_colleges", "churches", "mark_fuller", "textbooks", "classroom", "educational_leadership", "new_school", "the_city", "curriculums", "buettner", "proficiency", "students'", "instructional"

sports "volleyball", "college_football", "baseball", "football_team", "rugby", "hockey_team", "field_hockey", "tournament", "league", "big_ten", "championships", "b._y._u.", "postseason", "football_games", "women's_basketball", "playoffs", "championship", "players", "finals", "basketball's", "football's", "game", "nfl", "little_league", "tournaments", "foot_ball", "northwestern_wildcats", "n._b._a.", "games", "horizon_league", "coach", "n._f._l.", "n._c._a._a.", "wildcats", "national_championship", "player", "byu", "golf", "houston_dynamo", "game_on", "collegiate", "coaches", "highschool", "n._h._l.", "march_madness", "atlanta_united", "head_coach", "high_school", "southern_league", "athletic_conference", "hawkeye", "the_browns", "wrestling", "canadian_football", "northeast_conference", "college_hockey", "providence_bruins", "n._c.", "summit_league", "college_basketball", "a._c._c.", "boys'", "junior_college", "portland_timbers", "girls'", "football_player", "leagues", "american_football", "real_football", "football_league", "conference_usa", "the_cardinals", "exhibition_match", "x._f._l.", "baylor", "scoreboard", "steelers", "army_football", "professional_football", "lakers", "premier_league", "penn_state", "play", "hitting_streak", "fighting_irish", "night_game", "big_tent", "raiders", "pittsburgh_steelers", "notre_dame", "minor_league", "stadium", "buckeyes", "junior_varsity", "professional_baseball", "toronto_raptors", "bulldogs", "undefeated", "athletics", "dodgers", "miami_hurricanes", "huskies", "softball", "flag_football", "the_league", "state_champs", "philadelphia_flyers", "playing", "world_cup", "utah_football", "aggies", "urban_meyer", "indoor_soccer", "panthers", "birmingham_bowl", "playoff", "varsity", "central_michigan", "cardinals", "touch_football", "the_championship", "played", "champions_league", "world_series", "green_bay", "minnesota_lynx", "athletic_director", "teams", "women's_lacrosse", "tigers", "the_eagles", "steve_spurrier", "l._s._u.", "quarterback", "the_vikings", "quarterfinals", "neal_brown", "the_jets", "minnesota_twins", "badger", "league's", "men's", "lebron_james", "lsu", "clemson", "bass_fishing", "mccaffrey", "lane_stadium", "header", "women's", "game_seven", "the_game", "great_game", "seahawks", "professional_sports", "point_guard", "p._g._a.", "southeastern_conference", "old_dominion", "virginia_cavaliers", "atlanta_braves", "women's_tennis", "missouri_tigers", "semifinals", "tiger_woods", "gym-

nastics", "minnesota_vikings", "pga_tour", "champ", "the_packers", "game_one", "nashville_predators", "night_games", "devin_bush", "broncos", "baylor_bears", "embry", "king_philip", "college_baseball", "bruce_cassidy", "milwaukee_bucks", "kirby_smart", "quarterbacks", "enterprise_center", "top_seed", "s._e._c.", "coaching_staff", "miami_dolphins", "rupp_arena", "busch_stadium", "nfc_north", "drake_bulldogs", "the_stands", "ben_simmons", "larry_smith", "final_score", "celtics", "bradley_braves", "ryan_day", "jeremy_pruitt", "opener", "jaguars", "national_team", "mcleod_center", "ohio_state", "chip_kelly", "vikings", "peyton_manning", "bowl_game", "ice_hockey", "golf's", "big_dance", "duke", "the_opener", "cubs", "croquet", "the_penguins", "the_redskins", "lsu_tigers", "hockey's", "philadelphia_eagles", "wisconsin_badgers", "syracuse", "james_franklin", "soldier_field", "ben_jacobson", "training_camp", "remy_martin", "bangles", "detroit_tigers", "kyrie_irving", "super_bowl", "the_falcons", "pistons", "braves", "scott_frost", "trojans", "the_bangles", "anthony_davis", "bowl_games", "camp_randall", "buzz_williams", "panther", "houston_cougars", "falcons", "crosse", "chipper_jones", "u._c._l._a.", "umpires", "winter_sports", "calgary_flames", "derrick_rose", "student_athlete", "astros", "pat_fitzgerald", "ball_game", "drew_brees", "u._s._c.", "robert_thomas", "jim_leonard", "cowboys", "rugby_league", "special_teams", "novak_djokovic", "golden_knights", "redskins", "standings", "a_league", "badgers", "ryan_center", "chuck_long", "wheelchair_basketball", "coach's", "final_four", "tom_brady", "derek_mason", "white_sox", "caffrey", "the_quarterback", "fantasy_football", "college_softball", "kemba_walker", "cleveland_brown", "w._n._b._a.", "e._s._p._n.", "kurt_warner", "all_sport", "the_islanders", "chicago_bears", "baker_mayfield", "water_polo", "cleveland_browns", "blue_jays", "riddle", "jerseys", "auburn", "husky_stadium", "goalkeeper", "blue_bloods", "the_athletic", "rocky_top", "the_player", "pickup_game", "professional_basketball", "antonio_brown", "arenas", "women's_college", "cavaliers", "mel_allen", "washington_nationals", "serena_williams", "niners", "football_conference", "season's", "aaron_rodgers", "dustin_johnson", "razorbacks", "sixers", "the_cowboys", "milwaukee_brewers", "grass_court", "virginia_tech", "running_back", "nfl_playoffs", "villanova", "nfl's", "wiffle_ball", "small_forward", "defensive_team", "school_song", "wide_receiver", "robert_morris", "game_show", "hayden_fry", "sidney_crosby", "offseason", "rams", "athlete", "bruins", "skins_game", "linebacker", "philadelphia_union", "championship_week", "hank_gathers", "russell_wilson", "rutgers_football", "ball", "baseballs", "competitions", "aaron_brooks", "scotty_bowman", "bob_cousy", "marching_band", "usa_basketball", "max_kepler", "teammates", "chicago_blackhawks", "michael_potter", "lpga_tour", "golden_state", "first_period", "automatic_bid", "jerry_stackhouse", "thursday's_game", "nick_saban", "defensive_back", "padres", "dolphins", "the_varsity", "devin_white", "iowa_football", "hawks", "texas_tech", "andrew_mccutcheon", "freddie_kitchens", "buffalo_bills", "houston_astros",

"andy_pettitte", "longhorns", "luca", "dave_winfield", "elite_eight", "laker", "jim_harbaugh", "senior_bowl", "good_game", "roger_federer", "expansion_draft", "bulls", "the_cardinal", "roller_hockey", "racquetball", "championship_ring", "detroit_pistons", "game_time", "touchdowns", "baseball_tonight", "pat_riley", "mike_riley", "halftime", "season", "play_away", "boston_college", "soccer_city", "seattle_mariners", "game_day", "nfl_draft", "team's", "columbus_clippers", "sports", "alvin_kamara", "lion's", "boston_bruins", "ole_miss", "gary_woodland", "opening_night", "third_base", "atlanta_hawks", "twins", "florida_gators", "rodeo", "mike_lucas", "force_five", "blackhawks", "megan_rapinoe", "serino", "cavs", "defensive_line", "scrimmage", "richard_pitino", "detroit_lions", "defensive_coordinator", "beach_volleyball", "contact_sports", "game's", "clippers", "manchester_city", "troy_trojans", "sports_league", "clasby", "triple_jump", "heinz_field", "john_wooden", "mack_brown", "oakland_a's", "legion_field", "mark_loretta", "adam_silver", "rick_pitino", "olympic_sports", "expansion_team", "the_buccaneers", "boston_celtics", "bush_stadium", "m._v._p.s", "three_seasons", "kevin_callahan", "contact_sport", "barkley", "tailgating", "heisman_trophy", "figure_skating", "player_one", "fc", "winning_percentage", "team"

politics "republicans", "democrats", "g._o._p.", "us_senate", "senator", "mitch_mcconnell", "republican_party", "the_president", "democratic_party", "presidential", "congressional", "legislative", "majority_leader", "congressman", "chuck_schumer", "joe_biden", "nancy_pelosi", "legislature", "caucus", "minority_leader", "president_trump", "presidential_candidate", "bernie_sanders", "bipartisan", "donald_trump", "judiciary", "lawmakers", "impeachment", "reelection", "elections", "presidential_nomination", "trump", "john_cornyn", "partisan", "pelosi", "us_senator", "biden", "adam_schiff", "the_election", "party's", "president's", "elizabeth_warren", "state_senator", "congressional_district", "nomination", "front_runner", "democratic_congress", "congresswoman", "hillary_clinton", "mitt_romney", "white_house", "vote", "trump's", "the_vote", "presidents", "kamala_harris", "democrats'", "trump_administration", "congressmen", "inquiry", "voters", "presidential_nominee", "primaries", "political", "committee", "the_inquiry", "legislator", "the_house", "representatives", "elected", "politician", "special_election", "presidency", "doug_collins", "schumer", "john_hickenlooper", "incumbent", "campaigning", "cory_booker", "lindsey_graham", "state_legislature", "bill_weld", "democrat_party", "sanders", "votes", "speaker", "jim_sensenbrenner", "vice_president", "general_election", "charles_schumer", "opposition", "city_council", "capitol_hill", "supreme_court", "d._f._l.", "us_congress", "moderates", "dan_bishop", "governors", "brock_obama", "barack_obama", "primary_election", "the_congressman", "us_senators", "legislation", "executive_branch", "the_speaker", "senate's", "primary", "chief_justice", "appropriations", "state_senate", "impeached", "ted_cruz", "impeach", "cory_gardner",

"andrew_cuomo", "steny_hoyer", "special_session", "kevin_mccarthy", "trump_impeachment", "tony_evers", "mark_sanford", "committees", "debate", "governor's", "sensenbrenner", "joe_manchin", "minnesota_senate", "legislative_session", "romney", "david_ralston", "beto_o'rourke", "biden's", "committee's", "incumbents", "legislators", "arizona_senate", "representative", "hopefuls", "the_vice", "opponent", "bernie", "bob_menendez", "lawmaker", "sen", "wisconsin_senate", "ed_markey", "parliamentary", "political_parties", "ted_lieu", "state_assembly", "conservatives", "legislatures", "marco_rubio", "dick_durbin", "virginia_senate", "libertarian", "primary_challenge", "tim_ryan", "voting", "greg_stanton", "chuck_grassley", "gwen_moore", "debbie_stabenow", "national_convention", "wisconsin_legislature", "brendan_boyle", "chris_murphy", "establishment", "collin_peterson", "mccready", "eric_swalwell", "chairmanship", "lena_taylor", "treasury_secretary", "juppe", "jamie_raskin", "foreign_policy", "electoral_college", "senators", "darrell_issa", "mayor", "president_obama", "delegation", "campaign", "the_state", "district", "dems", "rob_bishop", "debates", "liz_cheney", "candidacy", "referendum", "cuomo", "angie_craig", "reelected", "ranking_member", "lieutenant_governor", "redistricting", "tom_o'halleran", "opponents", "david_cicilline", "terry_mcauliffe", "mayoral", "delegates", "hakeem_jeffries", "runoff_election", "abigail_spanberger", "veto_override", "joe_walsh", "green_party", "attorney_general", "voter", "casey_becker", "upper_chamber", "bill_clinton", "tammy_baldwin", "administration's", "bill_flores", "subpoena", "tim_walz", "super_tuesday", "ballot", "parliament", "gary_peters", "richard_neal", "corey_lewandowski", "electoral", "sylvia_garcia", "zell_miller", "schiff", "politicians", "governorship", "obama", "far_left", "voice_vote", "kevin_brady", "nadler", "guy_reschenthaler", "constituents", "richard_shelby", "ben_cardin", "fred_upton", "formalized", "andy_biggs", "whitney_williams", "tulsi", "jim_himes", "gun_control"

crime "homicide", "aggravated", "felony", "second_degree", "the_murder", "the_shooting", "aggravated_assault", "first_degree", "convicted", "manslaughter", "charged", "arrested", "attempted_murder", "a_crime", "carjacking", "suspects", "arson", "armed_robbery", "kidnapping", "hate_crime", "killings", "burglary", "false_imprisonment", "resisting_arrest", "assaults", "double_murder", "arrest", "crimes", "shooting", "charges", "bank_robbery", "capital_murder", "disorderly_conduct", "not_guilty", "police_officer", "suspect", "life_sentence", "child_pornography", "carjack", "conviction", "child_endangerment", "indecent_assault", "grand_larceny", "felonies", "arrests", "unarmed", "sexual_battery", "alleged", "voluntary_manslaughter", "reckless_driving", "incident", "police", "allegedly", "violent", "vehicular_homicide", "misdemeanor", "molestation", "assaulting", "aggravated_murder", "felon", "pistol_whipped", "fatally", "statutory_rape", "murdering", "stealing_cars", "criminal_mischief", "the_killing", "vulnerable_adult", "larceny", "traffic_stop", "violent_crime", "extortion", "murderer", "gang_related", "criminal", "official_misconduct", "guilty", "child_abuse", "ac-

cused", "gunfire", "possessing", "murders", "offenses", "homicides", "narcotics", "child_rape", "state_trooper", "custody", "arraignment", "strangling", "torture", "rape", "drug_possession", "negligent_homicide", "vandalism", "murdered", "sentencing", "detective_sergeant", "dog_attack", "crime_spree"

immigration "undocumented", "migrants", "illegal_immigration", "migrant", "refugees", "deportation", "asylum_seekers", "undocumented_immigrants", "refugee", "asylum", "illegal_alien", "citizenship", "emigration", "deportations", "green_card", "immigration_policy", "el_salvador", "the_immigrants", "visas", "us_immigration", "detention", "mexico", "the_asylum", "mexican_american", "border", "mexicans", "indigenous", "immigration_law", "amnesty", "guatemalan", "immigrate", "mexican", "deported", "illegal", "central_america", "detained", "detainees", "salvadoran", "discrimination", "open_borders", "illegally", "illegal_immigrants", "illegals", "foreign_workers", "immigration_enforcement", "anchor_baby"

weather "partly_cloudy", "mostly_sunny", "breezy", "showers", "skies", "windy", "sunshine", "clouds", "scattered", "cooler", "thunderstorms", "drizzle", "chilly", "thunderstorm", "partly", "temperatures", "patchy", "the_rain", "gusty", "three_degrees", "flurries", "blustery", "overcast", "cloud_cover", "unseasonably", "hive", "sunny_afternoon", "more_rain", "degrees", "warmer", "no_rain", "precipitation", "fog", "heavy_rain", "hides", "today_tonight", "freezing_drizzle", "downpours", "for_tomorrow", "rainfall", "chillier", "freezing_rain", "spotty", "chili", "storms", "milder", "forties", "humid", "the_hive", "six_degrees", "elevations", "chile", "the_clouds", "topping_out", "colder", "warm", "afternoon", "highs", "gusts", "shower", "thunder", "clouding", "snowfall", "slight", "the_overnight", "inland", "fifties", "daybreak", "hives", "north_wind", "easterly", "snow_fall", "upper", "tomorrow", "temps", "night_below", "wind_advisory", "taper", "raindrops", "dreary", "nine_tonight", "mid", "wins_out", "sun", "grand_junction", "the_thunderstorm", "night_alone", "scatter", "weather_system", "forecast", "south_wind", "freezing_fog", "warming_up", "rainy", "tapering", "sixties", "rain_fall", "heaviest", "after_midnight", "tapers"

B.4 N-gram Exclude Lists

This section lists the n-grams excluded from use as predictors in Chapter 6. There are two lists: the "base" list for prediction of ideology and follow community from show text, and the "extended" list for prediction of shows' election returns from show text. These lists, while they reflect the specific formatting of our radio data, include several kinds of n-grams: the names of hosts (e.g., "rush_limbaugh"), the names of stations (e.g., "_w _c" et al), the names of shows (e.g. "all_things considered"), certain other fixed phrases characteristic of particular radio stations (e.g., NPR's "your support" and "other contributors"), and a number of geographic indicators ("massachusetts", "in san_diego", etc).

The base list contains: '_a _b', '_b _c', '_c _a', '_c news', '_c news_radio', '_c radio', '_d _a', '_f _i', '_h _m', '_h', '_p _r', '_r and', '_r news', '_v _u', '_w _c', '_y', 'a _b', 'a _c', 'a r' 'abc_news', 'all things', 'all_things considered', 'all_things', 'and kathleen_collins', 'and npr', 'and w', 'are i', 'ari_shapiro', 'as npr', 'at cpr', 'at k', 'at npr', 'b _b', 'b', 'bbc', 'brian kill', 'catherine', 'cbs', 'cbs_news', 'coast am', 'the_coast to', 'coast to', 'to coast', 'collins wealth', 'colorado public_radio', 'com support', 'comes from', 'considered from', 'considered', 'contributors include', 'cpr dot', 'dana', 'dave', 'david_greene', 'david_folkenflik', 'donnell show', 'dot org', 'eight five', 'eighty nine', 'eighty w', 'eighty_eight point', 'eighty_eight', 'for npr', 'for on_point', 'fox is', 'fox', 'fox_nation', 'fox_news radio', 'fox_news', 'fox_sports radio', 'fox_sports', 'fresh_air', 'from npr', 'g p', 'georgia public_broadcasting', 'go fox_sports', 'gordon deal', 'ground_zero', 'hannity', 'heart radio', 'heart radio_station', 'here now', 'i heart', 'is all_things', 'is fresh_air', 'is made_possible', 'is morning_edition', 'is npr', 'is on_point', 'is supported', 'j c', 'j', 'josh', 'joshua_johnson', 'justin' 'k _f', 'k u', 'k', 'kate', 'kathleen_collins wealth_management', 'kathleen_collins', 'larry', 'larry_elder show', 'larry_elder', 'lawrence', 'm david_greene', 'm joshua_johnson', 'm noel_king', 'm rachel', 'm steve_inskeep', 'm sam_sanders', 'made_possible', 'marketplace morning', 'marketplace', 'martin and', 'martin', 'melissa', 'member station', 'morning_edition from', 'morning_edition on', 'morning_edition', 'n _b' 'n _p', 'n _p', 'nbc_news radio', 'nbc_news', 'new_england public_radio', 'news dot', 'news network', 'news supported', 'news talk', 'news_radio eleven', 'news_radio eleventh', 'news_radio nine', 'news_radio seven', 'news_radio ten', 'news_radio', 'next fresh_air', 'nine three', 'ninety am', 'ninety dot', 'ninety point', 'ninety_one point', 'ninety_one', 'noel_king', 'nouri', 'npr and', 'npr comes', 'npr i', 'npr news', 'npr s', 'npr station', 'npr stations', 'npr', 'o m', 'o o', 'o r', 'o', 'of new_england', 'on fox_nation', 'on fox_news', 'on fox_sports', 'on ground_zero', 'on morning_edition', 'on news_radio', 'on npr', 'on point', 'on_point comes', 'on_point', 'one a', 'org and', 'org or', 'org slash', 'org', 'other contributors', 'p b', 'p r', 'public media', 'public_broadcasting', 'public_radio and', 'public_radio comes', 'public_radio is', 'public_radio was', 'public_radio', 'r dot', 'r g', 'r', 'rachel martin', 'rachel', 'radio lab', 'radio music_festival', 'radio_network', 'radiolab', 'rush_limbaugh', 's morning_edition', 's npr', 'sam_sanders', 'sarah', 'sean_hannity show', 'sean_hannity', 'six one', 'sixty eight', 'slash npr', 'sports_radio', 'stations other', 'stephen', 'steve', 'steve_inskeep and', 'steve_inskeep', 'support comes', 'support for', 'supported by', 'talk thirteen', 'talk_radio seventy', 'terry_gross', 'the bbc', 'the fox_news', 'the npr', 'the sean_hannity', 'the_take', 'things considered', 'thirteen eighty', 'this npr', 'this weekend_edition', 'three three', 'to all_things', 'to fox_news', 'to morning_edition', 'to npr', 'unk fox_news', 'unk npr', 'vicki', 'w _a', 'w _b', 'w _v', 'w a', 'w b', 'w i', 'w j', 'weekend_edition from', 'weekend_edition', 'westwood_one', 'with george', 'y _w', 'your support'.

The extended list includes all of the base list, as well as: 'alabama', 'alaska', 'arizona', 'arkansas', 'california', 'colorado', 'connecticut', 'delaware', 'district_of_columbia', 'florida', 'georgia', 'hawaii', 'idaho', 'illinois', 'indiana', 'iowa', 'kansas', 'kentucky', 'louisiana', 'maine', 'maryland', 'massachusetts', 'michigan', 'minnesota', 'mississippi', 'missouri', 'montana', 'nebraska', 'nevada', 'new_hampshire',

'new_jersey', 'new_mexico', 'new_york', 'north_carolina', 'north_dakota', 'ohio', 'oklahoma', 'oregon', 'pennsylvania', 'rhode_island', 'south_carolina', 'south_dakota', 'tennessee', 'texas', 'utah', 'vermont', 'virginia', 'washington', 'west_virginia', 'wisconsin', 'wyoming', 'san_francisco', 'bay_area', 'in san_francisco', 'the san_francisco', 'san_diego', 'san', 'the california', 'oakland', 'el_paso', 'cal', 'of san_francisco', 'new_england', 'springfield', 'worcester', 'sacramento', 'atlanta', 'southern_california', 'in el_paso', 'pittsburgh', 'omaha', 'nashville', 'florida s', 'southeast', 'the florida', 'tampa', 'jacksonville', 'news washington', 'in jacksonville', 'in omaha', 'of houston', 'houston', 'des_moines', 'the bay_area', 'san_francisco and', 'berkeley', 'in san_diego', 'in atlanta', 'the san_diego', 'mesa', 'in new_hampshire', 'san_diego and', 'so cal', 'san_diego county', 'south_bend', 'of california', 'diego s', 'diego', 'of san_diego', 'to boston', 'albany', 'san diego', 'providence', 'the bay', '_a county', 'oklahoma_city', 'santa_monica', 'birmingham', 'in nashville', 'in pittsburgh', 'in florida', 'the tennessee', 'in nebraska', 'in california', 'bay', 'from florida', 'st_louis', 'in st_louis', 'of florida', 'to florida', 'california and', 'california s', 'california is', 'la', 'on san', 'tennessee star', 'on ninety_one', 'point seven', 'am eleven', 'dot o', 'the_journal', 'the_wild', 'an h', 'unk unk', 'the_city', 'new_yorker', 'a l', '_l', 'k a', '_w', 'a _g', '_b', 'o w', 'w e', 'l _a', 'n _y', 'w _i', '_a s', 'v i', 'c t', 'vi', 'w l', 'wor', 'u _c', 'm i', 'nine w', 'ypg', 'katie', 'cpr news', 'k r', 'point five', 'i a', 'russell', '_u', '_r', 'glenn_beck', 'ten k', '_w _w', 'p a', 'saint', '_f _l', 'first news', 'l l', 'k f', 'dave_anthony', 'w _g', 'ron', 'r c', 'i d', 'public radio', 'on cpr', 'ninety six', 'w p', 'on all_things', 'zero', 'cc', 'oh five', 'the fox', 'r a', 'lewis', 'from five', 'h _d', 'b o', 'franklin', 'w t', 'w _t', 'news_radio six', 'radio six', 'six seven', 'angela', 'thirteen ten', 'today_show', 'seven ten', 'nine eight', 'five eight', 'four eight', 'seven seven', 'point nine', 'radio five', 'eight eight', 'zero_zero', 'talk_radio ninety', 'radio ten', 'dave_anthony fox_news', 'rosenthal', '_l _w', 'k _l', 'l w', 'on k', 'w dot', 'the b', 'k o', 'from k', 'w o', 'h _b', '_b _o', 'w _m', '_r _c', '_t _o', 't r', 'u _a', 's c', '_b _r', 'w v', 'f l', 's t', '_s _u', 'w _w', '_p _d', 'p', 'v', 'x', 'c', 'f', 'y', '_g', 'five nine', 'to five', '_w dot', 'one seven', 'seven fm', 'eight to', 'eleven fifty', 'four four', 'badgers', 'county credit_union', 'seven or', 'on seven', 'seven nine', 'thirty a', 'seven fifty', 'eight four', 'five five', 'three seven', 'fifteen ten', 'eleven ten', 'five seven', 'five for', 'ten twenty', 'for nine', 'four oh', 'nine ten', 'five three', '_m on', 'from w', 'nine point', 'a dot', 'seven thirty', 'seven zero'.

Appendix C

Radio Show-Twitter Account Mapping

Here we present the full list of shows for which we collected Twitter accounts, together with the corresponding Twitter accounts themselves.

1A 1a, jejohnson322

All Things Considered jdeahl, andrea_c_hsu, Laur_npr, melissagray69, nataliewins1, ailsachang, WatsonCarline, csymrl, NPRKelly, NPRMichel, prairielaura, bridgetkelley, asilverman, arishapiro, npratc, nprAudie

Ben Maller benmaller

Beyond the Beltway DUMO

Bryan Suits darksecretplace

CATS Roundtable JCats2013, CatsRoundtable

Chris Baker Show CBakerShow

City Visions cityvisionsKALW

Coast to Coast AM with George Noory coasttocoastam, GeorgeNooryC2C, g_knapp, JChurchRadio

Consumer Team pwthomson

Dan Caplis Show DanCaplis

Dan O'Donnell DanODonnellShow

Dennis Prager Show pragershow, DennisPrager

Dr. Ronald Hoffman DrRonaldHoffman

Erick Erickson EWErickson

First Coast Connect With Melissa Ross MelissainJax

Fresh Air Nindoonjibaa, nprfreshair

Glenn Beck marissananos, DomTheodore, glennbeck

Ground Zero with Clyde Lewis ClydeLewis, RonParanoia

Here and Now MikeMoschetto, samraphelson, Cristinakim830, CikuTheuri, odowdpeter, Cementley, HutchinsMade, aashlock, ABaileyLocke, jeremyhobson, cassadyariel, Jillhereandnow, hereandnowrobin, TonyaMosley, toddmundt, KBMM, hereandnowchris, hereandnow, tinkuray, mainecook, soundkins, ebolinsky, serenaamcmahon, menegon

Hugh Hewitt Show hughhewitt, Radioblogger

It's Been a Minute samsanders, NPRItsBeenAMin

Jesse Kelly JesseKellyDC

Joe Walsh Show WalshFreedom

Jordan Levy Show JordanLevyShow

Larry Elder Show larryelder

Laura Ingraham Show IngrahamAngle, DaveKast, taywaltstweets, MikeMcDonald901, JessyCurry

Mark Belling MarkBellingShow

Markley & Vancamp Show Dave_vanC

Michael Medved Show TomlinMedia, MedvedSHOW, MarkDavis

Morning Edition Nancy_Pearl, rachelnpr, MorningEdition, mirandatk, ReenaAdvani, PhilHarrellNPR, NPRinskeep, nprkyoung, NoelKing, nprgreene, bgordemer

Next Level with Frankie Lane ItsFrankieLane

Norman Goldman normangoldman

On Point davidfolkenflik, OnPointRadio, MeghnaWBUR

Planet Money, How I Built This planetmoney, duffinkaren, RadioMalone, HowIBuiltThis, jacobgoldstein, guyraz, GonzalezSarahA

RadioLabInvisibilia aspiegelnpr, HannaRosin

Radio Open Source radioopensource

Red Eye Radio bkradio, garyredeye1, ericharley

Rush Limbaugh MarkSteynOnline, limbaugh, toddeherman, rushlimbaugh, KenMatthews, RogerHedgecock, BoSnerdley, yesnicksearcy

Talk With The Green Guy greenguymedia

Tech It Out marc_saltzman

The Al Sharpton Show TheRevAl

The Bernie and Sid Show bernieandsid

The Buck Sexton Show BuckSexton

The Dana Show INDIO_RADIO, DLoesch

The Dave Ramsey Show DaveRamsey, ablakethompson, jeremybreland, jensievertsen, eblackey, suz_simms, jameschilds, KellyDaniel1974, BrentSpicer

The Drive at 5 with Curtis Sliwa CurtisSliwa

The Jay Weber Show JayWeber3

The Jim Bohannon Show jimbotalks

The Joe Pags Show JoeTalkShow, PagsSam

The John Batchelor Show batchelorshow

The Justin Brady Show JustinBrady

The Mark Levin Show marklevinshow, RichSementa, LevinTV, RichValdes

The Mike Broomhead Show broomheadshow

The Mike Gallagher Show radiotalkermike

The Mike Siegel Show DrMikeSiegel

The Morgan Show with Morgan White, Jr. MorganWBZ

The Savage Nation - Michael Savage jjverdi, ASavageNation, bighairyclint

The Sean Hannity Show PrdcrJG, LyndaMick, benbrownmiller, f_treacher, seanhannity

The Stephanie Miller Show StephMillerShow, JimWardVoices

The Vicki McKenna Show VickiMcKenna

The Walton and Johnson Show WaltonNJohnson

The World alexandranewman, dyerworld, amulyats, globalcartoons, MarcoWerman, amandaemcgowan, AndreaCrossan, travelfarnow, pritheworld, JenniferGoren, joyhackel, apeavey, CCWoolf, jasonmargolis

This Morning with Gordon Deal jkushinka, ThisMorningShow, RadioGavin1, dduncRadio, GordonDeal

Watch Dog on Wall Street with Chris Markowski chrismarko

Weekend Edition NPRNedWharton, DGJourno, NPRWeekend, lourdesgnavarro, sososophia16, nprsaraholiver, yawnstewart, bhardymon, nprscottsimon, pbreslow

WOR Tonight with Joe and Lis JoeConchaTV, LisWiehl

Bibliography

[1] B. Rosenwald, *Talk Radio's America: How an Industry Took Over a Political Party That Took Over the United States*. Harvard University Press, 2019, ISBN: 9780674185012.

[2] P. Barberá, "Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data," *Political Analysis*, vol. 23, pp. 76–91, 1 Jan. 2015, ISSN: 1047-1987. DOI: 10.1093/pan/mpu011. [Online]. Available: https://doi.org/10.1093/pan/mpu011.

[3] G. King, B. Schneer, and A. White, "How the news media activate public expression and influence national agendas," *Science*, vol. 358, pp. 776–780, 2017. [Online]. Available: https://science.sciencemag.org/content/358/6364/776.

[4] N. Usher and Y. M. M. Ng, "Sharing knowledge and "microbubbles": Epistemic communities and insularity in US political journalism," *Social Media + Society*, vol. 6, 2 Apr. 2020, ISSN: 2056-3051. DOI: 10.1177/2056305120926639. [Online]. Available: http://journals.sagepub.com/doi/10.1177/2056305120926639.

[5] P. Barberá and G. Rivero, "Understanding the political representativeness of Twitter users," *Social Science Computer Review*, vol. 33, pp. 712–729, 6 Dec. 2015, ISSN: 0894-4393. DOI: 10.1177/0894439314558836. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0894439314558836.

[6] D. Beeferman, W. Brannon, and D. Roy, "RadioTalk: A Large-Scale Corpus of Talk Radio Transcripts," in *Interspeech 2019*, Graz, Austria: ISCA, Sep. 2019, pp. 564–568. DOI: 10.21437/Interspeech.2019-2714. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2714.html.

[7] The Nielsen Company, "The Nielsen Total Audience Report: Q1 2018," 2018. [Online]. Available: https://www.nielsen.com/us/en/insights/report/2018/q1-2018-total-audience-report/.

[8] J. R. Zaller, *The Nature and Origins of Mass Opinion*, ser. Cambridge Studies in Public Opinion and Political Psychology. Cambridge University Press, 1992, ISBN: 9780511818691. DOI: 10.1017/CBO9780511818691. [Online]. Available: http://dx.doi.org/10.1017/CBO9780511818691.

[9] A. Swasy, *How Journalists Use Twitter: The Changing Landscape of U.S. Newsrooms*, ser. G - Reference, Information and Interdisciplinary Subjects Series. Lexington Books, 2016, ISBN: 9781498532181.

[10] Muck Rack, *State of journalism 2019: How journalists find their news, use social media, and work with PR teams*, 2019. [Online]. Available: https://info.muckrack.com/stateofjournalism.

[11] P. Vijayaraghavan, S. Vosoughi, and D. Roy, "Twitter demographic classification using deep multi-modal multi-task learning," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada: Association for Computational Linguistics, Jul. 2017, pp. 478–483. DOI: 10.18653/v1/P17-2076. [Online]. Available: https://www.aclweb.org/anthology/P17-2076.

[12] Pew Research Center, "National politics on Twitter: Small share of U.S. adults produce majority of tweets," Oct. 2019. [Online]. Available: https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2019/10/PDL_10.23.19_politics.twitter_FULLREPORT.pdf.

[13] Y. Benkler, R. Faris, and H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*. Oxford University Press, 2018, ISBN: 9780190923624.

[14] T. Bonini and T. Sellas, "Twitter as a public service medium? A content analysis of the Twitter use made by radio RAI and RNE," *Communicacion y Sociedad*, vol. 27, pp. 125–146, 2 2014. [Online]. Available: https://search.proquest.com/docview/1537380841?accountid=12492.

[15] S. Herrera-Damas and A. Hermida, "Tweeting but not talking: The missing element in talk radio's institutional use of Twitter," *Journal of Broadcasting & Electronic Media*, vol. 58, pp. 481–500, 4 Oct. 2014, ISSN: 0883-8151. DOI: 10.1080/08838151.2014.966361. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2014.966361.

[16] D. A. Ferguson and C. F. Greer, "Local radio and microblogging: How radio stations in the U.S. are using Twitter," *Journal of Radio & Audio Media*, vol. 18, pp. 33–46, 1 Apr. 2011, ISSN: 1937-6529. DOI: 10.1080/19376529.2011.558867. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/19376529.2011.558867.

[17] S. H. Chiumbu and D. Ligaga, ""Communities of strangerhoods?": Internet, mobile phones and the changing nature of radio cultures in South Africa," *Telematics and Informatics*, vol. 30, pp. 242–251, 3 Aug. 2013, ISSN: 07365853. DOI: 10.1016/j.tele.2012.02.004. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0736585312000214.

[18] W. W. Xu and M. Feng, "Talking to the broadcasters on Twitter: Networked gatekeeping in Twitter conversations with journalists," *Journal of Broadcasting & Electronic Media*, vol. 58, pp. 420–437, 3 Jul. 2014, ISSN: 0883-8151. DOI: 10.1080/08838151.2014.935853. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2014.935853.

[19] S. J. Moon and P. Hadley, "Routinizing a new technology in the newsroom: Twitter as a news source in mainstream media," *Journal of Broadcasting & Electronic Media*, vol. 58, pp. 289–305, 2 Apr. 2014, ISSN: 0883-8151. DOI: 10.1080/08838151.2014.906435. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2014.906435.

[20] B. Schultz and M. L. Sheffer, "An exploratory study of how Twitter is affecting sports journalism," *International Journal of Sport Communication*, vol. 3, pp. 226–239, 2 Jun. 2010, ISSN: 1936-3915. DOI: 10.1123/ijsc.3.2.226. [Online]. Available: https://doi.org/10.1123/ijsc.3.2.226.

[21] A. O. Larsson, "Tweeting the viewer – use of Twitter in a talk show context," *Journal of Broadcasting & Electronic Media*, vol. 57, pp. 135–152, 2 Apr. 2013, ISSN: 0883-8151. DOI: 10.1080/08838151.2013.787081. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838151.2013.787081.

[22] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK: Association for Computational Linguistics, Jul. 2011, pp. 1589–1599. [Online]. Available: https://www.aclweb.org/anthology/D11-1147.

[23] Y. Wang, M. McKee, A. Torbica, and D. Stuckler, "Systematic literature review on the spread of health-related misinformation on social media," *Social Science & Medicine*, vol. 240, p. 112 552, 2019, ISSN: 0277-9536. DOI: https://doi.org/10.1016/j.socscimed.2019.112552. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0277953619305465.

[24] D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, and M. Dredze, "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate," *American Journal of Public Health*, vol. 108, no. 10, pp. 1378–1384, 2018, PMID: 30138075. DOI: 10.2105/AJPH.2018.304567. [Online]. Available: https://doi.org/10.2105/AJPH.2018.304567.

[25] L. Bode and E. K. Vraga, "See something, say something: Correction of global health misinformation on social media," *Health Communication*, vol. 33, no. 9, pp. 1131–1140, 2018, PMID: 28622038. DOI: 10.1080/10410236.2017.1331312. [Online]. Available: https://doi.org/10.1080/10410236.2017.1331312.

[26] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 554–559, 2016, ISSN: 0027-8424. DOI: 10.1073/pnas.1517441113. [Online]. Available: https://www.pnas.org/content/113/3/554.

[27] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018, ISSN: 0036-8075. DOI: 10.1126/science.aao2998. [Online]. Available: https://science.sciencemag.org/content/359/6380/1094.

[28] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, May 2017. DOI: 10.1257/jep.31.2.211. [Online]. Available: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[29] H. Allcott, M. Gentzkow, and C. Yu, "Trends in the diffusion of misinformation on social media," *Research & Politics*, vol. 6, no. 2, p. 2 053 168 019 848 554, 2019. DOI: 10.1177/2053168019848554. [Online]. Available: https://doi.org/10.1177/2053168019848554.

[30] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018, ISSN: 0036-8075. DOI: 10.1126/science.aap9559. [Online]. Available: https://science.sciencemag.org/content/359/6380/1146.

[31] W. Lippmann, *Public Opinion*. Harcourt, Brace, 1922. [Online]. Available: https://books.google.com/books?id=eLobn4WwbLUC.

[32] R. Y. Shapiro, "Public opinion and American democracy," *Public Opinion Quarterly*, vol. 75, no. 5, pp. 982–1017, 2011. [Online]. Available: https://www.jstor.org/stable/41345919.

[33] P. E. Converse, "The nature of belief systems in mass publics (1964)," *Critical Review*, vol. 18, no. 1-3, pp. 1–74, 2006. DOI: 10.1080/08913810608443650. [Online]. Available: http://dx.doi.org/10.1080/08913810608443650.

[34] S. Iyengar and D. Kinder, *News That Matters: Television and American Opinion*. University of Chicago Press, 1987, ISBN: 9780226388588.

[35] J. Zaller and S. Feldman, "A simple theory of the survey response: Answering questions versus revealing preferences," *American Journal of Political Science*, vol. 36, no. 3, pp. 579–616, 1992, ISSN: 00925853, 15405907. [Online]. Available: http://www.jstor.org/stable/2111583.

[36] G. J. Martin and A. Yurukoglu, "Bias in cable news: Persuasion and polarization," *American Economic Review*, vol. 107, pp. 2565–2599, 9 Sep. 2017, ISSN: 0002-8282. DOI: 10.1257/aer.20160812. [Online]. Available: http://dx.doi.org/10.1257/aer.20160812.

[37] K. H. Jamieson and D. Albarracin, "The relation between media consumption and misinformation at the outset of the SARS-CoV-2 pandemic in the US," *Harvard Kennedy School Misinformation Review*, Apr. 2020. DOI: 10.37016/mr-2020-012. [Online]. Available: http://dx.doi.org/10.37016/mr-2020-012.

[38] A. Simonov, S. Sacher, J.-P. Dubé, and S. Biswas, "The persuasive effect of Fox News: Non-compliance with social distancing during the COVID-19 pandemic," National Bureau of Economic Research, May 2020. DOI: 10.3386/w27237. [Online]. Available: http://www.nber.org/papers/w27237.pdf.

[39] J. M. Berry and S. Sobieraj, "Understanding the rise of talk radio," *PS: Political Science & Politics*, vol. 44, pp. 762–767, 04 Oct. 2011, ISSN: 1049-0965. DOI: 10.1017/S1049096511001223. [Online]. Available: http://www.jstor.org/stable/41319965.

[40] B. E. Drushel, "The Telecommunications Act of 1996 and radio market structure," *Journal of Media Economics*, vol. 11, pp. 3–20, 3 Jul. 1998, ISSN: 0899-7764. DOI: 10.1207/s15327736me1103_2. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1207/s15327736me1103_2.

[41] D. Hendy, *Radio in the Global Age*. Wiley, 2013, ISBN: 9780745667171. [Online]. Available: https://books.google.com/books?hl=en&lr=&id=TMUVkbPNvL8C.

[42] J. M. Berry and S. Sobieraj, *The Outrage Industry: Political Opinion Media and the New Incivility*. Oxford University Press, 2016, ISBN: 978-0190498467.

[43] S. Sobieraj and J. M. Berry, "From incivility to outrage: Political discourse in blogs, talk radio, and cable news," *Political Communication*, vol. 28, pp. 19–41, 1 Feb. 2011, ISSN: 1058-4609. DOI: 10.1080/10584609.2010.542360. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/10584609.2010.542360.

[44] K. H. Jamieson and J. N. Cappella, *The Echo Chamber: Rush Limbaugh and the Conservative Media*. Oxford University Press, 2008, p. 301, ISBN: 0195366824.

[45] C. R. Hofstetter and C. L. Gianos, "Political talk radio: Actions speak louder than words," *Journal of Broadcasting & Electronic Media*, vol. 41, pp. 501–515, 4 Sep. 1997, ISSN: 0883-8151. DOI: 10.1080/08838159709364423. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/08838159709364423.

[46] C. R. Hofstetter, D. Barker, J. T. Smith, G. M. Zari, and T. A. Ingrassia, "Information, misinformation, and political talk radio," *Political Research Quarterly*, vol. 52, p. 353, 2 Jun. 1999, ISSN: 10659129. DOI: 10.2307/449222. [Online]. Available: https://www.jstor.org/stable/449222?origin=crossref.

[47] D. C. Barker, "Rush to action: Political talk radio and health care (un)reform," *Political Communication*, vol. 15, pp. 83–97, 1 Jan. 1998, ISSN: 1058-4609. DOI: 10.1080/105846098199145. [Online]. Available: http://www.tandfonline.com/doi/full/10.1080/105846098199145.

[48] ——, "Rushed decisions: Political talk radio and vote choice, 1994-1996," *The Journal of Politics*, vol. 61, pp. 527–539, 2 May 1999, ISSN: 0022-3816. DOI: 10.2307/2647515. [Online]. Available: https://www.journals.uchicago.edu/doi/10.2307/2647515.

[49] D. Barker and K. Knight, "Political talk radio and public opinion," *The Public Opinion Quarterly*, vol. 64, pp. 149–170, 2 2000. DOI: 10.2307/3078813. [Online]. Available: http://www.jstor.org/stable/3078813.

[50] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," ACM Press, 2010, p. 591, ISBN: 9781605587998. DOI: 10.1145/1772690.1772751. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1772690.1772751.

[51] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? The structure of the Twitter follow graph," Association for Computing Machinery, Inc, Apr. 2014, pp. 493–498, ISBN: 9781450327459. DOI: 10.1145/2567948.2576939. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2567948.2576939.

[52] S. Yardi and D. Boyd, "Dynamic debates: An analysis of group polarization over time on Twitter," *Bulletin of Science, Technology & Society*, vol. 30, pp. 316–327, 5 Oct. 2010, ISSN: 0270-4676. DOI: 10.1177/0270467610380011. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0270467610380011.

[53] M. Conover, J. Ratkiewicz, M. Francisco, G. B., A. Flammini, and F. Menczer, "Political polarization on Twitter," 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2847/3275.

[54] Twitter, Inc., *Q4 and Fiscal Year 2018 Letter to Shareholders*, Feb. 7, 2019. [Online]. Available: https://s22.q4cdn.com/826641620/files/doc_financials/2018/q4/Q4-2018-Shareholder-Letter.pdf.

[55] R. Meyer, "Why Twitter may be ruinous for the left," *The Atlantic*, Jan. 2020. [Online]. Available: https://www.theatlantic.com/technology/archive/2020/01/how-twitter-harms-left/605098/.

[56] E. Klein, "The problem with Twitter, as shown by the Sarah Jeong fracas," *Vox*, Aug. 2018. [Online]. Available: https://www.vox.com/technology/2018/8/8/17661368/sarah-jeong-twitter-new-york-times-andrew-sullivan.

[57] D. Linker, "Twitter is destroying America," *The Week*, Jun. 2017. [Online]. Available: https://theweek.com/articles/702389/twitter-destroying-america.

[58] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on Twitter," ACM Press, 2011, p. 705, ISBN: 9781450306324. DOI: 10.1145/1963405.1963504. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1963405.1963504.

[59] L. Willnat and D. H. Weaver, "Social media and U.S. journalists: Uses and perceived effects on perceived norms and values," *Digital Journalism*, vol. 6, pp. 889–909, 7 Aug. 2018, ISSN: 2167082X. DOI: 10.1080/21670811.2018.1495570. [Online]. Available: http://dx.doi.org/10.1080/21670811.2018.1495570.

[60] U.S. Census Bureau, *American Community Survey*, 2016. [Online]. Available: ftp://ftp2.census.gov/geo/tiger/TIGER_DP/2016ACS.

[61] Theodric Technologies LLC. (2019). Radio-Locator, [Online]. Available: https://radio-locator.com/.

[62] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey, and S. Khudanpur, "JHU ASpIRE system: Robust LVCSR with TDNNs, ivector adaptation and RNN-LMs," in *Proceedings of 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, pp. 539–546.

[63] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Catalog No.: CFP11SRW-USB, Hilton Waikoloa Village, Big Island, Hawaii, US: IEEE Signal Processing Society, Dec. 2011.

[64] Twitter Inc., *Decahose stream*. [Online]. Available: https://developer.twitter.com/en/docs/tweets/sample-realtime/overview/decahose.

[65] MIT Election Data and Science Lab, *U.S. President Precinct-Level Returns 2016*, version V11, 2018. DOI: 10.7910/DVN/LYWX3D. [Online]. Available: https://dx.doi.org/10.7910/DVN/LYWX3D.

[66] M. Bloch, L. Buchanan, J. Katz, and K. Quealy, "An extremely detailed map of the 2016 presidential election," *New York Times*, 2018, Based on data provided by Ryne Rohla <http://rynerohla.com/>. [Online]. Available: https://www.nytimes.com/interactive/2018/upshot/election-2016-voting-precinct-maps.html.

[67] L. Bursztyn, A. Rao, C. Roth, and D. Yanagizawa-Drott, "Misinformation During a Pandemic," Becker Friedman Institue for Economics, University of Chicago, Tech. Rep., 2020. DOI: 10.2139/ssrn.3580487. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3580487.

[68] R. Alghamdi and K. Alfalqi, "A survey of topic modeling in text mining," *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 1, 2015. DOI: 10.14569/IJACSA.2015.060121. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2015.060121.

[69] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, "Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15 169–15 211, Nov. 2018, ISSN: 1573-7721. DOI: 10.1007/s11042-018-6894-4. [Online]. Available: http://dx.doi.org/10.1007/s11042-018-6894-4.

[70] D. Beeferman, *Topic indexing: Tools for generating indexes of a corpus of documents or conversations*, https://github.com/social-machines/topic-indexing, 2019.

[71] X. Li, J. Chi, C. Li, J. Ouyang, and B. Fu, "Integrating topic modeling with word embeddings by mixtures of vMFs," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 151–160. [Online]. Available: https://www.aclweb.org/anthology/C16-1015.

[72] F. Esposito, A. Corazza, and F. Cutugno, "Topic modelling with word embeddings," in *CLiC-it/EVALITA*, 2016.

[73] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, Jul. 2020, ISSN: 2307-387X. DOI: 10.1162/tacl_a_00325. [Online]. Available: http://dx.doi.org/10.1162/tacl_a_00325.

[74] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NeurIPS*, 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

[75] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta: ELRA, May 2010, pp. 45–50. [Online]. Available: http://is.muni.cz/publication/884893/en.

[76] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, ICLR workshop paper, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781.

[77] T. Mikolov, S. W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*, Association for Computational Linguistics, May 2013. [Online]. Available: https://www.microsoft.com/en-us/research/publication/linguistic-regularities-in-continuous-space-word-representations/.

[78] Policy Agendas Project at the University of Texas at Austin, *Topics codebook*, 2019. [Online]. Available: https://www.comparativeagendas.net/us.

[79] R. A. Harder, J. Sevenans, and P. V. Aelst, "Intermedia agenda setting in the social media age: How traditional players dominate the news agenda in election times," *The International Journal of Press/Politics*, vol. 22, no. 3, pp. 275–293, 2017. DOI: 10.1177/1940161217704969. [Online]. Available: https://doi.org/10.1177/1940161217704969.

[80] H. Yan, A. Lavoie, and S. Das, "The perils of classifying political orientation from text," 2017, pp. 38–50. [Online]. Available: https://www.cse.wustl.edu/~sanmay/papers/political-orientation.pdf.

[81] J. Green, "Identifying and estimating the ideologies of Twitter pundits," Data for Progress, Nov. 2019. [Online]. Available: https://www.dataforprogress.org/blog/2018/11/19/identifying-and-estimating-the-ideologies-of-twitter-pundits.

[82] J. Timm, *Twitter, political ideology and the 115th US Senate*, Nov. 2018. [Online]. Available: https://www.jtimm.net/2018/11/03/twitter-political-ideology-and-the-115-us-senate/.

[83] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, pp. 515–554, 3 May 2012, ISSN: 1384-5810. DOI: 10.1007/s10618-011-0224-z. [Online]. Available: http://link.springer.com/10.1007/s10618-011-0224-z.

[84] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, P10008, Oct. 2008. DOI: 10.1088/1742-5468/2008/10/p10008.

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

[86] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer-Verlag Berlin Heidelberg, 2005, ISBN: 978-3-540-40172-8. DOI: 10.1007/978-3-540-27752-1. [Online]. Available: https://www.springer.com/gp/book/9783540401728.

[87] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with Python," in *9th Python in Science Conference*, 2010.

[88] B. A. Conway, K. Kenski, and D. Wang, "The rise of Twitter in the political campaign: Searching for intermedia agenda-setting effects in the presidential primary," *Journal of Computer-Mediated Communication*, vol. 20, no. 4, pp. 363–380, May 2015, ISSN: 1083-6101. DOI: 10.1111/jcc4.12124. [Online]. Available: https://doi.org/10.1111/jcc4.12124.

[89] H. S. Lee, "Analyzing the multidirectional relationships between the president, news media, and the public: Who affects whom?" *Political Communication*, vol. 31, no. 2, pp. 259–281, 2014. DOI: 10.1080/10584609.2013.815295. [Online]. Available: https://doi.org/10.1080/10584609.2013.815295.

[90] S. Stern, G. Livan, and R. Smith, "A network perspective on intermedia agenda-setting," *Applied Network Science*, vol. 5, no. 1, Jun. 2020. DOI: 10.1007/s41109-020-00272-4. [Online]. Available: https://appliednetsci.springeropen.com/articles/10.1007/s41109-020-00272-4.

[91] S. Martelle, "Opinion: Wait, Trump wants to 'LIBERATE' Michigan but not Georgia?" *Los Angeles Times*, Apr. 23, 2020. [Online]. Available: https://www.latimes.com/opinion/story/2020-04-23/trump-liberate-michigan-but-not-georgia (visited on 07/29/2020).

[92] U.S. Census Bureau. (2019). U.S. Census Bureau QuickFacts: United States, [Online]. Available: https://www.census.gov/quickfacts/fact/table/US/PST045219.

[93] M. McCombs, *Setting the Agenda: The Mass Media and Public Opinion*. Wiley, 2013, ISBN: 9780745637136. [Online]. Available: https://books.google.com/books?id=oN2PKXMJYjkC.

[94] T. P. Boyle, "Intermedia agenda setting in the 1996 presidential election," *Journalism & Mass Communication Quarterly*, vol. 78, no. 1, pp. 26–44, 2001. DOI: 10.1177/107769900107800103. [Online]. Available: https://doi.org/10.1177/107769900107800103.

[95] W. Wanta and J. Foote, "The president-news media relationship: A time series analysis of agenda-setting," *Journal of Broadcasting & Electronic Media*, vol. 38, no. 4, pp. 437–448, 1994. DOI: 10.1080/08838159409364277. [Online]. Available: https://doi.org/10.1080/08838159409364277.

[96] B. Horvit, A. J. Schiffer, and M. Wright, "The limits of presidential agenda setting: Predicting newspaper coverage of the weekly radio address," *The International Journal of Press/Politics*, vol. 13, no. 1, pp. 8–28, 2008. DOI: 10.1177/1940161207312674. [Online]. Available: https://doi.org/10.1177/1940161207312674.

[97] J. S. Peake and M. Eshbaugh-Soha, "The agenda-setting impact of major presidential tv addresses," *Political Communication*, vol. 25, no. 2, pp. 113–137, 2008. DOI: 10.1080/10584600701641490. [Online]. Available: https://doi.org/10.1080/10584600701641490.

[98] D. Kreiss, "Seizing the moment: The presidential campaigns' use of Twitter during the 2012 electoral cycle," *New Media & Society*, vol. 18, no. 8, pp. 1473–1490, 2016. DOI: 10.1177/1461444814562445. [Online]. Available: https://doi.org/10.1177/1461444814562445.

[99] A. Jungherr, "Twitter use in election campaigns: A systematic literature review," *Journal of Information Technology & Politics*, vol. 13, no. 1, pp. 72–91, 2016. DOI: 10.1080/19331681.2015.1132401. [Online]. Available: https://doi.org/10.1080/19331681.2015.1132401.

[100] Federal Communications Commission, "FCC Broadcast Ownership Rules," Jan. 17, 2020, A summary of the official regulation codified at 47 CFR §73.3555. [Online]. Available: https://www.fcc.gov/sites/default/files/fcc_broadcast_ownership_rules.pdf.

[101] G. Bouma, "Normalized (pointwise) mutual information in collocation extraction," *Proceedings of the Biennial GSCL Conference 2009*, Jan. 2009.