ABSTRACT

SHESHADRI, KARTHIK. Toward a Unified Model of News Influence on Public Interest and Legislation. (Under the direction of Dr. Munindar P. Singh).

News has been shown to influence public perception, affect technology development, and increase public expression. However, existing research has not uncovered a mathematically expressible system of laws that predictably govern and explain this relationship. This research attempts to establish a computational science that yields predictive utility between specific news characteristics and public reaction. We contribute a set of novel observations that represent substantial progress toward a predictive model of how influential news patterns arise and the likely quantum of public and legislative reaction to them. Accordingly, we make three major contributions:

Firstly, we demonstrate that framing, a subjective aspect of news, appears to influence both significant public perception changes, and federal legislation. We show that specific features of news, such as publishing volume, appear to influence sustained public attention, as measured by annual Google Trends data, and federal legislation. We observe that federal legislative activity is often foreshadowed by periods of high news volume and similarity between articles, which we call *hyper-concentrated news periods*. Finally, we contribute the measures of *framing density* and *framing polarity*, which provide a quantitative assessment of news framing in a domain. We demonstrate that these measures appear to correlate substantially with the results of earlier human surveys.

Having established that hyper-concentrated periods Granger-cause federal legislation, we address the likelihood of origination of such influential news and reaction patterns across the distribution of news topics. We observe a uniform and systematic *selection bias*, common to news outlets and the public, that operates with rapidly decaying effect over the range of news topics. We show that the set of news topics partitions into three *interest regions*, which govern the nature of news selection from a topic, and the likely public reaction to it. We identify the news topics that inhabit each region. We find that the distributions of news volume and public interest (measured based on Google Trends) over topics follow rapidly decaying distributions, resembling Zipfians. We further find that the orders of the topics in these two distributions are substantially similar. We introduce a predictive model of *news prominence* that identifies key influential factors. Notably among these factors, we find that prominence varies predictably

across our three interest regions. We further demonstrate that the following three aspects of news, namely, news volume, concentration, and prominence may exhibit predictable covariance, that is, are not mutually independent. We show that news prominence can Granger cause federal legislative activity.

Finally, we address the problem of framing change detection. Changes in the framing of topical news have been shown to have public, legislative, and commercial consequences. Automated detection of framing changes is therefore an important problem, which existing research has not considered. Previous approaches are manual surveys, which rely on human effort and are consequently limited in scope. We make the following contributions. We systematize discovery of framing changes through a fully unsupervised computational method that isolates framing change trends over several years. We demonstrate our approach by identifying previously known framing changes. We have prepared a new dataset, consisting of over 12,000 articles from seven news topics or *domains* in which earlier surveys have found framing changes. Our work highlights the predictive utility of framing change detection, by identifying two domains in which framing changes foreshadowed substantial legislative activity.

Toward a Unified Model of News Influence on Public Interest and Legislation

by
Karthik Sheshadri

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina
2019

APPROVED BY:

Dr. William Enck	Dr. Arnav Jhala
Dr. Joann Keyton	Dr. Munindar P. Singh Chair of Advisory Committee

DEDICATION

Ukkādhāro Manussānam Niccam Apacito Mayā
"Always do I revere the torchbearer of mankind"

— Sutta Nipata 3.36

I also dedicate this thesis to my parents and sister.

ACKNOWLEDGMENTS

This work has benefited from collaborations and guidance from many colleagues and advisors. Firstly, I thank my research advisor, Dr. Munindar P. Singh, for his guidance and support. I am grateful to my committee members Dr. William Enck, Dr. Arnav Jhala, and Dr. Joann Keyton for their valuable advice and for many useful discussions during the course of my program.

In respect of Chapter 2, I thank Chung-Wei Hang for valuable discussions about the contributions and for his advice in ensuring the reproducibility of the results. I further thank Pradeep Murukannaiah for useful discussions, and the anonymous reviewers for helpful comments on a previous version.

I thank my labmates and collaborators Chaitanya Shivade, Nirav Ajmeri, Chinmaya Dabral, and several others. I have learned much from each of them and sincerely appreciate their guidance and support.

I am immensely grateful to my parents without whose constant encouragement I would not have come this far. My sister has been a role model and an inspiration to me throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Thesis Statement: *If the computational tools used are sufficiently precise, automated analysis of news yields predictive utility for trends in public and legislative response.*

Shared communication is a central aspect of human society. A prominent medium by which humans communicate is through *news*. In a democratic society, news is therefore a potential mechanism to influence legislation, since legislation is enacted by popular representatives of the public.

In the present heavily inter-connected and data-driven world, scientists have the ability to computationally model news patterns and subsequent societal changes. Such computational modeling enhances reach and coverage over the work of human social scientists, for whom it is infeasible to manually study patterns over large scales.

Further, whereas Political and Communications scientists have long studied the effect of news on human behavior Gunther (1998); Soroka (2006); King et al. (2017), existing approaches are limited to human surveys that study the effect of individual characteristics of news (such as *framing* Entman (1993), *selection*, and *slant* Gunther (1998)) on a study population.

If one were to model news as a dynamic system, we find that each news characteristic is an independently variable factor affecting the range of states in which the overall system may exist. Further, human behavior in itself is an extraordinarily complex phenomenon. In this work, all uses of the term "human behavior" refer to aggregate behavior of humans as a collective.

We additionally motivate a data-driven study of the problem of news influence on human in-

terest and reaction as a possible approach to cope with its complexity.

We posit that the level of precision provided by our analysis is sufficiently precise as to enable predictive utility for trends in public and legislative response.

Accordingly, we address the following major challenges. Firstly, we aim to establish the nature of news influence on human reaction. We seek to identify patterns in news publishing that elicit such responses, and investigate the origin of such "influential" news patterns. Secondly, we aim to combine different news characteristics into a "unified" model of news influence that yields predictive utility for human response to the of the dynamic system that news represents. Finally, we attempt to create computational methods and tools that enable *sufficiently precise* measurement and analytics of the various news characteristics we consider.

1.1 Motivating Example

In 2011, security vulnerabilities in Facebook's use of HTML5 allowed third-party applications to steal personal data from approximately 59 million users Crossley (2011). The framing of news on the topic "Markup Languages" changed from a neutral narrative to one focusing on personal privacy. Notably, fourteen times as many news articles on the topic "markup languages" were published in 2011 as in 2010. Revenues of Merix Games, Wimi5, and other HTML product vendors declined to the tune of four million dollars (over all companies) over the course of 2012 Fitzsimmons (2014). We conjecture that the decline was caused by negative coverage of news about HTML5 following the data leak.

Further, in 2013, the Personal Data Protection and Breach Accountability Act was promulgated in the US Congress (2014), under which Facebook was sued Fitzsimmons (2014).

1.2 Research Questions

Motivated by the above example, and by our overall challenges, we consider three primary research questions:

RQ_1 Legislative Activity: Do news patterns such as those involving news volume and framing influence public and legislative activity?

RQ_2 News Distributions: Are certain news topics more likely to elicit public and legislative reaction than others?

RQ_3 Framing Changes: Can framing changes be detected computationally, and if so, how accurately?

1.3 Contributions

We introduce and summarize each of our three primary contributions below. We list our contributions in the following order. First, we attempt to establish that news patterns systematically foreshadow legislation. We then consider the problem of evaluating how likely influential news patterns are to arise in different domains. We finally our contribute our computational treatment of the problem of detecting framing changes, which we demonstrate to have commercial and legislative import.

1.3.1 Hyper-concentrated Periods of Topic News

Our investigation of the influence of news on legislative activity uncovered a frequently occurring pattern in topical news publishing, which we term a *hyper-concentrated* news period. Such periods are characterized by high news volume, occurring simultaneously with high median similarity between articles from the same topic. We measure volume and similarity at the resolution of a year. We find that hyper-concentrated periods consistently Granger-cause federal legislation, and changes in the level of interest the public displays for a topic.

1.3.2 Three Interest Regions of News Publishing

Our investigation of the relative likelihood of different news topics led to discovery of the following patterns. We find that over sufficiently large time periods, the distributions across topics of the number of news articles and Google Trends responses (defined as $\frac{|\delta x|}{x}$ of the raw Google Trend volume) to significant events are rapidly decaying, and resemble Zipfian distributions. Moreover, we find that the topic orders of these distributions are *substantially correlated*. This finding is surprising in that it is the *percentage increase* in Google Trends that correlates with news volume, not the magnitude of the Google Trends volume. Taken

together with the fact that the distribution of the number of events across all topics is a uniform distribution, these results present evidence of a *systematic selection bias* in news and public interest. Further, news topics partition into three *interest regions*, which govern the nature of news selection from a topic, and the likely public reaction. We find that the region memberships of each topic are remarkably stable over time periods on the order of a decade. We find that news volume, Google Trend responses, and news prominence vary predictably across our three interest regions. This finding establishes that public reaction to events follows a power law with respect to the set of possible topics.

1.3.3 An Unsupervised Natural Language Approach to Framing Change Detection

We systematize discovery of framing changes through a fully unsupervised computational method that isolates framing change trends over several years. We introduce a new dataset, consisting of over 12,000 articles from six news topics or *domains* in which earlier surveys have found framing changes. We demonstrate our approach by identifying previously known framing changes. Our work further demonstrates the predictive utility of framing change detection, by identifying two domains in which framing changes foreshadowed substantial legislative activity.

1.4 Related Work

We discuss related work pertaining to each of our three main contributions below.

1.4.1 RQ_1: Do news patterns such as those involving news volume and framing influence public and legislative activity?

Our approach is similar in spirit to King et al. (2017) in that both their work and this paper examine the effect of news on public attention. However, our work yields several novel results. We posit the hyper-concentrated news period, and show that hyper-concentrated news is a Granger causal precursor to legislation. Our analysis applies to a larger population than the

outlets used by King et al., since our data sources Wikipedia (2001, 2002) enjoy wide readership. We measure public perception annually rather than over a period of weeks, as King et al. do. We distinguish between fact-based reporting and framing, and demonstrate that framing in itself is a Granger causal predictor of public approval and legislation.

In addition, we note that King et al. (2017) artificially created localized short duration hyper-concentrated news periods in their work. The fact that these short duration hyper-concentrated news periods did not Granger-cause legislation motivates the question of how long a hyper-concentrated news period must last in order to have such an influence. In our data, we found Granger-causality occurring between hyper-concentrated news and federal legislation over periods lasting at least a year.

Our conception of a hyper-concentrated news period is consistent with the idea of punctuated equilibria of media attention introduced by Baumgartner et al. (2014), and the notion of an availability cascade posited by Kahneman (2011). Finally, we note that Jacoby (2004) observes correlation between news coverage and legislation in a particular domain (Bankruptcy). We present several novel results that build on this body of work. We show that periods of macro-mutation (Baumgartner et al.) between punctuated equilibria, and availability cascades, may be brought about by news framing, without prominent event-based drivers. Whereas earlier work, such as by Baumgartner et al. and Edwards and Wood (1999), discuss Granger causal effects between media coverage and Congress, we demonstrate that punctuated equilibria extend to sustained public attention and legislative reaction.

Existing literature investigates various aspects of framing, and the terms “frame” and “framing” are consequently used to refer to various levels of analysis. For instance, Benford and Snow (2000) identify three core framing tasks: diagnostic, prognostic, and motivational framing. Further, collective action frames have been defined corresponding to the generation of interpretive frames that differ from and challenge existing ones. Existing work further studies “injustice frames” (Benford and Snow) as a particular subset of collective action frames that call attention to the victims of a given perceived injustice, and amplify their perceived suffering. Frame amplification (beyond injustice frames) and extension in particular have also been studied (Benford and Snow). The term “framing” is sometimes used to refer to tactics (Benford and Snow) that invoke human mental processes that lead members of the public to selectively focus on certain problems rather than on others. We acknowledge that our measures of framing do not probe such fine-grained processes, and our use of the word “frame” does not refer to such analyses.

1.4.2 RQ_2: Are certain news topics more likely to elicit public and legislative reaction than others?

Our investigation of this research question encompasses various areas of Social and Political Science. We discuss related work pertaining to each relevant area below, and describe our findings in each.

Selection Biases

Various selection biases in news coverage have been individually studied. Previous work includes examining bias in the coverage of protests McCarthy et al. (1996); Oliver and Maney (2000) to identifying race-specific biases in reports of violent crime Lundman (2003). In addition, earlier work posits selective selection bias in the coverage of individual domains such as privacy Sheshadri et al. (2017), but neither offers any explanations for the hypothesized bias nor compares it to biases in other domains. The first attempt at a large-scale modeling of selection bias was made by Bourgeois et al. (2018), who use a Matrix Factorization scheme to produce a latent representation of media selection between events and sources. Whereas Bourgeois et al. uncover interesting geographical dependencies, they do not uncover a systematic bias in domain selection as we do. Additionally, our work models the distribution of both news and Google Trends query volume with respect to various news domains.

Prominence

Existing work studies aspects of the problem of predicting and understanding prominence in isolation but does not model prominence as a whole. Recent work has examined how international relations affect prominence in US news coverage. Further, recent research Lee (2007) examines what makes specific countries feature prominently in global news coverage. Culbertson and Stempel (1984) hint at an analysis similar to ours in that they mention that prominence and article volume are correlated, but they restrict their analysis to mentioning this correlation, without quantifying or modeling it.

Macromutation and Hyper-Concentration

We already discussed the concept of macromutation and its analogy in the study of public reaction to news, namely, the hyper-concentrated period. We now discuss the additional impact of this concept for our specific research question.

Sheshadri and Singh (2019) introduce the hyper-concentrated period of domain news, which they characterize as a sustained burst of high domain news volume and concentration. They show that hyper-concentrated periods Granger cause changes in public opinion, as well as federal legislation. We build on this formulation by demonstrating that in general, domain prominence is an additional independent variable that characterizes such periods.

1.4.3 RQ_3: Can framing changes be detected computationally, and if so, how accurately?

The Media Frames Corpus, compiled by Card et al. (2015), studies three topics (Immigration, Smoking, and same-sex marriages), and identifies fifteen *framing dimensions* in each. We identify two major limitations of their work. Firstly, Card et al. study framing as a static detection problem, identifying which dimensions appear in a given news article. However, research in sociology Benford and Snow (2000) shows that most news topics feature a *dominant frame* (or dominant dimension in the terminology of Card et al. (2015)). Further, for a generic news topic, the dominant frame is not necessarily one of fifteen previously chosen dimensions, but can instead be an unknown arbitrary frame specific to the topic under consideration. For example, in the case of the Introduction and Contributions section, the dominant frame related to the privacy of individuals, which is not one of the fifteen dimensions described in Card et al. (2015).

Secondly, Sheshadri and Singh (2019) showed that public and legislative reaction tend to occur only after *changes* in the dominant frame. This finding motivates an approach to framing that focuses on identifying and detecting changes in the dominant frame of a news domain.

Sheshadri and Singh further propose two simple metrics that they motivate as measures of domain framing: framing polarity and density. They define framing polarity as the average frequency of occurrence in a domain corpus of terms from a benchmark sentiment lexicon. Framing density is measured using an entropic approach that counts the number of terms per

article required to distinguish a current corpus from an earlier one.

We identify the following limitations of the aforementioned measures (introduced in Sheshadri and Singh (2019)). Firstly, both measures make no effort to associate a given news article with a particular frame. Prior work does not support the inherent assumption that all articles in a given domain belong to a particular frame Benford and Snow (2000); Card et al. (2015). We enhance understanding by analyzing each domain using several distinct frames.

Secondly, framing density does not distinguish between a subjective choice made by a news outlet to frame a domain differently, and events that necessitate media coverage. Our work provides this distinction by analyzing framing using patterns of change in the adjectives that describe cooccurring nouns. Since adjectives are artifacts of subjective framing, they are not affected by events, as framing density is.

1.5 Preliminaries

We now provide a brief background on certain necessary concepts used in this work. We create domain datasets for each of our three primary research questions.

News Domain A broad collection of related news topics such as politics or sports is a *domain*.

Annual Domain Volume The number of news articles in a specific domain published in a given calendar year is its annual domain volume. As as example, in the *child privacy* domain of Fig. 2.2, the annual domain volume of the year 1990 is approximately 100, and the annual domain volume of year 2011 is approximately 300.

Domain Time Series A vector of annual news characteristics (such as domain volume defined above) is referred to as a domain time series. As an example, consider Fig. 2.2. The dotted (blue) curve depicting annual news volume over our period of interest is a domain time series.

Period of Interest The contiguous set of calendar years over which we analyze news publishing in a domain is its domain period of interest.

For the first research question on legislation, we would ideally like our period of interest to be ten years before a law was enacted. For our second research question on news distributions, we

examine the period 2004 to 2019. For the third research question on the detection of framing changes, we ideally consider ten years before and after a candidate framing change.

For many domains, our APIs do not provide data for the ideal period of interest as defined above. In these cases, we describe the specific period of interest used in the relevant chapters below.

As an example, our period of interest for the *Surveillance* domain is 2003 to 2016, twelve years before the enactment of the USA Freedom Act. However, our period of interest for the *Housing* domain is restricted to 2000 to 2003.

Granger Causality We use Granger causality Granger (1969) as our primary means of demonstrating correlations between domain time series that we posit may influence each other.

Granger causality uses predictive ability to conceptualize a statistical concept of causality. Granger causality posits that if a signal $x(t)$ "Granger-causes" (or "Granger-causes") a signal $y(t)$, then past values of $x(t)$ should contain information that helps predict $y(t)$ more accurately than the information contained in past values of $y(t)$ alone. Granger causality also consequently infers directionality, i.e., which of each of $x(t)$ or $y(t)$ (or both) Granger cause the other.

Granger causality is normally tested in the context of linear regression models. For illustration, consider a bivariate linear autoregressive model of two variables $x(t)$ and $y(t)$:

$$y(t) = \sum_{i=1}^{p} \alpha_i y(t-i) + c_1 + v_1(t) \tag{1.1}$$

$$y(t) = \sum_{i=1}^{p} \alpha_i y(t-i) + \sum_{j=1}^{p} \beta_j x(t-i) + c_2 + v_2(t) \tag{1.2}$$

where p is the maximum number of lagged observations included in the model (the model order), α_i and β_j are coefficients of the model (i.e., the contributions of each lagged observation to the predicted values of $x(t)$ and $y(t)$), and $v_1(t)$ and $v_2(t)$ are residuals (prediction errors) for each time series. If the variance of $v_1(t)$ (or $v_2(t)$) is reduced by the inclusion of the $x(t)$ terms in the first equation, then it is said that $x(t)$ Granger-causes $y(t)$.

The two models are denoted as restricted (M_R) and unrestricted (M_{UR}) respectively, accordingly as whether $x(t)$ is used as a predictor of $y(t)$ or otherwise. M_R is used as the critical

value, whereas the F measure is obtained as $F = \frac{M_R - M_{\text{UR}}}{p} \div \frac{M_{\text{UR}}}{n-k}$.

Our Granger causality searches are conducted on pairs of (hypothesized) Granger causal domain time series and (hypothesized) Granger caused ones.

A small fraction of our Granger causality searches suffer from rank deficiency due to the fact that our time series are relatively small. However, our searches appear to produce plausible results even in such cases.

Since Granger causality determines directionality in correlation between two time series, but does not measure "causality" in the strict sense of the term, we do not make causal claims based on our results.

Therefore, whereas we observe Granger causally significant correlations in our data, we are unable to state categorically that a (hypothesized) Granger causal time series influenced a (hypothesized) Granger caused one. We therefore qualify our claims by stating that Granger causal time series "appear to influence" Granger caused ones.

However, we posit that the correlations we estimate are strong enough to warrant future causal analyses.

1.6 Organisation

Chapter 2 introduces the hyper-concentrated period of topic news, and discusses correlations between such periods and federal legislation. We find that federal and state legislation is often foreshadowed by hyper-concentrated periods. We also show that news characteristics such as concentration, measured either using paragraph vector similarity, or using entropic keywords, can have a Granger causal relationship with legislation. Chapter 3 estimates relationships between the distributions of news and human interest over topics, and further ties in these findings to a computational analysis of news prominence. It describes our discovery of a substantial correlation between these distributions, which reveals a selection bias that is uniform to news portals and the public. This bias appears to operate with a rapidly decaying magnitude over the range of domains that we examine. This chapter also posits an extension of the Granger causal frame work we introduce in Chapter 2. Chapter 4 presents an unsupervised natural language approach to detecting framing changes in topical news publishing, and compares results from the approach to earlier human surveys. We find substantial correlations between the results of

our automated framework and the findings of these surveys. Further, our approach identifies periods with substantial legislative import, even in cases where earlier surveys were not conducted. Chapter 5 discusses the ramifications of our work and concludes with important future directions.

Chapter 2

The Public and Legislative Impact of Hyper-Concentrated Topic News

This chapter addresses our first research question, RQ_1. The chapter describes our discovery of the hyper-concentrated news period, and our experiments on Granger-causality that associate it with federal legislative activity. The chapter is based on a paper of the same title soon to appear in *Science Advances*.

2.1 Introduction and Contributions

The effect of news on public behavior has been the subject of considerable scientific interest. Prior work has established that news framing influences public perception Gunther (1998); Mutz and Soss (1997), affects technology development Hoadley et al. (2010); Taylor (2016), and contributes to setting agendas Iyengar and Kinder (2010). Most recently, publishing from small news outlets has been shown to increase short-term public involvement in specific domains King et al. (2017).

Our work enhances understanding by explicitly modeling the *Granger causal (G-causal)* Granger (1969) link between specific news characteristics, public opinion, and federal legislation. We note that Granger causality captures directionality in correlation between time series but does not correspond to "true" causality. In this work, we restrict ourselves to Granger causality and indicate every use of the term with the qualifier "Granger" or "G." Firstly, we demonstrate a

predictive relationship between news characteristics and federal legislation.

Secondly, we show that public and legislative reaction to news follows a punctuated equilibrium model Baumgartner et al. (2014). The punctuated equilibrium model, adopted from evolutionary biology, posits long periods of equilibrium during which there is little change, punctuated with short durations of macromutation. Similarly, we observe that the public and the federal legislature tend to react substantially at discrete intervals (analogously to macromutation in the above model), rather than uniformly and gradually. We identify a defining characteristic of news periods that appear to elicit such substantial reactions, namely, that they have high news-volume occurring simultaneously with high similarity between articles. We term such periods *hyper-concentrated news periods*. We note that the approach introduced in King et al. (2017) artificially created such news conditions for short time periods and reader subsets.

Thirdly, news reporting in general introduces subjective biases, referred to as framing. We adopt the formulation presented in Entman (1993) in this paper. Whereas news publishing is ordinarily event driven, we demonstrate that hyper-concentrated news periods, combining high article volume and similarity, can occur spontaneously, without event-based drivers, as a Granger causal effect of news framing (see Fig. 2.4 for a compelling example). We find that hyper-concentrated periods brought about by framing can be equally influential in predicting public approval (defined as the fraction of the public that approves of a particular position) and legislation. This finding demonstrates that the framing of news is as influential as the events and facts reported on in the news.

Additionally, we demonstrate that news publishing volume within a specific domain can be a reliable long-term predictor of public attention (the number of people who demonstrated interest in a domain by conducting an Internet search), measured annually using Google Trends data (Fig. 2.3).

The Granger causal flow we deduced is depicted in Fig. 2.1. We confirmed each link using a directional Granger causality test, which evaluates the influence of a G-causal time series on a G-caused one. All tests were conducted at the $\alpha = 0.05$ (or lower) significance level. Our choice of Granger causality over a structural model was deliberate, since we wished to infer rather than assume structure and direction. We note that prior research Edwards and Wood (1999) agrees with this choice.

The details of the parameters we use are listed in Table 2.2. To the best of our knowledge, we use the most stringent possible parameters to evaluate our hypothesis. We note that if our pa-

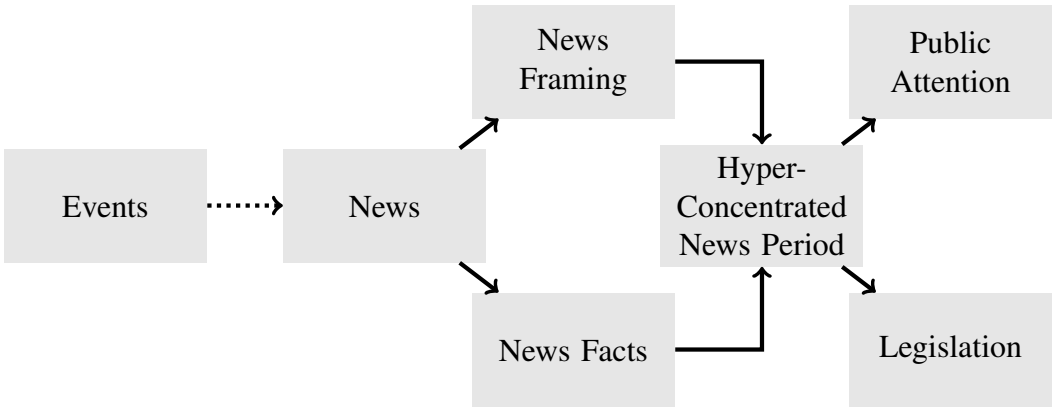


Fig. 2.1: We posit the *hyper-concentrated* period of domain news, characterized by high article volume and similarity, which G-causes public attention changes and legislation. Hyper-concentrated periods arise either due to news events, or independently of events, due to *news framing*. We observe and model every link in the figure, except the Events to News link, which is shown with a dotted arrow.

rameters were relaxed to admit higher lag values and minimum counts, significant results may be obtained for domains beyond those listed in the paper. Despite these conservative choices, we demonstrate that our hypothesis appears to hold consistently over our set of domains.

2.1.1 Hyper-Concentrated News Periods

Our observations stem from a remarkable pattern that holds reliably over the set of domains and articles we examined (we highlight several compelling cases in the text and in Section 2.4 on Detailed Results, and present a full list in Table 2.1).

We posit the idea of a *hyper-concentrated period* of domain news as one that is characterized by high article volume occurring simultaneously with high median similarity between articles. We study legislative reaction to news and find that Granger causally significant changes in legislative activity are often foreshadowed by hyper-concentrated periods. Fig. 2.5 illustrates a hyper-concentrated news period for the Surveillance domain.

We define the median similarity of a domain corpus of size n as the median cosine similarity Tan et al. (2006) in paragraph vector Le and Mikolov (2014a) space between all $\binom{n}{2}$ pairs of articles in the corpus. In each domain, periods of high article volume also tend to have high median similarity between articles (in multiple domains, this correlation was Granger causally

significant). This finding is surprising, since one would expect a larger volume of articles to discuss a larger variety of subjects. Instead, we found that domain news publishing tends to be event driven, and influential events appear to increase not only the median similarity of the corpus but also its *volume*. For example, the number of Surveillance articles increased by 282% in 2013, with 65% of the total (31 of 48) being primarily about Snowden. Although it is well known that news is event driven, the discovery of a Granger causal relationship between article volume and median corpus similarity is a novel finding of our work.

2.2 Materials and Methods

This section describes novel algorithms and methods introduced in our work. The Detailed Results section (Section 2.4) provides additional details.

2.2.1 Dataset Collection

We describe our data sources, method for domain dataset generation, and dataset quality evaluation below.

Data Sources

We used publicly accessible Application Programming Interfaces (APIs), specifically those of The New York Times NYT (2016) and The Guardian (2016), to create our news datasets. In addition to the large volume of relevant news made available by these two publications, our choice is motivated by their well-documented influence on public attitudes and perception Althaus and Tewksbury (2002); Drezner and Farrell (2004); Golan (2006); Kiousis (2004); Mutz and Soss (1997). We note that The New York Times has previously been shown to influence legislation Rahbar (2016), making it an ideal choice for our study.

The New York Times API provides a lead paragraph and/or a summary snippet for each news article. The Guardian provides full article text.

We note that the results returned by our APIs for the same queries can change over time. We expect that analyses conducted with later retrievals from the APIs should preserve the trends

in our data.

Domain Dataset Generation

As in earlier work King et al. (2017); Sheshadri et al. (2017), we use a standard term search procedure to create our datasets. For each domain, our APIs were used to extract news data during the time period b (denoting the beginning) to e (denoting the end), of the domain period of interest.

Ideally, we would like to use a period of interest of at least ten years before each federal law was enacted. For some domains, we were unable to retrieve data from our APIs for this full period. The period of interest used varies from a minimum of three years (for the domain Abortion), and a maximum of twelve years (for the domain Surveillance) before federal legislative activity.

Our hypothesis suggests that during periods of macromutation, publishing in a given domain may focus on a particular frame, such as the Snowden revelations in the domain Surveillance. In such cases, we have some pre-existing knowledge of our frame of interest. We attempt to use this knowledge in our term search procedure by focusing our search accordingly. As an example, we use the search term "surveillance+privacy" in the domain Surveillance, knowing that it is the aspect of personal privacy that macromutated in this domain.

In general, two factors limit our ability to focus our term search as described above. Firstly, we do not have advance knowledge of the frames in each domain. Secondly, even in domains for which we do have such knowledge, a particularly focused term search often saturates our paragraph vector training procedure (see the Corpus and Document Similarity section (section 2.2.3)), making it impossible for us to compare similarity across years. In such cases, we resort to using a more generic term search (such as "drones"). We acknowledge that the resultant dataset from such a generic term search may have lower precision in that articles discussing aspects of drone use not related to our frame of interest may be included. As an example, articles that discuss military drone use rather than civilian use (which is the aspect of drones that was legislated upon in the United States in 2012) may be included in such a search. However, we observe that the similarity of such generic datasets may tend to increase sharply during years in which legislation is enacted (see Fig 2.2 for an example, in which similarity increased in 1998, coinciding with COPPA legislation as described in the Main Findings section (Section 2.3.1)). We posit that this increase may be due to the fact that articles from such

a generic search tend to focus around the relevant frame of interest preceding legislation in the domain. We posit that the inclusion of articles from frames other than our frame of interest may thus help our paragraph vector model produce a sharper contrast during years preceding relevant legislation.

We provide a list of the terms used in Table 2.2.

We further note that the validity of our hypothesis does not appear to depend on the use of either a focused or generic term search procedure.

Dataset Quality

A random sample of articles from each domain dataset was coded by two raters. An article is considered as belonging to a domain if and only if a component of the article discusses the domain under consideration.

As an example, consider the article "Vivien Leigh lights a cigarette. Sigmund Freud lights a cigar. That's what they should be doing, isn't it? Miss Leigh is a glamorous movie star of a bygone era, and everyone knows about Dr. Freud and cigars." from the domain Smoking. We code it as a negative because whereas the article mentions smoking, it primarily discusses movies, and does not discuss any aspect pertaining to the prevalence or control of Smoking, which is our frame of interest.

In some domains, (such as Child Privacy), we slightly relaxed this criterion to allow the inclusion of articles from related domains such as Child Abuse, which we posit were G-causally influential in predicting legislation in this domain.

We obtain median per-domain accuracies of 0.83 according to coder 1 and 0.80 according to coder 2. We measured inter-annotator agreement using Cohen's κ Viera and Garrett (2005). Our median agreement was $\kappa = 0.67$, considered "substantial agreement" by Landis and Koch (1977). We acknowledge that the estimated precision may vary according to the specific sample used, and further may vary by coder.

We did not directly measure recall. However, since news publications have a strong incentive to broadly cover events, and The New York Times and The Guardian have the largest and fifth largest circulations in America and the world, respectively Wikipedia (2001, 2002), we assume that sufficiently many relevant articles are included in our corpus.

2.2.2 Discriminative Keywords

We are interested in identifying and summarizing those aspects of a domain's current framing that distinguish it from the domain's framing at a previous time period. To this end, we adopt the idea of an entropic formulation of discriminative keywords, as proposed by Sheshadri et al. (2017).

Below, a corpus T is a set of news articles. Specifically, given two disjoint sets of news articles T_1 and T_2, we identify a set of k n-grams that yield the largest Cross Entropy Harris (2002) in the combined corpus $T = T_1 \cup T_2$. Let A be an article in corpus T. Let x_i represent any of the possible m n-grams in T. Let $S(x_i, T) = \{A \in T | x_i \in A\}$ be the set of articles in corpus T in which the n-gram x_i appears. We use a $|T| \times m$ term frequency (TF) matrix representing the corpus to calculate H, the information entropy of T. We use MATLAB's fitctree and predictor importance functions with a split criterion parameter of 'deviance' to estimate the utility of each n-gram.

$$IG(T, x_i) = H(T) - \frac{S(x_i, T)}{|T|} H(S(x_i)) \tag{2.1}$$

Following Entman's Entman (1993) formulation, this approach weights n-grams that are specific to a particular corpus more highly than n-grams that are common to both corpora. A quick intuition for the approach is obtained by considering that the unigram "Snowden" may have a high utility in distinguishing Surveillance articles published after January 1^{st} 2014 from those before then, but the unigram "surveillance" is common to articles from both periods and therefore may not.

Since keywords from a particular news corpus distinguish it from others, they may be said to represent the "concentration" of news in that corpus.

2.2.3 Corpus and Document Similarity

We estimate the similarity of a corpus of documents as the median of its pairwise document similarities, using all $\binom{n}{2}$ combinations from the corpus. In order to estimate similarity between two documents, we adopt doc2vec Le and Mikolov (2014a), a well-known tool that generates a vector representation (called a "paragraph vector") of a document. Specifically, we use a

standard doc2vec model Lau (2017), trained on each domain corpus, to compute a vector for each document in our corpus. We define the pairwise similarity of two documents as the cosine similarity of their respective document vectors Wikipedia.

Whereas we do not in general deny that high median similarities can occur in annual corpora with low news volume (see fig. 2.8), we found that legislative activity tends to correlate with periods in which news volume and median similarity are simultaneously high. We therefore employ a threshold whereby the similarity of an annual corpus is considered to be zero if it contains less than $c\%$ of the articles from the respective domain corpus. We use a threshold of $c = 5\%$ in this dissertation.

We note that since cosine similarity has a range of $[-1, 1]$, and our models are learnt on datasets that discuss a common topic, the variation in similarities we obtain is relatively small compared to a metric with a larger range.

Despite this conservative choice, we demonstrate consistent G-causality with legislation (Table 2.2). We note that stronger significance may be obtained if we were to use similarity measures with larger ranges. However, the results obtained with our conservative approach inspire confidence in the validity of our hypothesis.

2.2.4 Framing Density

We contribute the notion of framing density, measured by entropic news keywords. We use entropy between pairs of temporally disparate news corpora (as described earlier) to rank individual n-grams for their effectiveness in distinguishing the later corpus from the earlier one. Entropic keywords therefore represent the "concentration" of a news domain at a given time. We define the annual framing density of a given domain as the number of keywords per article required to attain $K\%$ of dataset entropy between the present annual corpus and the preceding one. We examined values of K from 50% to 99%, and found that the resulting trend appeared to be fairly consistent across this range, though the specific values varied. Our intention is to capture the bulk of the probability mass while ignoring the long tail. We use a value of $K = 50\%$ in Fig. 2.5. We posit, as in Fig. 2.5, that framing changes tend to be characterized by low values of framing density.

We scale our values of framing density by a constant factor to enable visibility in figures.

2.2.5 Framing Polarity

We are interested in measuring the net polarity of the adjectives and adverbs within a corpus. Since adjectives and adverbs cannot be used to state underlying facts or events, they represent artifacts of how an event is framed.

Ideally, we would like to use the average sentiment polarity of all the adjectives and adverbs within a corpus as its framing polarity. However, we note that 75.27% of words from Sentiwordnet Baccianella et al. (2010), a benchmark lexical resource for opinion mining, have both a positivity and negativity score of zero. Therefore, an approach based on averaging polarities would not yield meaningful results.

Instead, we use an exhaustive list of manually curated sentiment adjectives and adverbs Breen (2011). We restrict ourselves to negative sentiment words in this paper, since the framing changes we examine are known to be associated with negative sentiment news, and previous work has established that negative news is more influential than positive news Soroka (2006); Sheshadri et al. (2017).

We measure the frequency of occurrence of each of these words within the corpus of interest, and sum them. Finally, we divide this sum by the number of articles in the corpus to represent its framing polarity. We calculate annual framing polarity within each domain, by constructing annual corpora from the full domain corpus.

Our domains tend to belong to one of two categories: (i) domains in which there is no substantial publishing in the absence of a hyper-concentrated period (such as Surveillance), and domains in which there is always substantial publishing (such as LGBT). In the former case, when a domain's annual article volume is close to zero, it does not represent a reliable factor with which to scale polarity. In this case, we present the sum of the number of negative sentiment words within an annual corpus as its framing polarity.

2.2.6 Measuring Domain Framing

We use our measures of framing polarity and concentration to assess domain framing. We show that the results obtained using these measures tend to correlate substantially with the findings of earlier human surveys.

2.3 Experiments and Discussion

We summarize our findings in the subsection below. Next, we describe comparisons with political framing. Finally, we discuss validation of our hypothesis using the comparative method in succeeding subsections.

2.3.1 Main Findings

To establish the Granger causal effect of hyper-concentrated news on legislation, we considered all federal legislation enacted beginning from 1991 up to 2016. Our choice was motivated by the fact that we were unable to achieve credible coverage using our APIs for legislation that occurred before this period. We found eight cases (seven American and one British) of federal legislation in this period that were Granger-caused by hyper-concentrated news periods. We acknowledge, however, that there may be further examples beyond those identified by our search. Whereas we do not claim hyper-concentrated news periods to be a necessary condition for legislation, we conclude that the probability of legislation being Granger caused by a hyper-concentrated period is statistically significant.

We illustrate our approach and results in Fig. 2.2, using a compelling example from the domain of Child Privacy. We use the abbreviation "HC Period" to refer to hyper-concentrated news periods in this and other figures. The primary laws governing children's privacy protection in the United States are COPPA FTC (1998) and FERPA U.S. Department of Education (1974). COPPA was enacted in the US Congress in 1998, and took effect in April 2000. Since then, a series of amendments have been proposed and enacted. We retrieved a list of COPPA amendments and subsequent press statements from www.ftc.gov. Due to the unavailability of children's privacy news articles before 1974 (a keyword search via The New York Times API returns zero articles), we restrict our analysis to COPPA. The Granger causal variables of interest in Fig. 2.2 are annual news volume (dotted blue), and median pairwise article similarity (dashed red). We represent the volume of COPPA legislative activity by a time series depicted with solid brown in Fig. 2.2. We represent the primary year of COPPA legislation, 1998, using a value of ten. Other years are represented according to the number of relevant FTC press statements during the year. Our Granger causality tests are therefore conducted between pairs of independent and (hypothesized) dependent time series, such as between news volume (dotted blue) and COPPA legislation (solid brown) in Fig. 2.2. We observe that the number of

news articles published on the topic more than doubled between 1991 and 1998, coupled with a simultaneous increase in median article similarity. Coinciding with this hyper-concentrated period, COPPA legislation was promulgated through the period ending in 2000. Another hyper-concentrated period occurs before the revival of interest in COPPA, as seen in the large number of amendments in the period 2011 to 2013.

We tested the Granger causal flow depicted in Fig. 2.1 over the set of domains obtained as described in the previous paragraph (using news volume and similarity as our G-causal variables and legislation as our G-caused ones), yielding statistically significant results in each case (see Table 2.1 and Table 2.2 for a full list). Our results motivate the predictive utility of news as a Granger causal set of independent variables that influence legislation.

Google Trends Trasborg (2018) estimate public interest in a topic of interest by measuring related searches worldwide over chosen time periods. Since 89% of US Research (2018) and 82% of UK residents of National Statistics (2017) use the Internet and 74% of Internet users use Google as their primary search Mangles (2018), we posit that Google Trends are a representative measure of public attention. For one of our domains (see Fig. 2.3) we found significant G-causality between article volume and Google Trend volume. This correlation is also observable in other domains (such as Cyberbullying), but yields G-causal measures that are slightly below the $\alpha = 0.05$ threshold in these domains.

The LGBT rights domain, depicted in Fig. 2.4, illustrates the Granger causal influence of framing on public opinion. Note that the negativity of framing drops in the year 2004–2005, after which public approval begins to climb steadily. Further, we note that an earlier survey Engel (2013) found that in 2003, print and media coverage of LGBT rights underwent a *change in framing*, during which coverage began to focus on the issue of marriage equality. We conjecture that the focus on marriage equality may have resulted in less negative news articles, which coincides with our findings based on framing polarity. This motivates the possible utility of framing polarity as a mechanism to isolate *changes in news framing*. Fig. 2.4 demonstrates an inverse relationship between framing negativity and public approval. We note that following this trend, major LGBT legislation legalizing same-sex marriage in fifty states was promulgated in 2016.

This result is noteworthy in that it is the polarity of news framing in the area, rather than specific news events, that Granger causes public approval at the 0.05 significance level. This finding is reinforced by the fact that event-based drivers cannot influence framing polarity, since only adjectives and adverbs, taken here to be artifacts of how a domain is framed, contribute to

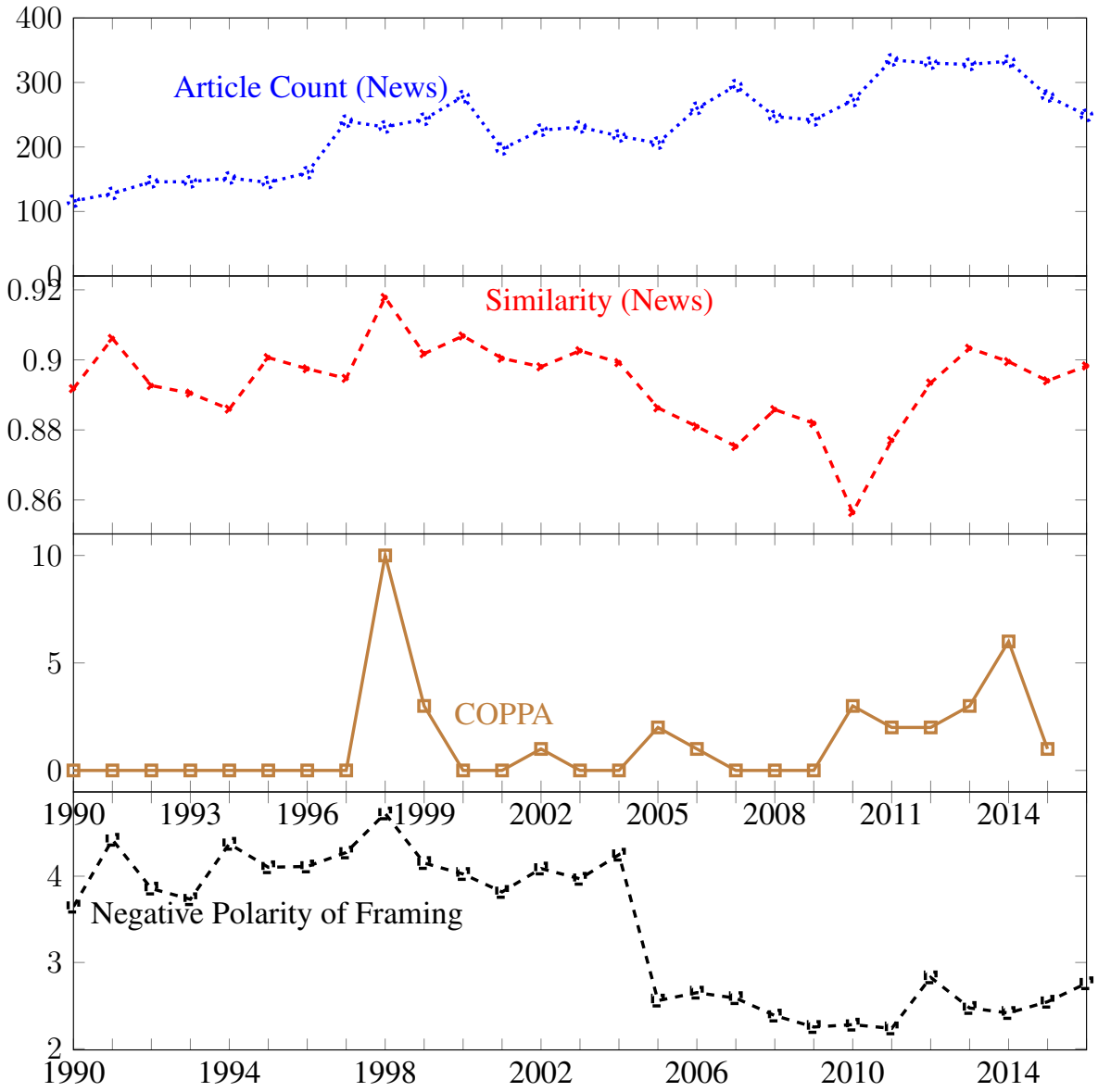


Fig. 2.2: News characteristics and legislation for the domain *Child Privacy*. Note that during the period 1996 to 1999, news volume and similarity sharply increase together, foreshadowing COPPA legislation. Notice further that news volume, similarity, and negative polarity of framing reach peaks in 1998, corresponding to the year that COPPA was promulgated.

framing polarity.

However, we note that we did also find G-causality between news volume in the LGBT rights domain and the number of state LGBT laws enacted per year, during the period 1996 to 2015 (see the Legislation section (Section 2.4.2)).

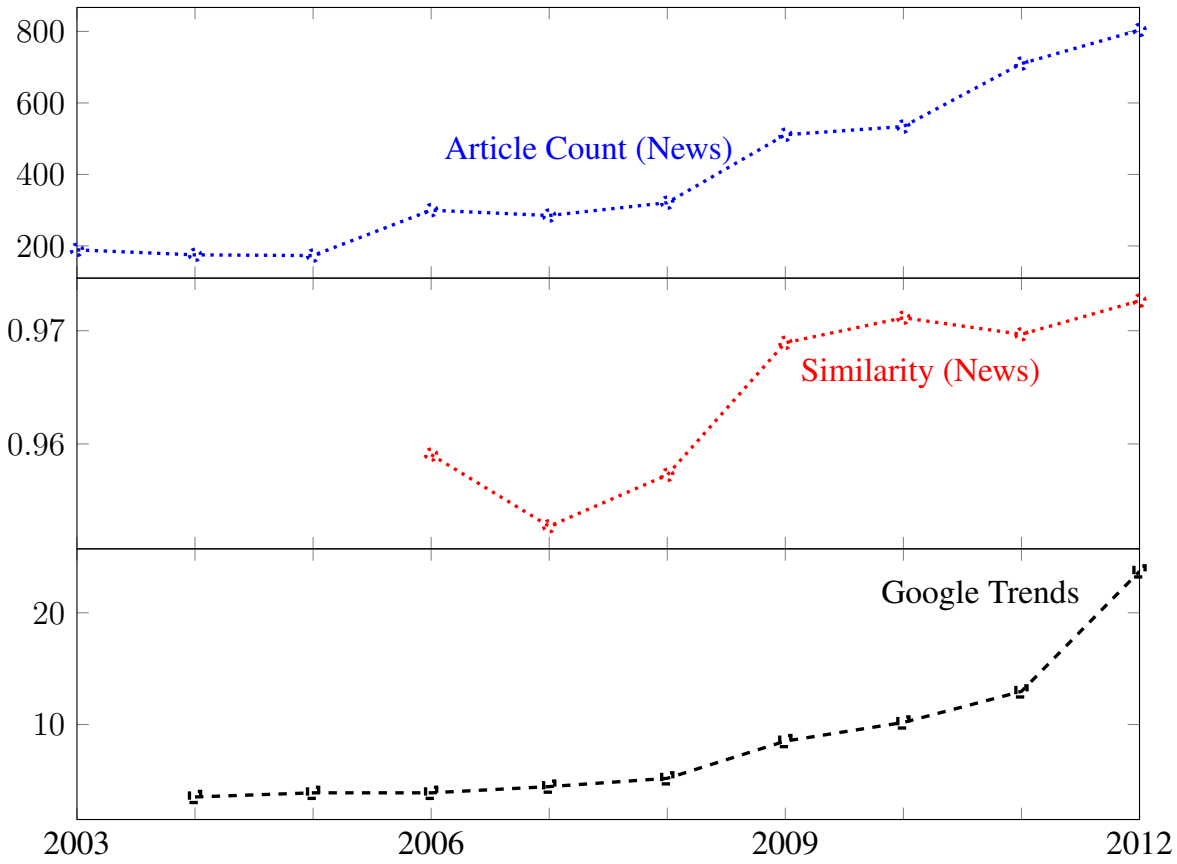


Fig. 2.3: News volume and median article similarity as predictors of public attention in the domain *Drones*. Note that public attention (measured by Google Trends) climbs sharply with news volume and median similarity, foreshadowing legislation in 2012.

To gain confidence in our findings, we address an alternative hypothesis of note, namely, that political framing Granger causally influences news framing, and not vice versa. We do not in general deny that such a Granger causal direction may exist—indeed such an effect has been demonstrated in prior work using news data collected from print newspapers Flores (2015). However, we did not find that this effect is Granger causally significant for our data over the domains we examine.

In order to do so, we downloaded the Republican and Democratic Party Platforms from 1996 to 2016, and used a simple term search procedure to identify the number of mentions of the domain in each platform. Since party platforms are issued every four years, we used linear interpolation to estimate values for the intervening years between two successive platforms. For the case of LGBT Rights, we also estimated framing polarity of the paragraphs mentioning

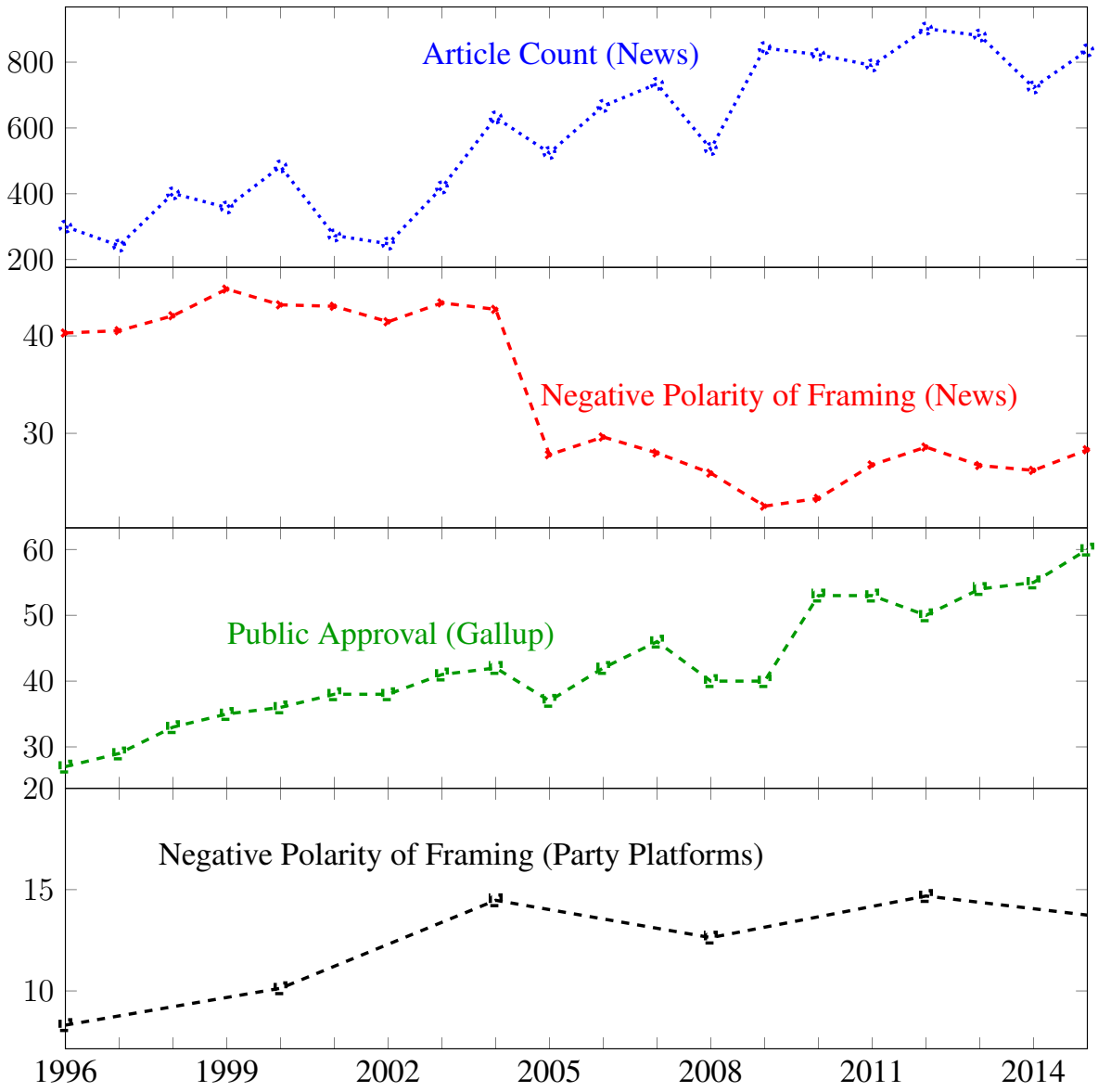


Fig. 2.4: News framing as a Granger causal precursor to public approval changes foreshadowing legislation in the domain *LGBT Rights*. Public approval increases as negative framing declines. Note that the decline in framing polarity after 2004 coincides with a *change in framing* during 2003 described in an earlier survey Engel (2013)

this domain in each platform. Fig. 2.4 depicts the results. G-causality for the (Political Framing, Public Approval), and (Political Framing, Legislation) tuples were insignificant for this example, in contrast to the (News Framing, Public Approval), (News Volume, Legislation) tuples, consistent with our hypothesis. We refer the reader to Table 2.1 for a full list comparing the Granger causal effect of news with the effect of political framing on legislation, for the

domains we consider. We describe full details of this study in the subsection below.

Fig. 2.5 depicts framing density versus time for the Surveillance domain, around the period of the Snowden revelations. Note that framing density is at its lowest in 2014, corresponding to the onset of the Snowden revelations. For illustrative purposes, we use a high minimum count of 20 to depict framing density in this figure. Results with other minimum counts appear to preserve the essential trend (as in fig. 2.6). Further, fig. 2.6 depicts framing density for three domains (Smoking, Surveillance, and LGBT Rights), in which we found earlier studies suggesting that the domain had undergone a framing change. The figure also depicts the framing density of random news. We assume that since random news has no particular "concentration" at any time, it does not undergo changes in framing. Whereas the three domains shown in fig. 2.6 appear to have low values of framing density during periods in which earlier studies found framing changes, the framing density of random news appears generally constant. We take this as evidence that our measure of framing density appears to successfully identify news concentrations that are suggestive of framing changes. For the Surveillance domain, we found Granger causation between framing density (as computed in Fig. 2.5) and legislation.

In fig. 2.6, we used a uniform minimum count of 5 for all domains, to enable a consistent comparison across domains.

It is worthwhile to point out that the Snowden revelations, which we use in Fig. 2.5 to depict framing density, were an event-based driver of news, and not in themselves a framing change. However, the Columbia Journalism Review Vernon (2018) found that following the Snowden revelations, news coverage of Surveillance changed to a narrative focusing on individual rights and digital privacy. We further note that event-based news drivers have often been found to cause framing changes Baumgartner et al. (2014).

Further, we point out that whereas the event of the Snowden revelations took place in late 2013, the legislative response (The Freedom Act) was enacted two years later, in 2015. We show that polarity of negative framing in Surveillance increased following the Snowden revelations (Fig. 2.5), and remained high until 2015, corresponding exactly with our hyper-concentrated period, after which legislation was promulgated and framing polarity increased.

Further, we note that in the domain Child Privacy (Fig. 2.2), framing polarity is at it's highest in 1998, coinciding with the introduction of COPPA.

Since news events cannot affect framing polarity (since framing polarity depends purely on adjectives and adverbs), and we show that both framing polarity and framing density appear

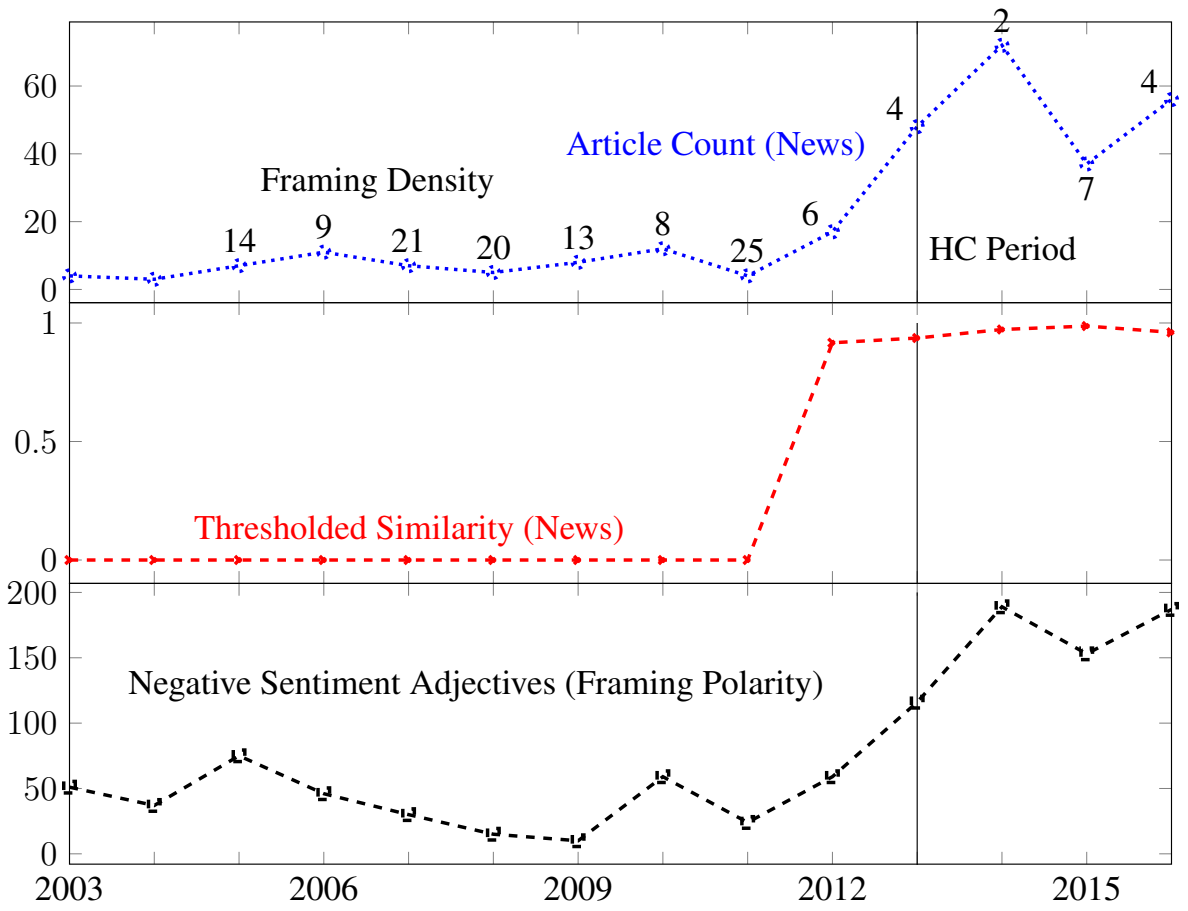


Fig. 2.5: Framing changes may be characterized by low framing density and changes in framing polarity. The figure shows news volume, median article similarity, and framing density in the domain *Surveillance* spike during a hyper-concentrated (labelled HC in the figure) period, foreshadowing legislation.

to have distinctive patterns during framing changes (figs. 2.6 and 2.7), we conclude that news framing can Granger cause legislation.

2.3.2 Hyper-Concentrated News versus Political Framing as a Granger Cause of Legislation

This section details the full results of our Granger causality study. We begin with the complete list of federal legislation promulgated from 1991 to 2016.

From this list, we manually identified domains for which we were able to obtain data, and for

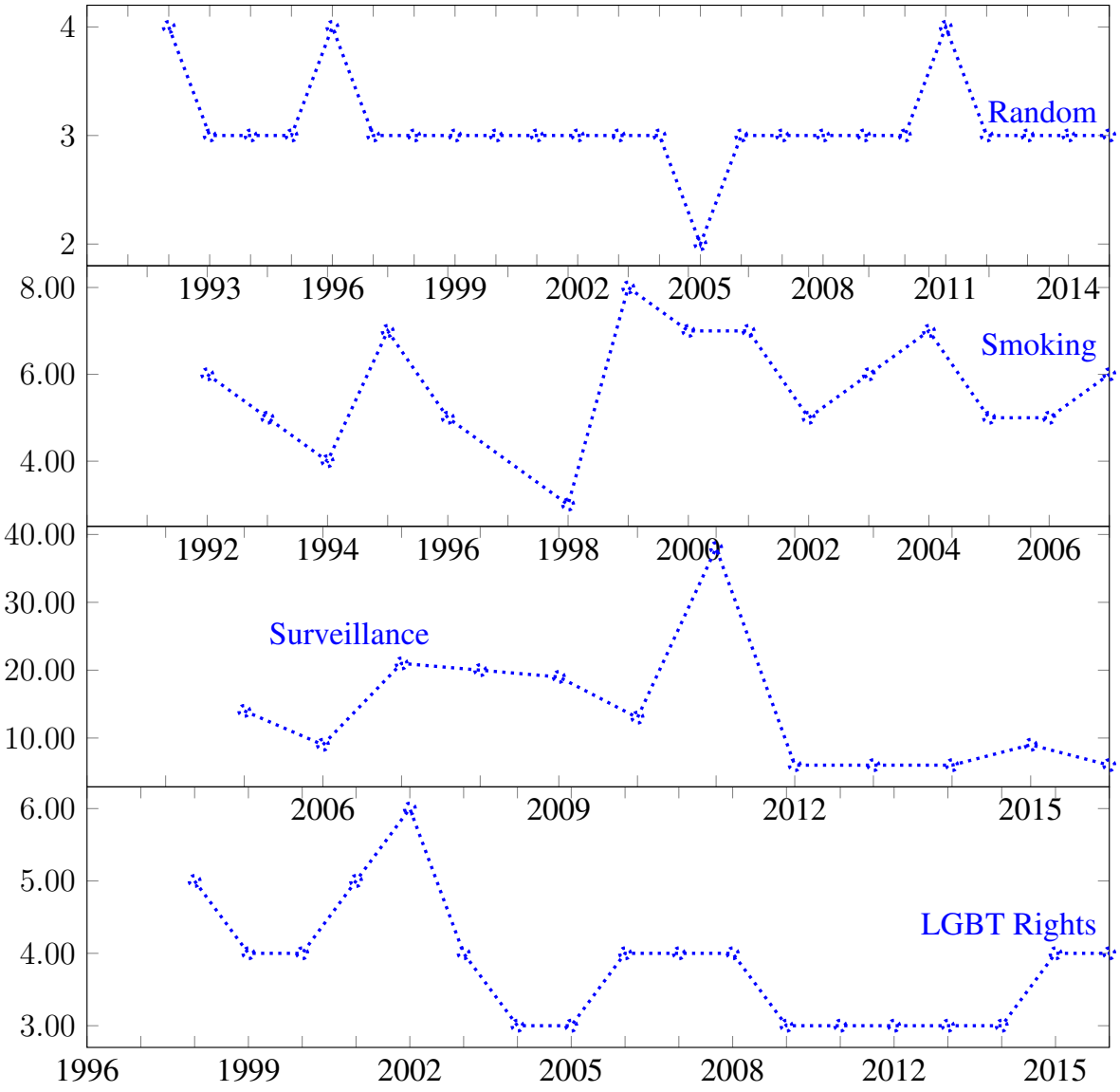


Fig. 2.6: Framing density for random news, versus for framing change positives (Smoking, Surveillance, and LGBT Rights). The values are scaled for visibility, and are not comparable across domains. Notice that the framing density of random news varies much less than that of the framing change positives. Notice further that low values of framing density appear to correlate with periods of framing changes described in earlier studies.

which our data suggested the presence of a hyper-concentrated news period. Table 2.1 depicts this list. We found eight such cases, seven American and one British. We note that there may be further cases which were not identified by our search. We then conducted Granger causality tests between news volume and similarity in these periods (the posited G-causal variables), and the corresponding federal legislation (the posited G-caused ones). We find a Granger causally significant result in each case. Our threshold for significance is $\alpha = 0.05$. For each domain, Table 2.1 lists the smallest significance level at which we obtain a G-causally significant result.

We address the alternative hypothesis that political framing G-causes legislation. In order to do so, we downloaded the Democratic and Republican party platforms from 1996 to 2016, and measured political interest in the relevant domain as the number of mentions of the domain retrieved by a term search in an annual platform. We then conducted G-causality tests with federal legislation in the same domains. For all eight cases, we discovered that political framing did not G-cause legislation at the $\alpha = 0.05$ level. Interestingly, for two of these domains, we obtained a $p = 0.20$ for the hypothesis that political framing G-causes legislation. However, we note that this result is much weaker than the G-causal significance we obtain for hyper-concentrated news.

Some domains remain unmentioned through the relevant period in both party platforms, such as Cyberbullying, Drones, and Child Privacy. For such domains, since the political parties do not mention the domain, we conclude that there was no measurable political framing of these domains (Table 2.1). Therefore, these domains do not affect our hypothesis, given significant G-causal measures between hyper-concentrated news characteristics and federal legislation.

2.3.3 Measuring Domain Framing

Fig. 2.7 shows framing polarity, and fig. 2.6 depicts framing density for three domains (Surveillance, LGBT Rights, Smoking), as well as for a random control. To generate the random control, we retrieved a sample of 991 articles from the NYT API with a null query, for each year between 1990 to 2016.

As is evident, framing polarity of the three domains appears to correlate substantially with the periods of framing change discussed in earlier surveys (see Sec. 2.4.1). As an example,

consider that whereas the framing polarity of LGBT news between 1990 and 2000 remains fairly similar to that of random news in that period (fig. 2.7), it drops between 2004 and 2005, corresponding to the framing change of late 2003, which was reported in Engel (2013). Note also that consistent with our hypothesis, the framing polarity of random news remains close to constant between the years 1990 to 2005.

To depict the framing polarity of Surveillance news in fig. 2.7 on approximately the same scale as that of the other domains (framing polarity of the Surveillance domain is not normalized to the annual article count as described in section 2.2.5 on Framing Polarity, we normalize each entry to the overall sum of entries in this domain over our period of interest.

It is important for us to acknowledge that in multiple domains (Child Privacy, Smoking, and LGBT Rights), framing polarity shows a characteristic drop between the years 2004 and 2005. Since this pattern is apparent across multiple domains, we conjecture that it may be specific to our data source, and not a pattern with particular significance for any given domain. However, the correlations we observe with earlier studies are mostly independent of this observation. For example, the drop in framing polarity of Smoking news between 2000 and 2003 correlates with the findings of National Cancer Institute (2019) (as described in section 2.4.1 on Smoking, before the year 2004. Further, framing polarities in the domains Child Privacy and Surveillance peak during periods corresponding to legislation in these domains. Whereas in the LGBT rights domain, we acknowledge that the drop in the year 2004 to 2005 immediately succeeds the documented framing change of 2003, we believe that the correlation between low framing polarity and increased public approval in this domain is nonetheless worthy of note.

Similarly, our measure of framing density for these three domains (shown in fig. 2.6), depicts a generally constant value for random news, whereas also demonstrating that the framing density of specific domains appears to be low during periods with framing changes.

This observation corroborates our finding that framing polarity and density appear to successfully measure framing.

2.3.4 Comparative Evaluation

Finally, we evaluate the validity of our hypothesis using the comparative method Arend (1971). We conduct tests using the *most different* research design, and explain that the *most similar* research design cannot be used for our data. Full results are presented in section 2.4 on Detailed

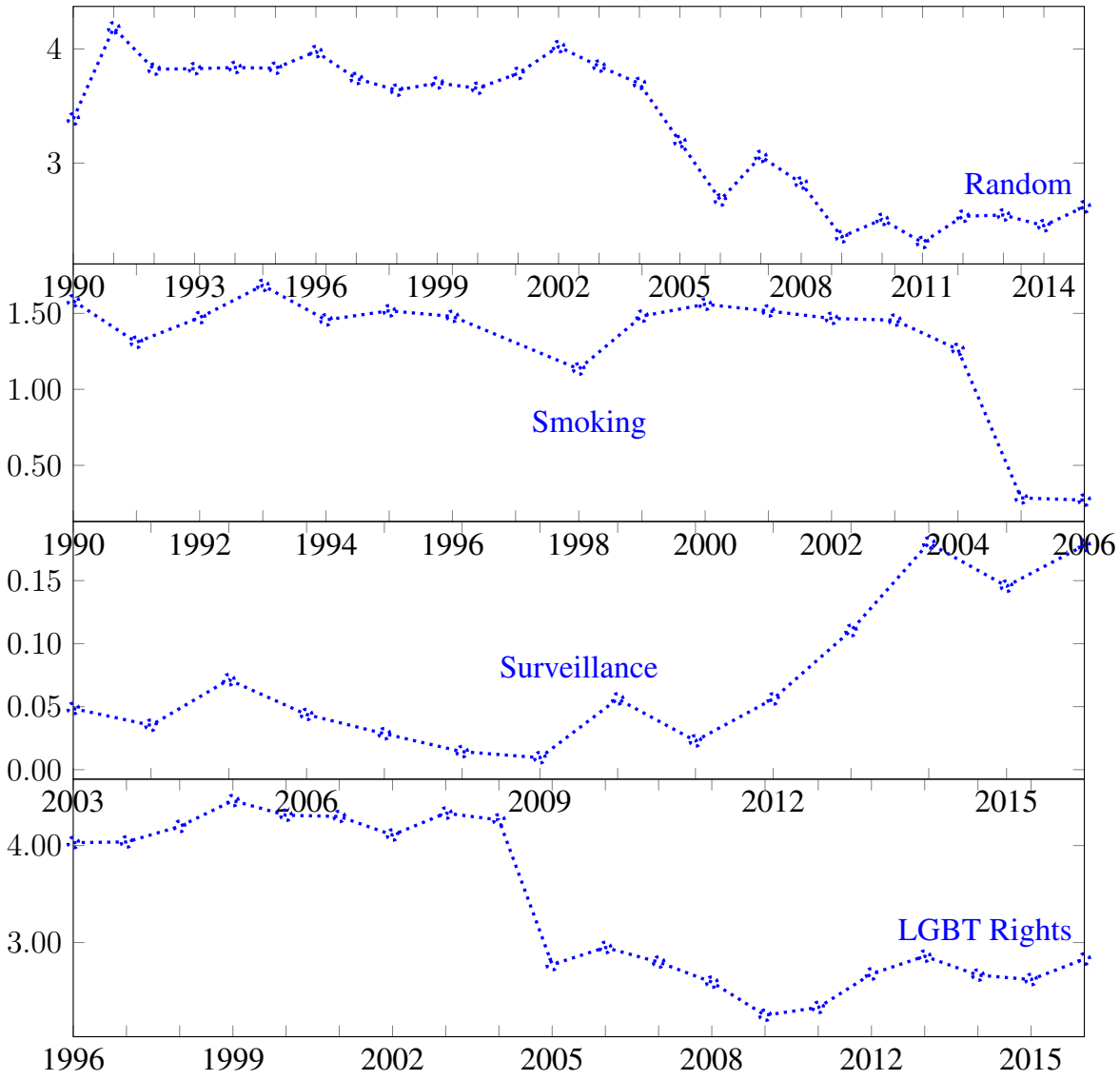


Fig. 2.7: Framing Polarity: Random versus LGBT, Surveillance, and Smoking news.

Results. We summarize our research design and findings here.

We evaluate our hypothesis using the *most different* research paradigm Arend (1971), which relies on comparing strongly different cases, all of which however have in common the same dependent variable, so that any similarity in the independent variables must explain the common value of the dependent variable. In order to estimate "difference" between our domains, we define a custom distance function (Euclidean over our features) based on our news features. We use the following news features as descriptors of each domain: (i) maximum, minimum, and mean annual article volume (used as three separate features), (ii) maximum, minimum, and mean framing polarity (used as three separate features), (iii) maximum, minimum, and mean framing density (used as three separate features). Note that we do not normalize the raw values of our features, since they characterize the domain and we are making inter-domain comparisons. However, we normalize our overall distance to a scale of zero to one. Our data contains ten domains with hyper-concentrated periods. We compute all $\binom{10}{2} = 45$ distances, and pick the top ten to represent our most different domain pairs, shown in Table 2.3. Since in each of these domains, federal legislation was enacted, and further since each domain contains a hyper-concentrated news period (the only common independent variable), we conclude that our hypothesis holds under the *most different* research paradigm. Our domain set changed slightly since our analysis on domain distances was conducted. However, the pattern demonstrating wide variation in our domains remains consistent.

The *most similar* paradigm Arend (1971) relies on comparing highly similar cases that differ only in the dependent variable, as well as in a single or only a few independent variables. Given that the dependent variables differ, the paradigm assumes that the few differing independent variables must be responsible. To use the *most similar* paradigm, we would take advantage of the fact that a domain is most similar to itself. Therefore, to evaluate our hypothesis that hyper-concentrated news periods Granger cause legislation, we would evaluate Granger causality of the domain's news patterns with legislation, both with and without the presence of a hyper-concentrated period.

We were unable to use this research design, since, for many of our domains, there was little or no legislative activity during non hyper-concentrated periods. This in fact supports our hypothesis.

In this context, let us address a concern that our results rely on a particular choice of domains. Note that we exercised no explicit choice in collecting our original set of domains (we examined all federal legislation in the periods for which the NYT and Guardian APIs provide

data). We then analyzed eight of these domains for which our data indicated the presence of a hyper-concentrated news period. We acknowledge however that there may be additional domains with hyper-concentrated periods that our search omitted. As we show in Table 2.1, all of these domains Granger cause legislation at the $\alpha = 0.05$ level. However, our comparative analysis demonstrates through the *most different* (Table 2.3) paradigm that our hypothesis remains valid despite wide variation in the domains.

2.4 Detailed Results

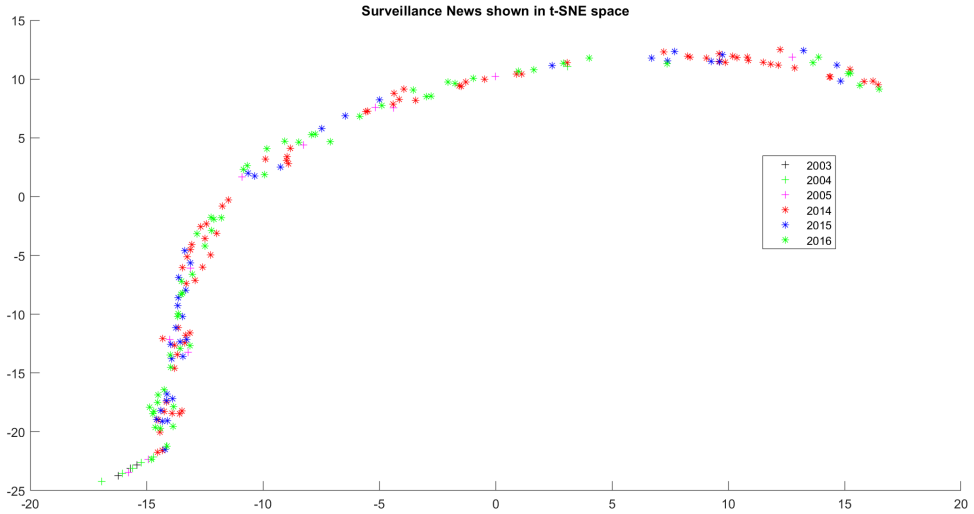


Fig. 2.8: The figure visualizes six years from our Surveillance dataset. Each has high median similarity. Three correspond to years with low article volume (2003 to 2005), and three to years with high article volume (2014 to 2016). The latter years are hyper-concentrated, combining high article volume and high median similarity. We used t-SNE to project our Surveillance dataset to two dimensions. Since t-SNE uses a random seed, different runs of our code will produce slightly differing visualizations, though the trend should be preserved.

2.4.1 Measuring Domain Framing

Section 2.3.1 on Main Findings and section 2.3.3 on Measuring Domain Framing describe how our measures of framing polarity and density appear to correctly measure domain framing.

The aforementioned sections discuss correlations between our measure of framing polarity and the results of earlier surveys in the domains of Surveillance and LGBT Rights.

Here, we depict our measures of polarity and density for four domains (including the aforementioned, Surveillance and LGBT Rights). We note that these domains have been the subjects of study in earlier work such as Baumgartner and Jones.

We also provide a discussion of the results of our approach on the domain *smoking*. We omit descriptions of results in other domains such as *surveillance* and *LGBT Rights* which were previously discussed in section 2.3.1 on Main Findings.

Smoking

The National Cancer Institute recently published a monograph on the influence of the news media on tobacco use National Cancer Institute (2019). On page 337, the monograph describes how, between 2000 and 2003, American news media progressed towards pro-tobacco control frames, with 55% of articles reporting progress on tobacco control, whereas only 23% reported setbacks.

We contrast this to the monograph's earlier finding (on the same page) that between 1985 to 1996, pro-tobacco control frames (11) were fairly well balanced with pro-tobacco frames (10).

Our framing polarity of Smoking news, depicted in fig. 2.7 shows a consistent trend between the year 2000 and 2004, coinciding with National Cancer Institute (2019). This finding supports our hypothesis that our measure of framing polarity appears to correctly measure framing.

2.4.2 Legislation

This section provides a complete list of cases for which we found G-causality between hyper-concentrated news and legislation.

Surveillance

Following the Snowden revelations of June 2013, the USA Freedom Act was introduced in the US Congress in October 2013, and was finally passed into law in 2015.

To test G-causality in this domain, we estimate legislation using a binary time series beginning in the year 2003. Whereas all entries except those pertaining to 2014 and 2015 are represented with zeros, these two years with legislative action are represented with ones.

Cyberbullying

Although there are no federal Cyberbullying laws as of 2018, we compiled a comprehensive list of statewise Cyberbullying laws. We harvested news articles from 2003 when reports of Cyberbullying begin to appear, until 2016. Fig. 2.9 visualizes the number of state Cyberbullying laws enacted in a given year, alongside Cyberbullying news volume and median article similarity.

Drones

Drone legislation in America was first promulgated in 2012 Vanian (2015), and Senate debate on the subject has been active since. Fig. 2.3 depicts news patterns for this domain. We tested our approach on this dataset by using the data shown in Fig. 2.3. As can be seen from the figure, publishing volume and mean similarity reach a peak for the year 2012.

We tested our hypothesis using two configurations of keyword search for this domain, as described in the main text. We consider a period of interest between 2003 and 2012. Using the generic keyword search (“drones”), we retrieved approximately 4,000 articles. Using a more focused keyword search (“drones+privacy”), we obtained nearly 300 articles.

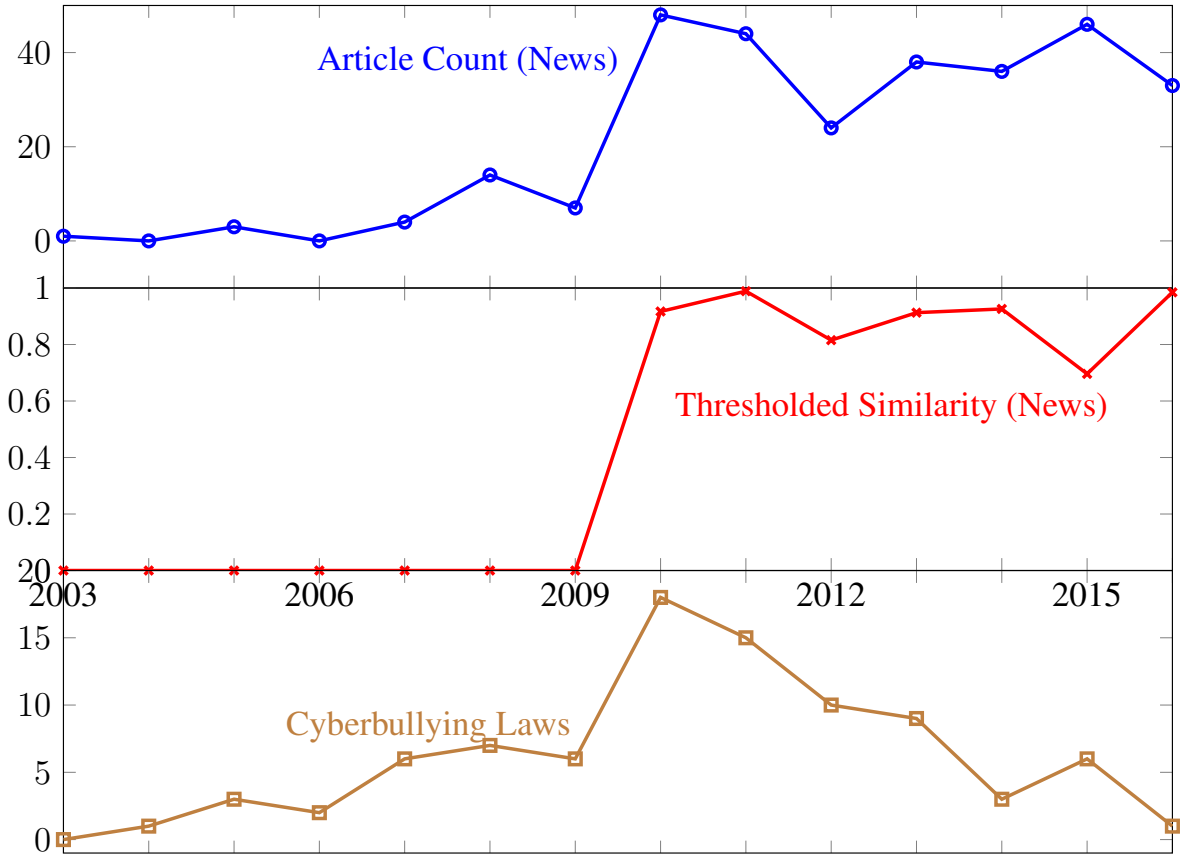


Fig. 2.9: News volume and similarity as predictors of legislation in Cyberbullying.

Child Privacy

We discuss this domain in detail in the Main Findings (section 2.3.1) section of the main text. For the benefit of the reader, we restate the main points and findings here. We collected children's privacy news articles from 1990 to 2016 from the New York Times API (a total of 2,011 articles). Fig. 2.2 in the main text visualizes news volume together with similarity. The figure displays a clear correlation between news volume and similarity, and legislation.

Our legislation time series for this domain is determined by the number of FTC press statements following major legislative activity in the domain. We use a value of ten for the year 1998 in which COPPA was enacted.

We find G-causality in this domain for the first ten years of our analysis (between 1990 and 1999), corresponding to a substantial period before, and one year immediately succeeding leg-

islation.

Housing

We analyze news patterns preceding the Housing and Economic Recovery Act of 2008. We analyze news publishing beginning in the year 2004. We generated a dataset of 753 articles for this domain.

Since legislation in this domain focused on subprime mortgages, we used the term "housing+economy" to cncentrate our article search on both topics.

Our legislation time series in this domain is a binary vector of four zeros (corresponding to 2004 to 2007) and one positive entry (corresponding to 2008).

Abortion

The Partial-Birth Abortion Ban Act was enacted in 2003. We compiled a dataset of 239 articles over the years 2000 to 2004.

We were unable to obtain data from our API for the period preceding the year 2000. Legislation in this domain is represented by a binary time series with three zeros and one positive entry.

Interestingly, we note that in this domain, whereas our measure of news volume oes not yield significant G-causality with legislation, we do obtain a significant G-causal measure between median news similarity and legislation.

This highlights the utility of the similarity characteristic introduced in this work, as an enhancement over the two main news characteristics introduced by Baumgartner et al. (2014) (*attention* and *tone*).

LGBT rights

We use the list of state LGBT laws presented in https://en.wikipedia.org/wiki/Same-sex_marriage_legislation_in_the_United_States as our dependent variable. We analyze G-causality between news and legislation over the period 1996 to 2015 in this domain. We retrieved a total of 3368

articles on this domain. We used two open source Matlab implementations to evaluate Granger causality. Once again, we note that whereas one implementation yielded significant G-causal results with legislation for both news volume and similarity, the second implementation found G-causality only between median similarity and legislation. This result further underlines the contribution of our similarity measure.

Criminal Justice and Courts Act 2015 (UK)

The October 2014 amendment to the Criminal Justice and Courts Act made the act of revenge pornography an offence. The law was passed under the Criminal Justice and Courts Act of 2015.

We retrieved 375 articles from the Guardian API for this domain between 2003 and 2015. We used the full text of the article to train and evaluate paragraph vector similarity.

We found G-causality between news volume in the domain Cyberbullying (in the Guardian) and these legislative acts.

Table 2.1: Comparing the Granger causal effect of hyper-concentrated news against that of political framing for legislation in our domains. The abortion entry, shown with a *, refers to G-causality between similarity and legislation. All other entries refer to G-causality between news volume and legislation.

Domain	**Hyper-concentrated News**			**Political Framing**		
	F-Statistic	**Critical Value**	**P-value**	**F-Statistic**	**Critical Value**	**P-value**
Child Privacy	96.99	12.25	< 0.01	NaN	NaN	NaN
Cyberbullying	13.13	4.46	< 0.05	NaN	NaN	NaN
Drones	7.29	5.59	< 0.05	NaN	NaN	NaN
Cyberbullying (United Kingdom)	39.23	7.56	< 0.01	NA	NA	NA
LGBT	4.27	3.10	< 0.05	2.78	1.78	< 0.20
Surveillance	136.32	9.65	< 0.01	2.39	2.35	< 0.20
Abortion Ban 2003*	26.87	21.19	< 0.05	Inadequate data	NA	NA
Housing 2008	22.94	18.51	< 0.05	0.33	0.26	< 0.70

2.4.3 Results from the Comparative Method

Table 2.3 lists the full results from our comparative study using the *most different* paradigm.

Table 2.2: Details of our Granger causality study. The table shows the value of the min count used for training each domain's doc2vec model and the lowest lag at which either the domain's news volume or its median similarity G-caused legislation. The table further lists which of news volume and/or median similarity G-caused legislation in the domain. Entries with a star indicate that significance was found using a lag value greater than 2 (and therefore possibly signify a weaker correlation). We note that in the LGBT domain, the value of the lowest lag for volume was 6.

Domain	**Search Term**	**Min Count**	**Lowest Lag**	**Volume**	**Similarity**
Child Privacy	"child+privacy"	1	1	✓	×
Cyberbullying	"cyberbullying"	1	1	✓	✓
Drones	"drones"	30	1	✓	×
Cyberbullying (United Kingdom)	"cyberbullying"	50	2	✓	✓*
LGBT	"gay+rights"	1	1	✓*	✓
Surveillance	"surveillance+privacy"	1	1	✓	×
Abortion	"abortion"	1	1	×	✓
Housing	"housing+economy"	1	1	✓	×

Table 2.3: A comparative evaluation of our hypothesis using the *most different* research design.

Domain Pair	**Domain Distance**
LGBT, Cyberbullying	1.00
LGBT, Drones	0.97
LGBT, Surveillance	0.97
LGBT, HTML	0.97
LGBT, Abortion 2003	0.91
LGBT, HIPAA	0.78
LGBT, Housing	0.69
LGBT, COPPA	0.64
COPPA, Cyberbullying	0.38
COPPA, Drones	0.36

2.5 Conclusion

Our data supports our conclusion that hyper-concentrated news periods in news, brought about both by driver events and framing changes, Granger causally influence public attention and federal legislation. We acknowledge, however, that our analysis does not disprove reverse causality, and we do not model confounding factors beyond those discussed in the paper.

Chapter 3

Toward a Unified Model of News Influence on Public Interest and Legislation

This chapter addresses our second research question, RQ_2. It describes our discovery of the three interest regions in topic news publishing, and posits a systematic selection bias that governs the nature of news selection in each topic, and the likely nature of public reaction to it. It is based on a paper of the same title in review at the *Proceedings of the National Academy of Sciences*.

3.1 Introduction and Contributions

The question of what drives public interest in a particular topic has long been the subject of investigation by social and political scientists. Recent advances King et al. (2017); Sheshadri and Singh (2019) have shown that news patterns measurably influence interest in a specific domain. Our use of the phrase "news pattern" refers to a consistently observable pattern in a news variable or variables. As an example, consider the increase of news volume shown in Fig. 3.6, from about 200 articles in 2004 to about 800 articles in 2011. We refer to this pattern (of increasing article volume) as a news pattern. The hyper-concentrated period posited by Sheshadri and Singh (2019) is another example of a news pattern that involves two variables.

Existing research tends to focus on individual aspects of this relationship. For example, King et al. (2017) shows that news publishing in small outlets causes local readers to engage in public

discussion on the topics covered by the news. Earlier efforts Soroka (2006); McCarthy et al. (1996); Oliver and Maney (2000); Lundman (2003) similarly focus on identifying particular relationships that may influence public interest in a specific case.

News characteristics such as framing Sheshadri and Singh (2019), selection Bourgeois et al. (2018), prominence Lundman (2003), and sentiment Soroka (2006); Sheshadri et al. (2017) have previously been shown to influence public reaction. How these characteristics interact in a given situation, which of them are likely to be influential under a given set of conditions, and what reaction they elicit remain open questions. A unified model of news influence with generic predictive utility should, given the state of news in a particular domain, be able to estimate values for news characteristics such as those listed above, and model their interaction to predict the nature and quantum of public reaction to a given news pattern. Our work represents substantial and promising progress in this direction. We describe our main contributions below.

3.1.1 Contribution: Uniform Selection Bias of Three Fundamental Interest Regions

We refer to a broad collection of related news topics such as politics or sports as a *domain*. We identify three fundamental domain-specific factors that appear surprisingly stable relative to others: news volume, news prominence, and public interest.

Consistently across these factors, we posit that any domain predictably belongs to exactly one of three interest regions. The *primary* region includes domains such as *war* and *politics*, the *secondary* region domains such as *economy* and *sports*, and the *tertiary* region domains such as *privacy*. Surprisingly, we find that whereas domains may shift positions within a region, domains usually do not migrate to a different region within the timescale of the few decades that we examine.

We observe that news prominence increases from the primary region, where an article from a given domain is on average least likely to be prominently featured, to the tertiary region where the corresponding likelihood is high. We measure interest from news portals in a given domain using its domain volume. We measure interest from the public in a given domain using the relative percentage increase in number of web searches of relevant queries reported on Google Trends (see the Distribution of Google Trends Response over Domains section (Section 3.2.2)).

We observe that interest both from news portals and from the public decreases systematically from the primary to the tertiary region.

This finding may appear counter-intuitive at first since there is a prevalent view The Washington Post (2014) that domains such as *politics* and *sports* (domains mainly from our primary region and sometimes from our secondary region) tend to be those most prominently featured. However, we explain that this finding is consistent with our conception of the three interest regions.

We point out that our conception of interest regions bears a direct relationship to *news selection*, which refers to the process of a news source (web portal or newspaper) choosing which events to report as news.

We show that in any sufficiently large time period, the distribution of domain volume (the number of news articles published in a domain) across all domains and the distribution of public interest change across all domains follow a rapidly decaying distribution, resembling a Zipfian. We use the relative percentage increase in the number of relevant web searches retrieved via Google Trends as a surrogate estimate for public interest.

Surprisingly, we find that the domains within each region are largely the same across these two distributions (see Fig 3.3). This is a notable finding, since it is not the raw volume of Google Trends queries across domains that correlates with domain news volumes, but instead, the *percentage of change*.

One might intuitively expect that a domain that features more frequently than others in the news (in other words, has a large news volume), would be more frequently searched for on the internet (and consequently have a correspondingly high Google Trends query volume).

Surprisingly, we found that this was not the case. As an example, the domains *movies* and *sports* have consistently higher raw values of Google Trends queries than the domains *politics* and *war* (with an average of 77 for movies and 48 for *sports* versus only 3 for *politics* over the period 2004 to 2018). However, following a significant event in the domain *politics*, such as the Presidential election of November 2016, Google Trends query volume of this domain increased from 21% to 49%. Whereas by comparison, after the release of *Star Wars: The Last Jedi* in December 2017, the corresponding increase for the domain *movies* was only 14%, despite the fact that *The Last Jedi* grossed over a billion dollars worldwide, making it the seventh largest grossing movie of all time Wikipedia (2017).

Thus, we find that it is the *percentage increase* in Google Trends query volumes after such significant events that correlates with the orders of domains in the news distribution.

This suggests that the public reacts preferentially to certain topics over others, according to the rapidly decaing distribution shown in Fig. 3.2. This is a novel finding of our work.

Whereas existing work has established a Granger causal Granger (1969) link between news and Google Trends query volume Sheshadri and Singh (2019), we posit an underlying theory that explains this link.

Motivated by the results of the Experiments section (Section 3.3), we posit that the level of interest the public displays for a domain relates to the immediacy of its effect. Accordingly, domains such as *war* and *politics* that have a direct immediate effect on society tend to elicit the strongest level of interest. These domains comprise our primary interest region, characterized by both high news volume and a high percentage of change of Google Trends query volumes.

The secondary interest region consists of domains that are potentially impactful for society, but not immediately and directly so. Domains such as the *economy*, *nutrition*, *real estate*, and *employment* comprise this region, which is characterized by moderate volume and prominence.

Finally, the tertiary region comprises domains that are unlikely to immediately impact society. lives. Domains such as *privacy* inhabit this region.

These observations serve to elucidate a uniform *selection bias* that applies both to news outlets and readers. This bias therefore systematically affects both news publishing and public reaction to news. We observe that the number of news articles from the domain *war* in 2003 (about 16,000) was still well below the number of articles on politics (about 24,000), despite the onset of the war in Iraq. Similarly, the events of September 11^{th} 2001 resulted in only 624 articles on terrorism from September 11^{th} to December 31^{th}, considerably fewer than the number of politics articles in the same period (about 9,516). This suggests that humans (including news outlets) tend to favor certain domains over others with decaying attention across domains, as shown in the distribution of Fig. 3.2. We confirm the existence of this systematic selection bias by demonstrating that the overall number of events that take place across these domains in any given year is substantially closer to a uniform distribution than to an exponential one (Fig. 3.2).

We note that this selection bias applies uniformly from news selection to public reaction. Our conception of the three interest regions, which is inferred from the observed news and Google Trends data, is suggestive of how the public may react to news in a specific domain as opposed to another domain. A notable example is that whereas media coverage of political corruption and Government influence on corporations (*politics*, and hence in the primary interest domain) preceded the Occupy Wall Street protests, the Snowden revelations (*privacy*, and hence in the tertiary interest domain) did not generate a similar public reaction.

The fact that articles from the tertiary region are featured most prominently is consistent with our hypothesis of selection bias. Since news portals select many more events from the primary region than the tertiary to report as news, it is correspondingly harder for an event from the tertiary region to be selected. Events from the tertiary region that satisfy this threshold are therefore correspondingly more influential, and are thus displayed more prominently.

3.1.2 Contribution: A Unified Model of News Prominence and Legislation

We adopt a data-driven approach that infers a unified model of prominence. Using this model, we demonstrate that we can predict news prominence on unseen data with an area under the Receiver-Operating Characteristics curve of 0.81. These results are notable in that whereas previous efforts are restricted to identifying individual factors that may influence prominence, our work is the first to produce a full predictive model that performs accurately on unseen data.

Our model yields the following interesting findings. We find that domains in the tertiary region are most likely to be prominently featured. We observe that articles with a high polarity score, either positive or negative, are less likely to be featured prominently by newspaper outlets than those with a mild sentiment. We show that concentrated domains (with a *dominant frame* characterized by a large number of points occurring within a tight cluster) tend to be more prominently featured. We find, in addition, that newspaper outlets tend to feature movie and product reviews more prominently on average than news.

We note the relationship prominence bears to the problem of news selection. Whereas selection refers to the process of choosing an event to report in a news article, prominence is the measure of how conspicuously an article is displayed in a print or online newspaper. Although selection

and prominence refer to distinct processes, we posit that they are conceptually similar, in that both involve evaluating how important an event is to a newspaper. Therefore, the conclusions from our analysis of prominence serve to explain selection.

Our model of prominence relies on several independent news variables, which we show may exhibit predictable covariance (rather than being independent). One of the variables we examine is *news concentration*, as introduced by Sheshadri and Singh (2019).

We show that within a domain, news volume and prominence may increase and decrease together. We further show that periods with high concentration may tend to occur simultaneously with high prominence. We show that prominence in itself tends to correlate with, and can Granger cause, federal legislation.

Motivated by these observations, we posit a novel Granger causal flow, which enhances understanding over recent models King et al. (2017); Sheshadri and Singh (2019). We obtain initial evidence for Granger causality between the (Volume, Concentration) and (Concentration,Prominence) pairs. We posit Granger causality between all of the (Volume, Concentration), (Volume, Prominence), and (Concentration, Prominence) pairs. Whereas we acknowledge that limitations in our data prevent us from systematically establishing every link in our hypothesized flow, we believe that our existing data provides a sufficient basis for further investigation. We also show that news prominence in itself can correlate with, and may Granger cause, legislation. We further show that the three news characteristics we consider (Volume, Concentration, and Prominence) vary predictably across our interest regions. Our Granger causal flows are shown in Fig. 3.1.

3.2 Materials and Methods

This section describes the novel methods we introduce in this paper.

3.2.1 Dataset Collection

We describe our data sources, method for domain dataset generation, and inter-annotator agreement below.

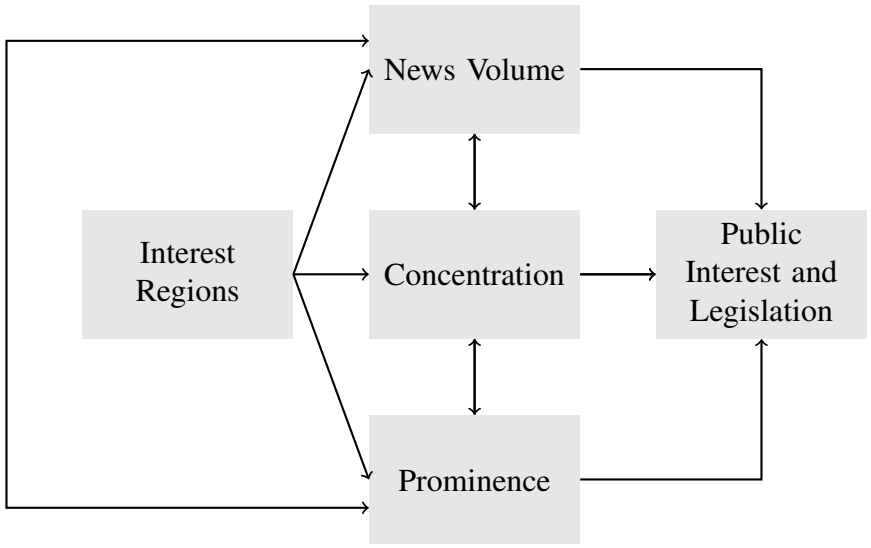


Fig. 3.1: Motivated by our data, we posit granular Granger Causal relationships between news volume, concentration, and prominence (the G-causal variables) and public interest and legislation (the G-caused ones).

Data Sources

We use three primary Application Programming Interfaces (APIs) in our analysis.

The New York Times API:
The New York Times (NYT) Developer's API NYT (2016) provides access to news data from the NYT, Reuters, and Associated Press (AP) newspapers (both print and online versions) beginning from 1985. The NYT has the second largest circulation of any newspaper in the United States Wikipedia (2001).

The data returned by the API includes fields such as the article type (news, reviews, summaries, and so on), the news source (NYT, Reuters, or AP), the article's word count, the date of its publication, article text (in the form of the abstract, the lead (first) paragraph, and a summary). Importantly, the data provides the article's print page, which we use to compute an NYT article's prominence score.

The Guardian API:
The Guardian Xplore API The Guardian (2016) provides access to news data from The Guardian, a prominent UK newspaper, which is read by approximately 23 million UK adults Wikipedia (2002).

The Guardian API returns full-length articles along with such metadata as the article type (similar to NYT) and a general section name (such as sports, politics, and so on). Although these section names are manually annotated by humans, we do not use them in our analysis, but rely instead on our Domain Dataset Generation approach (Section 3.2.5) to annotate our datasets.

Google Trends API:

The Google Trends API Trasborg (2018) provides monthly Google Trends data on a scale of 1–100 for any input keyword. Whereas the number returned is ordinarily at a scale that is specific to the domain, the Google Trends API supports a comparison across domains by simultaneously querying multiple keywords. This comparison enables us to obtain an annual Google Trends index for each domain on a scale that is uniform across domains. We describe our computational method in the Distribution of Google Trends Response over Domains section (Section 3.2.2).

The Google Trends API may be queried to obtain five independent data types: Web, Image, Youtube, News search, and Google Shopping Trends. We ignore Google Shopping since our hypotheses do not directly address this dimension of human behavior. Since News search returns only domain volume and not the data itself (and therefore cannot be used to compute other news characteristics such as concentration, sentiment, and so on), we ignore it and rely instead on our dedicated news APIs for our news analysis.

The Web, Image, and Youtube search function of the Google Trends API is popularly referred to by the public as "Google Trends," since it returns the volume of search queries relevant to a given domain. We restrict our analysis to these dimensions of the Google Trends API.

Event Registry:

The Event Registry Leban et al. (2014) is an open access dataset of worldwide activities, compiled from over 30,000 sources worldwide.

The Event Registry actively monitors a wide range of news sources (broadcast, print, and web), recording and annotating global events and their coverage. Since the Event Registry does not return any measure of news prominence, we are unable to use its data in our prominence analysis. We use it to estimate the number of events that took place in the US and UK within a given domain over our period of interest.

Domain Set Generation from Google Trends

The Google Trends API lists 25 commonly used categories. These include Entertainment,

Beauty and Fitness, Business and Finance, Food, Health, Jobs, Law and Government, Real Estate, Sports, and Travel.

3.2.2 Estimating Distributions

To estimate our distributions, we use the set of domains obtained using the domain set generation procedure described earlier.

Our news and Google Trends query volume distributions are computed over the period 2004 to 2018, which we refer to as our *period of interest* below.

For uniformity, the experiments of the Distribution of Prominence over Domains and Distribution of Prominence over Sentiment sections (Section 3.2.2) were computed over the year 2015 from the NYT, for which we were able to retrieve data for many of our domains.

Universal News Dataset

We compile a universal news dataset for our period of interest by querying both news APIs with a null query over this period.

Distribution of News Volume over Domains

We query our news APIs for our keyword set (see section 3.2.5 on Domain Dataset Generation) over the period 2004 to 2018. We compute the average annual news volume within each domain to serve as its overall volume.

Distribution of Google Trends Response over Domains

We are interested in comparing the percentage increase of Google Trends query volume after an influential event in a given domain, relative to the corresponding percentage increase in other domains. We rank domains based on this relative magnitude of increase in public interest after an influential event. As described in Section 3.2.1 on Data Sources, the values returned by the Google Trends API for each domain are on a scale that is specific to the domain. We

therefore use pairwise comparisons between domains, which we then combine to arrive at a total ranking.

We do not attempt to compile an exhaustive list of events in each domain over our period of interest, since such a procedure may result in important events being omitted. Further, we have no well-defined metric to rank events in order of their relative influence. Instead, we identify the largest percentage increases of Google Trends query volume for a given domain within our period of interest.

Thus, for each domain pair over our period of interest, we collect monthly Google Trends query volumes. We compile a list of *monthly differences* for every month in this period, for each domain pair. We do so by calculating the absolute value (since we are interested in the magnitude of change between a month in which the public is demonstrating interest in the topic and a month in which public interest has changed) of the difference between the present monthly value and the previous one, for every month.

We thence arrive at a pairwise monthly difference vector $V = [V_1, V_2]$ for each domain pair. V therefore has n rows, where n is the number of months in our period of interest, and two columns. We separately sort each vector and collect the top k pairs V_k. We count k_1, the number of rows in V_k where the first column has a value greater than the second, and k_2, the number of rows where it has a value less than or equal to the second. We rank as d_1 greater or lesser than d_2 in our distribution according to whether $k_1 \geq k_2$ or vice versa.

We illustrate this approach using an example domain pair (*economy* and *healthcare*) in Tables 3.1 and 3.2.

Notice the Google Trends query volumes for the domain *economy* from January to August 2004, as shown in Table 3.1. Interest increases from January to its peak in May, after which it dies away during the next three months. We note that this surge in interest is contemporaneous with the 2004 State of the Union address by President Bush Gongloff (2004), in which he introduced several new economic policy proposals. The succeeding economic surge is recorded to have continued through April Uchitelle (2004).

Our method captures the change in interest in this domain by computing the difference in Google Trends query volume between a period during which public attention was high (May 2004) and a period when it had died away (June 2004). We note that we are interested in the absolute value of this difference, since a period of high interest may either precede or succeed a period of low interest in a generic domain.

Table 3.1: Illustrating our method for comparison of Google Trends query volume between domain pairs.

	Domains		Absolute Domain Differences	
Month	Healthcare	Economy	Healthcare	Economy
2004-01	36	14	NA	NA
2004-02	37	15	1	1
2004-03	35	17	2	2
2004-04	40	17	5	0
2004-05	35	18	5	1
2004-06	36	10	1	8
2004-07	37	8	1	2
2004-08	35	9	2	1
2004-09	38	14	3	5
2004-10	40	17	2	1
2004-11	36	16	4	1
2004-12	31	12	5	1
2005-01	34	12	3	2
2005-02	34	14	0	3
2005-03	35	15	1	1

Table 3.2: (Continued from above) Illustrating our method for comparison of Google Trends query volume between domain pairs.

Sorted Differences (top k)		
Healthcare	Economy	Greater Rank Domain
5	8	Economy
5	5	NA
5	3	Healthcare
4	2	Healthcare
3	2	Healthcare

Domain Prominence

We refer to prominent articles as those that were published on the first page of an online news portal, or on either the first or last page of a print newspaper.

We compute the prominence of a domain corpus as the fraction of articles within it that are prominent.

We briefly discuss an alternative definition of domain prominence that is not used in the present work. Whereas our above definition relies on corpus volume, section 2.3.1 of Chapter 1 shows that domain dataset accuracy can vary according to the specific search terms used. Further, credible coverage with sufficient recall is difficult to guarantee. Therefore, we hypothesize that the raw volume of articles from a domain that are featured prominently, without normalization to overall domain volume, may also represent a useful definition of domain prominence.

Distribution of Prominence over Domains

For each domain, we collect all articles published in that domain during our period of interest. We obtain the domain prominence as defined above for each of these domains, to arrive at our distribution.

For uniformity, this distribution was computed over the year 2015.

Distribution of Prominence over Sentiment

We use the Stanford sentiment analyzer Socher et al. (2013) to obtain a sentiment polarity score on a scale of zero to one for every article in our universal set. We then create ten nonoverlapping clusters by binning our articles according to their polarity scores, in intervals of width 0.2 from -1 to 1.

We compute the total number of prominent articles as defined above in the universal set, which we refer to as P_u. We also compute the total number of prominent articles in each bin i, which we refer to as P_i.

We compute the prominence score of each bin i as $\frac{P_i}{P_u}$.

For uniformity, this distribution was computed over the year 2008.

3.2.3 Distribution Fitting

For each distribution, we fit the data it contains to several candidate models using Matlab's fit function Mathworks (2019b). We use the following candidate models: polynomials of degrees one and two, a normal distribution, a power law, and a Beta distribution.

The fit obtained for each candidate distribution estimates the values of the distribution coefficients (such as the exponent of the power law, the constant multipliers in a polynomial fit, and so on), and other data such as the R^2 statistic. We use this data to obtain a Root Mean Square (RMS) error for each candidate distribution. We choose the fit that minimizes the RMS error.

3.2.4 Event Registry Event Analytics

The Event Registry Analytics Tool returns a monthly intensity score representing the number of global events that took place in a given domain in the concerned month. For each of our domains, we compile a median event score over several months, which serves as our overall domain event score. We note that the values obtained were fairly uniform across different months. We restrict our geographical locations to the US and UK which our news sources focus on.

3.2.5 Domain Dataset Generation

As in earlier work King et al. (2017); Sheshadri et al. (2017); Sheshadri and Singh (2019), we use a standard term search procedure to create our datasets. For each domain, our APIs were used to extract news data during the time period b (denoting the beginning) to e (denoting the end), of the period of interest.

3.2.6 Variance in API Results

We note that the results returned by our APIs for the same queries can change over time. For concreteness, we finalized our datasets on December 17, 2018. We provide all our datasets

and code along with our submission. Similar analyses conducted with our model for other snapshots of the data should preserve the trends in our results.

3.2.7 Predicting Prominence

We follow a supervised learning approach using Random Forests Breiman (2001) to predict prominence. We use our universal dataset to train and test the supervised model. The input features to the classifier include all news characteristics that we found to affect prominence. These include the article text, its domain, sentiment polarity, word count, article type as well as the domain's volume, concentration, sentiment, and prominence. We represent article text using paragraph vectors Lau (2017). Our forest consists of 100 independently trained trees.

3.2.8 Corpus Visualization

A domain corpus is a set of news articles from a given domain. We are interested in visualizing annual domain corpora over our period of interest. For a particular domain, let this period have m years with annual domain corpora T_1, T_2, $\ldots T_m$. Let $T = T_1 \cup T_2 \ldots ... \cup T_m$ denote the combined domain corpus. We learn a word vector representation W Le and Mikolov (2014b) of T.

For every $1 \leq i \leq m$, we extract all nouns in T_i, and visualize their word vectors in W using t-SNE van der Maaten and Hinton (2008).

Our visualizations demonstrate that periods of macromutation correlate with the appearance of dense noun clusters in the data, as shown in Fig. 3.5.

Corpus Concentration

We describe a domain corpus T using a Term Frequency matrix U of $|T| \times s$ dimensions, arranged by year, where $|T|$ is the number of articles in T and s is the number of n-grams used to describe the corpus.

Each column of U therefore represents a particular n-gram, and is a list of observations of the term-frequency of this n-gram over the set of articles in U. Each column is therefore a random variable over the term-frequency of an n-gram over the set of articles in U.

For each year i in the corpus, we extract U_i, all the rows of U corresponding to articles from year i. We compute the vector v_i of Pearson correlations Benesty et al. (2009) between all $\binom{s}{2}$ pairs of columns (n-grams) in U_i, and use it as our annual measure of corpus concentration. Since we are interested in the magnitude of the correlations, we convert each entry in v to its absolute value, and then evaluate the mean of v_i to represent our annual measure of corpus concentration.

Inter-Annotator Agreement

A random sample from each domain dataset was coded by two raters. An article is considered as belonging to a domain if and only if the article could not be published with the domain component removed. As an example, consider the article entitled "Was Lyndon B. Johnson a civil rights mastermind, or a reluctant follower pulled along by activists led by the Rev. Dr. Martin Luther King Jr.?" of the domain *war*. We coded it as a negative since, whereas it contains a reference to the Vietnam War, it primarily discusses the presidency of Lyndon Johnson, and could therefore have been published with the component on the Vietnam war removed. We obtain a median per-domain accuracy of 0.83. We measured inter-annotator agreement using Cohen's κ. Our median agreement was $\kappa = 0.82$, considered "substantial agreement" by Landis and Koch (1977).

3.3 Experiments

We describe experiments corresponding to each of our contributions below.

3.3.1 Selection

Fig. 3.2 depicts our estimated distributions. The distributions of news and Google Trends query volume over domains are shown in the dotted (orange) and solid (black) lines, respectively. As is evident from the figure, both distributions depict a pattern of rapid decay.

We note for technical correctness that since the X-axis of Fig. 3.2 is not a well-defined variable, our use of the word "distribution" is intended only to convey the pattern of decay in the Y-axis,

but does not refer to a well-defined mathematical function of the domains placed on the X-axis.

The dashed (red) curve of Fig. 3.2 depicts the number of events (compiled by Event Registry) within each domain. The distribution matches most closely with a uniform distribution in terms of goodness of fit.

Whereas the news volume and Google Trends query volume distributions are rapidly decaying distributions with respect to the set of domains, the number of events in each domain remains (relative to the aforementioned two distributions) uniform. This fact establishes our pattern of systematic selection bias with decaying interest.

Fig. 3.2 delineates our three interest regions. The primary region features domains considered to be of existential interest, such as war and political activity. Our secondary region of interest includes domains such as Nutrition, Economy, Real Estate, and so on.

Beyond our conservative claim that the domains that inhabit each region are the same, we point out that the order of the domains in our news and Google Trends query volume distributions show substantial similarity. This is illustrated in Fig. 3.3, which shows the two distributions, both ordered according to the domain ordering of the news distribution. The fact that the Google Trends responses distribution is still rapidly decaying up to the secondary region establishes this similarity between the two, and supports our hypothesis of rapidly decaying attention across the regions.

The dotted (blue) distribution depicts the probability of a domain being prominently featured by in news. Consistent with our hypothesis of the three interest regions, this prominence distribution features three distinct probability regions. Counter-intuitively, we find that prominence increases as we go from primary to tertiary regions, and that domains from our tertiary region tend on average to be featured most prominently.

Whereas increased prominence for tertiary region domains may appear surprising at first sight, it is in fact consistent with our hypothesis of selection bias. Since news portals select many more events from the primary region to report as news, it is correspondingly harder for an event from the tertiary region to be selected. Events from the tertiary region that satisfy this threshold are therefore correspondingly more impactful, and are thus displayed more prominently.

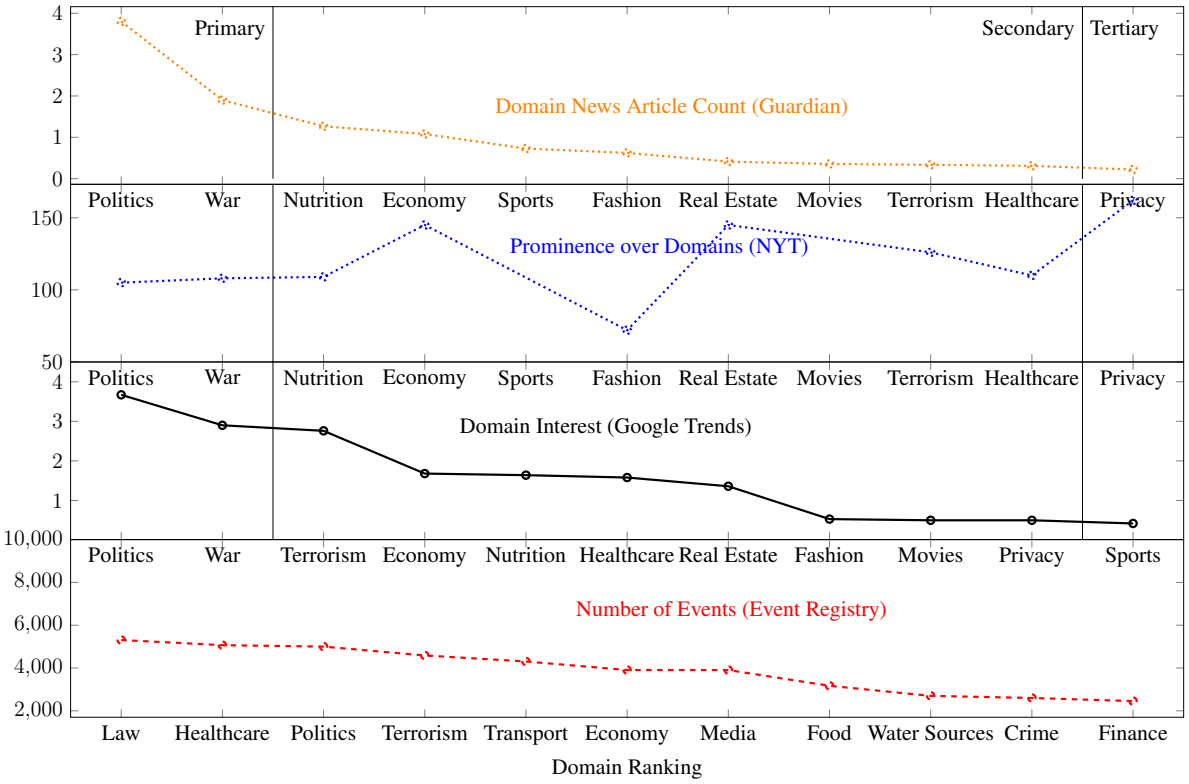


Fig. 3.2: Overall likelihood distributions computed over our period of interest of News (above) and Public Interest as measured by Google Trends (below) over selected domains.

3.3.2 Prominence

We summarize our findings and use them together to construct a predictive model of article prominence. Earlier, we noted that prominence increases from left to right on our volume-domain distribution. Further, we note that within a domain, prominence increases with domain volume as well as with concentration.

Variation of Prominence with Sentiment

Figs. 3.8 and 3.9 depict the variation of prominence with sentiment. Contrary to existing views Soroka et al. (2015), we find that positive sentiment news is more likely to be prominently featured than news with negative sentiment. We note, however, that since Soroka et al. (2015) restrict themselves to the study of news from a specific domain (politics), their findings cannot be generalized to the broad range of news domains examined in this study, and our findings are

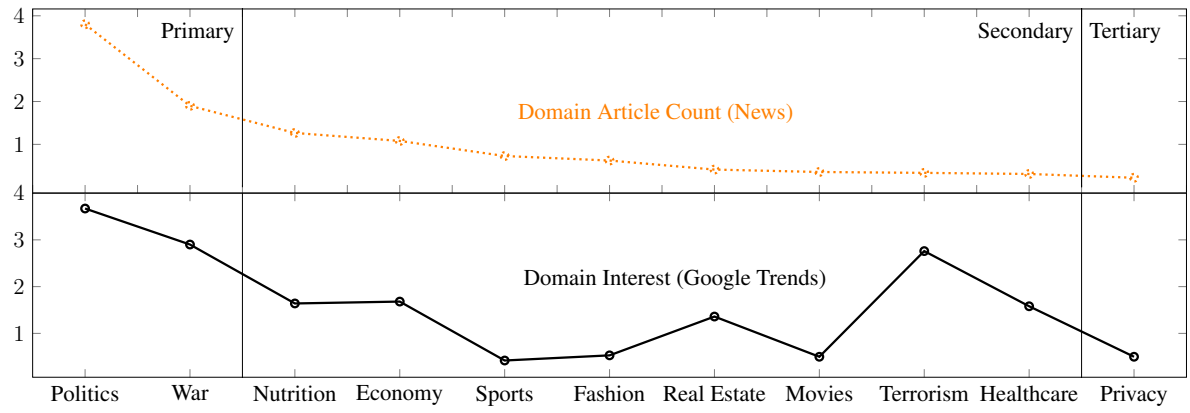


Fig. 3.3: Our News and Google Trends response distributions shown with a common domain order (taken from the news distribution of Fig. 3.2). Note that the rapidly decaying nature of the Google Trends distribution is preserved up to the Secondary region, despite the rearrangement of domains.

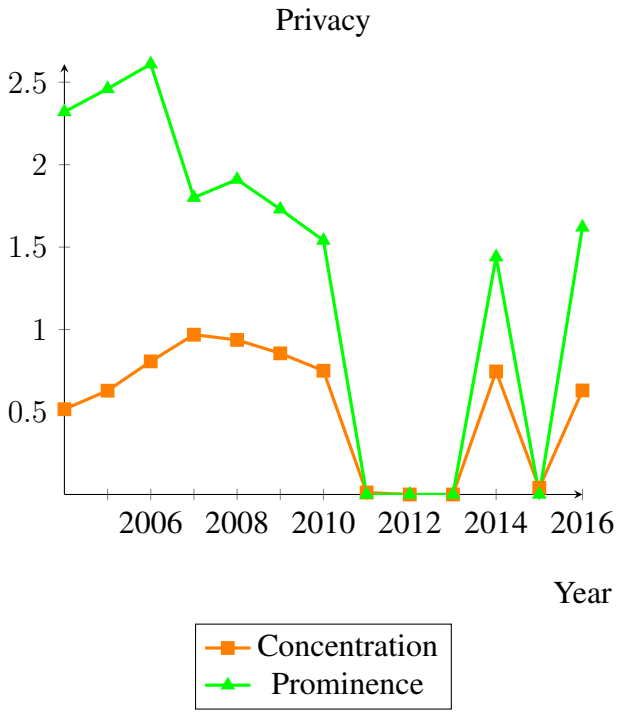
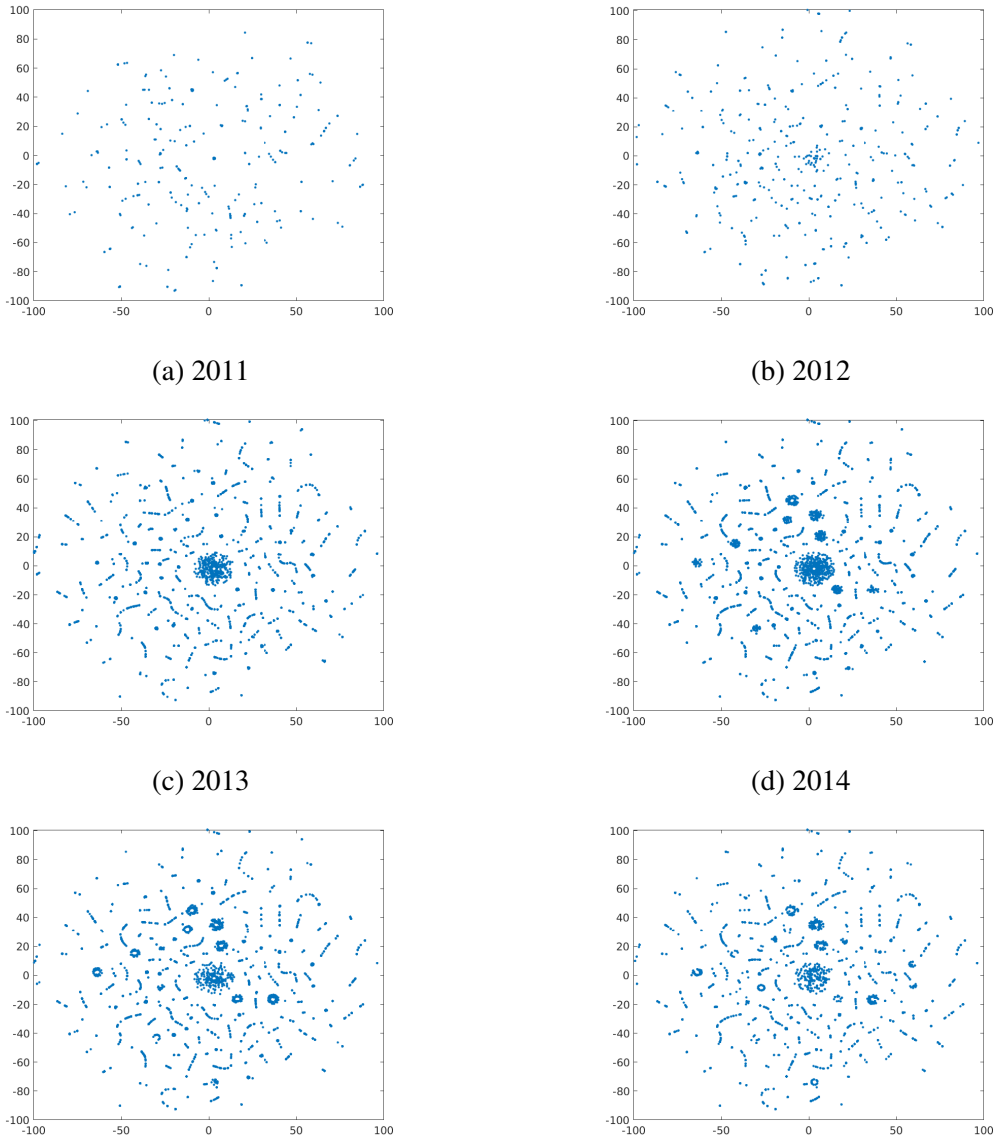


Fig. 3.4: Within a domain (*Privacy* shown here), prominence and concentration can increase and decrease together.

therefore not contradictory to theirs.

We show that sentiment polarity on average steadily becomes more positive as prominence



(a) 2011

(b) 2012

(c) 2013

(d) 2014

(e) 2015

(f) 2016

Fig. 3.5: *Surveillance* news from 2011 to 2016. We project our learned semantic space to two dimensions using t-SNE van der Maaten and Hinton (2008). Note that the Snowden revelations of 2013 coincide with the development of dense clusters in the noun corpus beginning in 2013, simultaneously with increasing prominence of Surveillance news (see Fig. 3.6).

decreases (see Fig. 3.9). We note, however, that this observation does not suggest that negative articles are never featured prominently. As Fig. 3.8 shows, negative sentiment articles do appear even on page one.

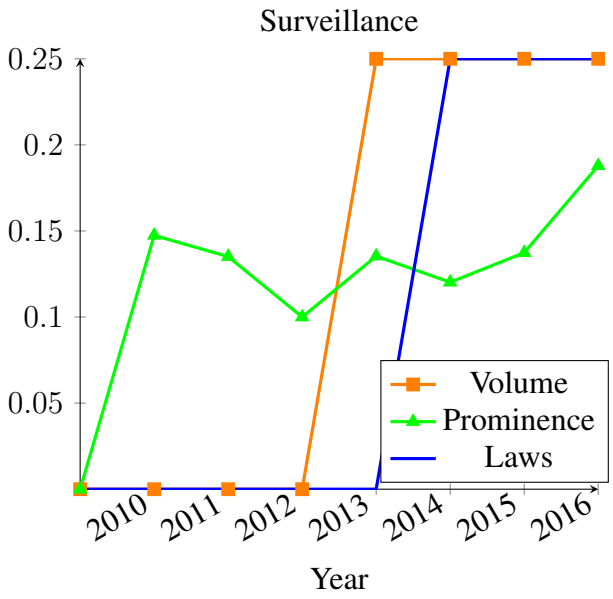


Fig. 3.6: The figure illustrates two distinct Granger causal relationships. Firstly, within a domain, prominence and news volume Granger cause each other. Secondly, prominence can Granger cause federal legislation.

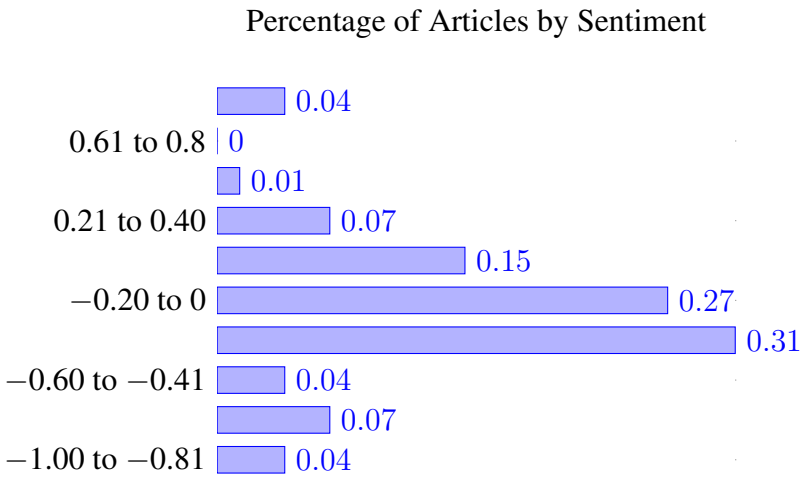


Fig. 3.7: The likelihood of an article being prominently featured versus sentiment polarity. We bin sentiment polarity into ten nonoverlapping divisions. Notice that articles with an extreme sentiment tend not to be featured prominently, and that positive articles are featured more prominently than negative ones.

However, we find that similar to the intuition that a tertiary domain event must be correspondingly more impactful than a primary domain event to be selected as news, a negative sentiment

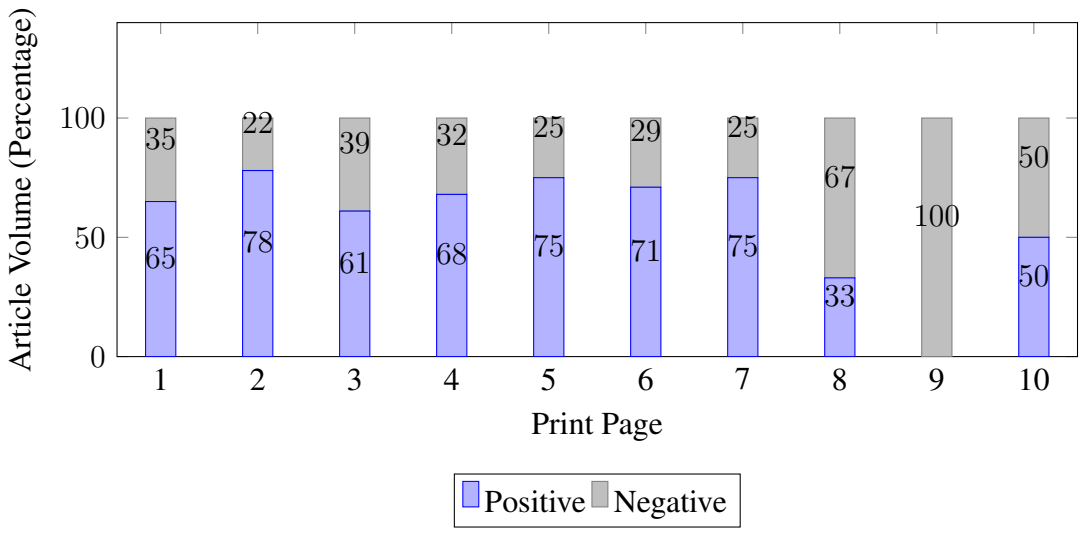


Fig. 3.8: The percentage of positive and negative sentiment articles per print page. Note that positive articles are twice as likely as negative ones to appear on page one. For this experiment, we assume that articles with polarity in $[-1,0)$ are negative and those with polarity in $(0,1]$ are positive.

article must be correspondingly more negative (than a positive sentiment article needs to be positive) to be prominently featured. Fig. 3.7 shows the distribution of prominence over sentiment. Notice that articles with an extreme sentiment, whether positive or negative, tend not to be featured prominently, and that positive articles are featured more prominently than negative ones.

Variation of Prominence with Article Type and Word Count

We now describe other factors that influence prominence. We find, as shown in Table 3.3, that the type of content an article contains influences how prominently it is featured. Surprisingly, we find that reviews (e.g., of movies, products, and literature) are twice as likely as news articles to be featured on page one.

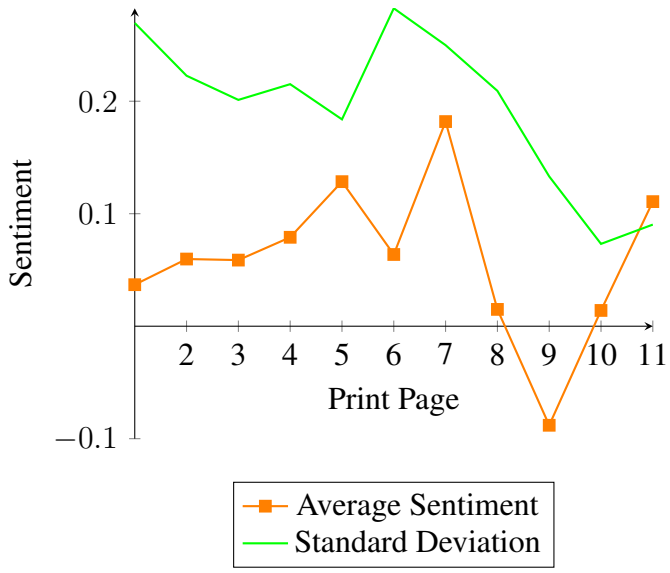


Fig. 3.9: Average sentiment per print page, on our scale of -1 to 1. Note that positivity generally increases with page number.

Table 3.3: Prominence by article type. The table shows the probability that an article of a particular type appears on pages 1 to 10. Prominent pages (1 and 10) are shown in bold. Note that reviews are twice as likely as news to appear on page one.

	Print Page									
Article Type	**Page 1**	Page 2	Page 3	Page 4	Page 5	Page 6	Page 7	Page 8	Page 9	**Page 10**
Reviews	**0.40**	0.10	0.10	0	0.20	0.10	0.10	0	0	0
News	**0.21**	0.15	0.12	0.08	0.05	0.05	0.02	0.02	0.01	**0.02**
Questions	0	0.50	0	0	0	0	0	0.25	0	0
Obituaries	0	0	0	0	0	0	0.43	0	0	0
Editorials	0	0	0	0	0	0	0	0	0	0
Summaries	0	1	0	0	0	0	0	0	0	0
Opinion Pieces	0	0	0	0	0	0	0	0	0	0
Letters	0	0	0	0.17	0	0	0	0	0	0

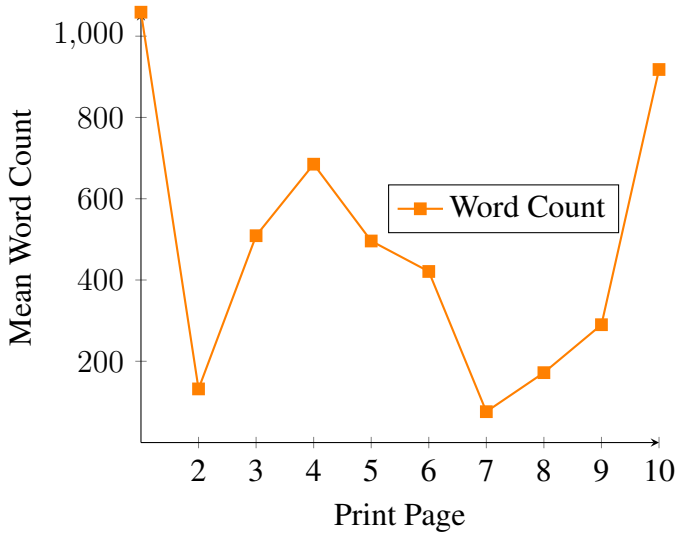


Fig. 3.10: Mean Word Count per print page, on our scale of -1 to 1. Note that positivity increases with page number.

Further, we find that prominence is correlated with an article's word count. Fig. 3.10 shows the variation of prominence with word count. As is apparent from the figure, prominently featured articles (featured on the first and last pages) also tend to have the largest word counts.

To rule out the competing hypothesis that an article's small word count may refer only to a snippet in a print newspaper or a link in an online article that is then continued on a subsequent page, we use the full word count of the article including the component that is continued on the later print or web page.

Prominence as a Granger Cause of Legislation

Finally, we show that just as news volume and similarity do, news prominence can Granger cause legislation. Fig. 3.6 shows a compelling example from the domain *Surveillance*. Surveillance prominence is depicted from 2009 to 2016, together with a binary time series depicting federal legislative activity in this domain. The USA Freedom Act was promulgated in 2013 in the wake of the Snowden revelations and was passed in 2015. We find a significant Granger causal F-statistic of 9.04 with a critical value 6.60 for this domain, at the $\alpha = 0.05$ level.

We found examples from other domains such as *drones* and *privacy* in which prominence correlates with federal legislative activity. However, we were unable to establish Granger causality

in these domains due to insufficient prominence data. We believe that we would find Granger causality in these domains, were more data to be available.

Predictive Model

Using these combined insights, we construct a supervised model of news prominence using the approach described in the Materials and Methods section (Section 3.2). We evaluate our model using ten train-test splits of 80% and 20% each, from our universal set with the prominence score as the dependent variable. Fig. 3.11 depicts the returned Receiver-Operating Characteristics curve, from which we obtain an Area Under the Curve (AUC) of 0.81.

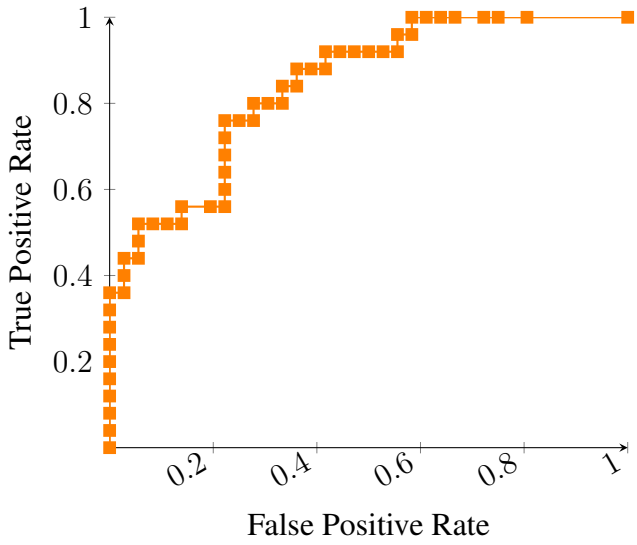


Fig. 3.11: Receiver-Operating Characteristics of our Prominence Model. We achieve an AUC of 0.81 on unseen data.

3.4 Discussion and Limitations

Taken together with existing understanding, our work constitutes progress toward a comprehensive and unified predictive model of public reaction to news. Earlier work Sheshadri and Singh (2019) discussed how public reaction tends to occur after periods of hyper-concentration,

and suggested how such periods may come about. The present paper enriches that understanding by quantifying the likelihood of such a period arising in a given domain (based on which of the three interest regions it belongs to), and, having arisen, the nature of the reaction it is likely to elicit.

The ultimate independent variables in news publishing are the events themselves, and how influential the events objectively are. The occurrence of events being a truly arbitrary process, this process is impossible to model. The relevant literature identifies selection and framing as the two primary subjective aspects of news Entman (1993). The present paper and Bourgeois et al. (2018) present insights into selection, whereas Sheshadri and Singh provide a framework that measures and explains framing.

We found two domains (out of twelve, representing about 16%) that represent counter-examples to our proposed interest regions. As can be seen from Fig. 3.2, the domain *sports* is in both the secondary news region and the tertiary domain interest region. The domain *privacy* is in the secondary domain interest region. However, *privacy* misses the tertiary region by only 2% of the overall distribution, and is clearly in the tertiary news and prominence regions. Given the uniformly consistent trends in all other domains, we conjecture that these exceptions may be due to imprecision in measurement, rather than counter-examples to our framework itself.

For uniform coverage across domains, our prominence distribution was computed over the year 2015. Based on incomplete data from other years, we believe that our hypotheses remain valid over our larger period of interest.

There is considerable scope for future research on this topic. Although we conjecture that the concepts we describe hold for all human collectives, we lack the data to evaluate them beyond the geographical regions of the US and UK. Whereas data sources such as GDELT Leetaru and Schrodt (2013) and Event Registry Leban et al. (2014) systematically compile events from all geographical regions, the data they provide is usually not rich enough for us to compute the news characteristics our model needs. Whereas our approach assumes that prominence is inversely proportional to an article's print page, prominence in general is more nuanced, being affected by factors such as the size of graphics or pictures in the article, the duration that an article stays on the homepage of an online news portal, and so on. Whereas our approach would benefit from the incorporation of these features, our APIs do not provide access to them. Additionally, our measure of news sentiment is at the level of individual articles and not at the level of individual aspects. Finally, we would like to replicate all aspects of our analysis on social media data, but lack the data sources to achieve credible coverage.

3.4.1 Variation of Agreement between Distributions with API

The news volume distribution of Fig. 3.2 was retrieved over our period of interest (2004 to 2018) from The Guardian. To confirm that the correlation obtained between our news and Google Trends query volume distributions is not API specific, we conduct a similar analysis with the NYT. We find that in the case of the Guardian, seven out of eleven (64%) domains in the news volume distribution are in positions that are within two spots of their corresponding position in the Google Trends query volume distribution. Four out of eleven positions are exactly the same.

For the NYT, we similarly find that five out of eight (63%) of domains are within two spots of their corresponding position in the Google Trends query volume distribution.

Whereas our data from both APIs support our hypothesis, we acknowledge that our formulation may undergo enhancements and modifications with the addition of data from sources with broader temporal and geographical ranges.

We are unable to evaluate our hypotheses beyond the data available to us, but motivate our work as an initial analysis of a hitherto unexplored area.

Chapter 4

Detecting Framing Changes in Topical News Publishing

This chapter addresses our third research question, RQ_3. The chapter describes our design and implementation of a fully unsupervised, data-driven approach to the automatic detection of framing changes in news data. It is based on a paper of the same title in review at *PLOS One*.

> *"For nearly four decades, health and fitness experts have* ***prodded and cajoled*** *and used other powers of persuasion in a* ***futile attempt to whip*** *America's youngsters into shape."*
>
> The New York Times, 1995

> *"The New York City Health Department has embarked on a new campaign to* ***persuade processed food brands to decrease sugar content*** *in a bid to curb obesity.*
>
> The New York Times, 2015

4.1 Introduction and Contributions

To motivate the problem and approach of this paper, let us investigate the primary causes of obesity in America. Public opinion and behavior on the subject have changed measurably since the late 1990s. As an example, Gunnars (2015) compiled a list in 2015 of ten leading causes, six of which suggest that the processed food industry is primarily responsible. By contrast, in the late 1990s, the public held Kim and Willis (2007); Flegal et al. (2012) that obesity was primarily caused by individual negligence and lifestyle choices. What led to this change in public opinion?

We posit that news publishing on the subject of *obesity* contributed to the change in the public's opinion. The above quotes from the New York Times (NYT) are representative snippets from news articles on *obesity* published in 1995 and 2015, respectively. Whereas both address the same topic, the 1995 snippet implies responsibility on the part of individuals, and the 2015 snippet implies responsibility on the part of the processed food industry. These subjective biases in news are collectively referred to as *framing*.

Framing theory Chong and Druckman (2007); de Vreese (2005) suggests that how a topic is presented to the audience (called "the frame") influences the choices people make about how to process that information. The central premise of the theory is that since an issue can be viewed from a variety of perspectives and be construed as having varying implications, the manner in which it is presented influences public reaction.

In general, understanding news framing can be a crucial component of decision-support in a corporate and regulatory setting. To illustrate this fact, we present a real-life example of the influence of framing on public perception and legislation. In 2011, security vulnerabilities in Facebook's use of HTML5 allowed third-party applications to steal personal data from approximately 59 million users Crossley (2011). The framing of news on the topic "Markup Languages" changed from a neutral narrative to one focusing on personal privacy. Revenues of Merix Games, Wimi5, and other HTML product vendors declined to the tune of four million dollars (over all companies) over the course of 2012 Fitzsimmons (2014). We conjecture that the decline was caused by negative coverage of news about HTML5 following the data leak.

Further, in 2013, the Personal Data Protection and Breach Accountability Act was promulgated in Congress US Congress (2014), under which Facebook was sued Fitzsimmons (2014). These

examples motivate the problem of *framing change detection*, which involves identifying when the dominant frame (or frames) Benford and Snow (2000) of a topic undergoes a change.

Contributions

This paper contributes a fully unsupervised and data-driven natural language based approach to detecting framing change trends over several years in domain news publishing. To the best of our knowledge, this paper is the first to address framing change detection, a problem of significant public and legislative import. Our approach agrees with and extends the results of earlier manual surveys, which required human data collection and were consequently limited in scope. Our approach removes this restriction by being fully automated. Our method can thus be run simultaneously over all news domains, limited only by the availability of real-time news data. Further, we show that our approach yields results that foreshadow periods of legislative activity. This motivates the predictive utility of our method for legislative activity, a problem of significant import.

Further, we contribute a collection of over 12,000 news articles from seven news topics or *domains*. In four of these domains, framing has been shown to change by surveys carried out in earlier research. In two domains, periods with significant legislative activity are considered. We collectively refer to these articles as the *Framing Changes Dataset*. Our individual domain datasets within the framing changes dataset cover the years in which earlier research found framing changes, as well as periods ranging up to ten years before and after the change. Our data are the first to enable computational modeling of framing change trends. We plan to release the dataset with our paper.

4.2 Materials and Methods

This section describes our datasets, data sources, and inter-annotator agreement.

4.2.1 Data Sources

We use two Application Programming Interfaces (APIs) to create our datasets.

The New York Times API:

The New York Times (NYT) Developer's API NYT (2016) provides access to news data from the NYT, Reuters, and Associated Press (AP) newspapers (both print and online versions) beginning from 1985. The NYT has the second largest circulation of any newspaper in the United States Wikipedia (2001).

The data object returned by the API includes fields such as the article type (news, reviews, summaries, and so on), the news source (NYT, Reuters, or AP), the article's word count, the date of its publication, article text (in the form of the abstract, the lead (first) paragraph, and a summary).

The Guardian API:

The Guardian Xplore API The Guardian (2016) provides access to news data from The Guardian, a prominent UK newspaper which reaches 23 million UK adults per month Wikipedia (2002).

The Guardian API returns full-length articles along with such metadata as the article type (similar to the NYT API) and a general section name (such as sports, politics, and so on). Although these section names are manually annotated by humans, we do not use them in our analysis, but rely instead on a simple term search procedure (see the Domain Dataset Generation section (Section 4.2.2)) to annotate our datasets.

4.2.2 Domain Dataset Generation

As in earlier work King et al. (2017); Sheshadri et al. (2017); Sheshadri and Singh (2019), we use a standard term search procedure to create our datasets. Specifically, an article belongs to a domain if at least a component of the article discusses a topic that is directly relevant to the domain Sheshadri and Singh (2019). For each domain, our APIs were used to extract news data during the time period b (denoting the beginning) to e (denoting the end), of the period of interest.

4.2.3 Inter-Annotator Agreement

A random sample of articles from each domain dataset was coded by two raters. We supply the per domain accuracy and inter-annotator agreement as Cohen's Kappa for sample domains in Table 4.1.

Table 4.1: Inter-Annotator agreement as Cohen's Kappa.

Domain	Dataset Accuracy Coder 1	Coder 2	Kappa
Surveillance	0.80	0.75	0.79
Smoking	0.84	0.82	0.93
Obesity	0.78	0.74	0.67
LGBT Rights	0.83	0.74	0.64
Abortion	0.80	0.80	0.50

4.2.4 Probability Distribution over Adjectives

Our approach relies on the key intuition that during a framing change, the valence of the adjectives describing co-occurring nouns changes significantly.

To measure this change, we create a reference probability distribution of adjectives based on the frequency of their occurrence in benchmark sentiment datasets.

Benchmark Datasets

We identified three open source benchmark review datasets from which to create our adjective probability distribution. Together, these datasets provide about 150 million reviews of various restaurants, services and products, with each review rated from one to five. Given the large volume of reviews from different sources made available by these datasets, we assume that they provide a sufficiently realistic representation of all adjectives in the English language.

We rely primarily on the Trip Advisor dataset to create our adjective probability distribution. We identified two other benchmark datasets, namely, the Yelp Challenge dataset and the Ama-

zon review dataset. Due to the fact that these datasets together comprise about 150 million reviews, it is computationally infeasible for us to include them in our learning procedure. Instead, we learnt distributions from these datasets for sample adjectives, to serve as a comparison with and verification of our overall learnt distribution. The resulting distributions for these adjectives appeared substantially similar to those of the corresponding adjectives in our learnt distribution. We therefore conclude that our learnt distribution provides a valid representation of all adjectives in the English language. We describe each dataset below.

Trip Advisor

The Trip Advisor dataset consists of 236,000 hotel reviews. Each review provides text, an overall rating, and aspect specific ratings for the following seven aspects: Rooms, Cleanliness, Value, Service, Location, Checkin, and Business. We limit ourselves to using the overall rating of each review.

Yelp

The Yelp challenge dataset consists of approximately six million restaurant reviews. Each entry is stored as a JSON string with a unique user ID, check-in data, review text, and rating.

Amazon

The Amazon dataset provides approximately 143 million reviews from 24 product categories such as Books, Electronics, Movies, and so on. The dataset uses the JSON format and includes reviews comprising a rating, review text, and helpfulness votes. Additionally, the JSON string encodes product metadata such as a product description, category information, price, brand, and image features.

4.2.5 Polarity of Adjectives

For each adjective in the English language, we are interested in producing a probability distribution that describes the relative likelihood of the adjective appearing in a review whose rating is r. For our data, r ranges from one to five.

We began by compiling a set of reviews from the Trip Advisor dataset for each rating from one to five. We used the Stanford CoreNLP parser Socher et. al. (2013) to parse each of the five sets of reviews so obtained. We thus obtained sets of parses corresponding to each review set. From the set of resultant parses, we extracted all words that were assigned a part-of-speech of 'JJ' (adjective). Our search identified 454,281 unique adjectives.

For each unique adjective a, we counted the number of times it occurred in our set of parses corresponding to review ratings one to five. We denote this by N_i, with $1 \leq i \leq 5$. Our probability vector for adjective a is then $\{\frac{N_a^1}{S_a}, \frac{N_a^2}{S_a}, \ldots, \frac{N_a^5}{S_a}\}$ where $S_a = N_a^1 + N_a^2 + N_a^3 + N_a^4 + N_a^5$.

Additionally, we recorded the rarity of each adjective as $\frac{1}{S_a}$. This estimates a probability distribution P, with 454,281 rows and six columns.

Table 4.2 shows example entries from our learned probability distribution. As can be seen from the table, our learned distribution not only correctly encodes probabilities (the adjective 'great' has nearly 80% of its probability mass in the classes four and five, whereas the adjective 'horrible' has nearly 80% of its mass in classes one and two), but also implicitly learns an adjective ranking such as the one described in De Melo et al. Bansal (2013). To illustrate this ranking, consider that the adjective 'excellent' has 60% of its probability mass in class five, whereas the corresponding mass for the adjective 'good' is only 38%.

Motivated by our learned probability distribution, we posit that classes 1 represents negativity, class 2 to 4 represent neutrality, and class 5 represents positivity.

4.2.6 Incorporating Adjective Rarity

Our measure of adjective rarity serves as a method by which uncommon adjectives, which rarely occur in our benchmark dataset, and whose learnt probability distributions may hence be unreliable, can be excluded.

However, in doing so, we run the risk of excluding relevant adjectives from the analysis. We manually inspect the set of adjectives that describe the nouns in each domain to arrive at a domain specific threshold.

For a majority of our domains (five out of seven), we use a threshold of $q > -\infty$, that is, no adjectives are excluded. For the remaining two domains, (*drones* and *LGBT rights*), we

Table 4.2: Sample entries from our learned probability distribution for positive and negative sentiment adjectives.

Adjective	Class 1	Class 2	Class 3	Class 4	Class 5	Rarity (Inverse Scale)
Great	0.039	0.048	0.093	0.274	0.545	4.495e-07
Excellent	0.019	0.028	0.070	0.269	0.612	2.739e-06
Attractive	0.095	0.125	0.192	0.296	0.292	0.0001
Cute	0.039	0.068	0.155	0.330	0.407	1.499e-05
Compassionate	0.068	0.020	0.010	0.038	0.864	0.0004
Good	0.076	0.095	0.185	0.336	0.308	3.459e-07
Horrible	0.682	0.143	0.076	0.042	0.057	7.453e-06
Ridiculous	0.461	0.180	0.125	0.116	0.118	2.033e-05
Angry	0.546	0.138	0.092	0.098	0.126	6.955e-05
Stupid	0.484	0.136	0.099	0.117	0.164	5.364e-05
Beautiful	0.043	0.049	0.085	0.222	0.599	6.233e-06

employ a threshold of $q > 10^{-4}$.

The trends in our results appeared to be fairly consistent across a reasonable range of threshold values.

4.2.7 Domain Period of Interest

We define a period of interest for each domain. Let t_f be a year in which a documented framing change took place in the domain under consideration. Then, our period of interest for this domain is $b = \min(t_f - 10, t_f - l)$ to $e = \max(t_f + 10, t_f + r)$, where the API provides data up to l years before, and r years after t_f. All units are in years.

4.2.8 Corpus-Specific Representations

A domain corpus is a set of news articles from a given domain. Let a given domain have m years in its period of interest with annual domain corpora $T_1, T_2, \ldots T_m$.

4.2.9 Corpus Clustering

An overall domain corpus is therefore $T = T_1 \cup T_2 \ldots \cup T_m$.

We assume that a corpus has k unique frames. We adopt a standard topic modeling approach to estimate frames. We use the benchmark Latent Dirichlet Allocation (LDA) Blei et al. (2003) approach to model k=5 topics (or frames) in each domain corpus. We extract the top l=20 terms v from each frame. We also extract the set of all unique nouns in T. We define a cluster as the set of nouns $v \cap T$. We thus generate k clusters, each representing a unique frame.

4.2.10 Annual Cluster Polarity

For each cluster c, we are interested in arriving at a vector of annual polarities for each year i, $1 \leq i \leq m$ in the domain period of interest.

Let x_c be the set of all nouns in c. For each noun $v \in x_c$, we use the Stanford dependency parser Socher et al. (2013) to identify all adjectives (without removing duplicates) that describe v in T_i. We extract the polarity vectors for each of these adjectives from P as the matrix A_i. A_i therefore has n rows, one for each adjective so identified, and five columns (see the Polarity of Adjectives section (Section 4.2.5)). We estimate the annual cluster polarity of c as the vector of column-wise averages of A_i. Let $P_c = \{P_1, P_2, \ldots, P_m\}$ be the set of annual cluster polarities so obtained.

Annual polarities for representative clusters from each of our domains are shown in figures 4.10 to 4.16. For visibility, each subplot in figures 4.10 to 4.16 is shown on a distinct scale.

4.2.11 Detecting Framing Changes using Periods of Maximum Correlation

Our five polarity classes serve as measures of framing within a domain. We conceive of a framing change as a trend, consistent across our five polarity classes, over a period of years.

We describe our intuition and approach to detecting framing changes below. Firstly, we show

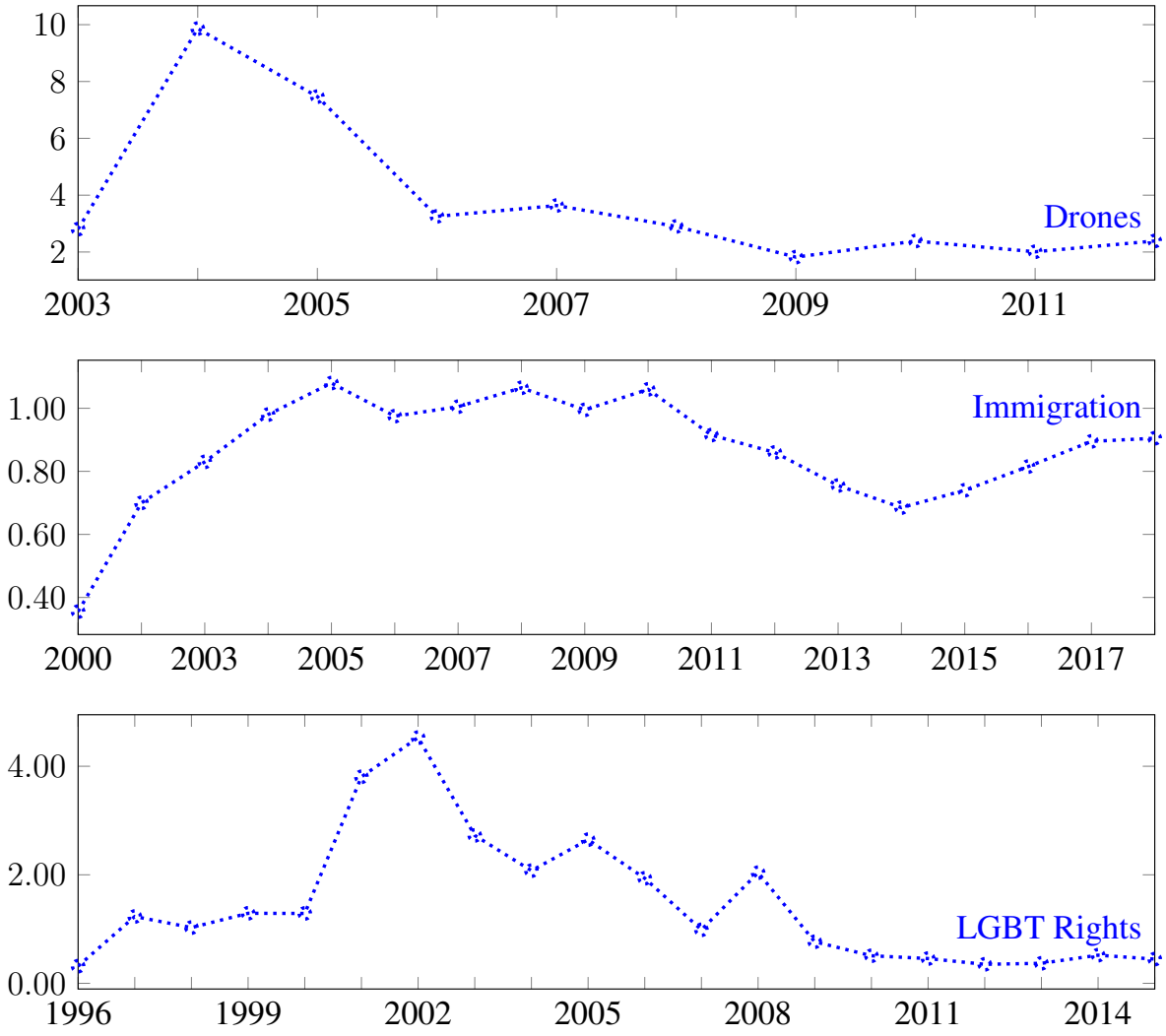


Fig. 4.1: The average number of adjectives per article, shown for our domains over their respective periods of interest. This measure serves as a measure of the subjectivity of news in a domain. Notice that in the domain *LGBT rights*, the peak in this measure immediately precedes a framing change from an earlier study Engel (2013).

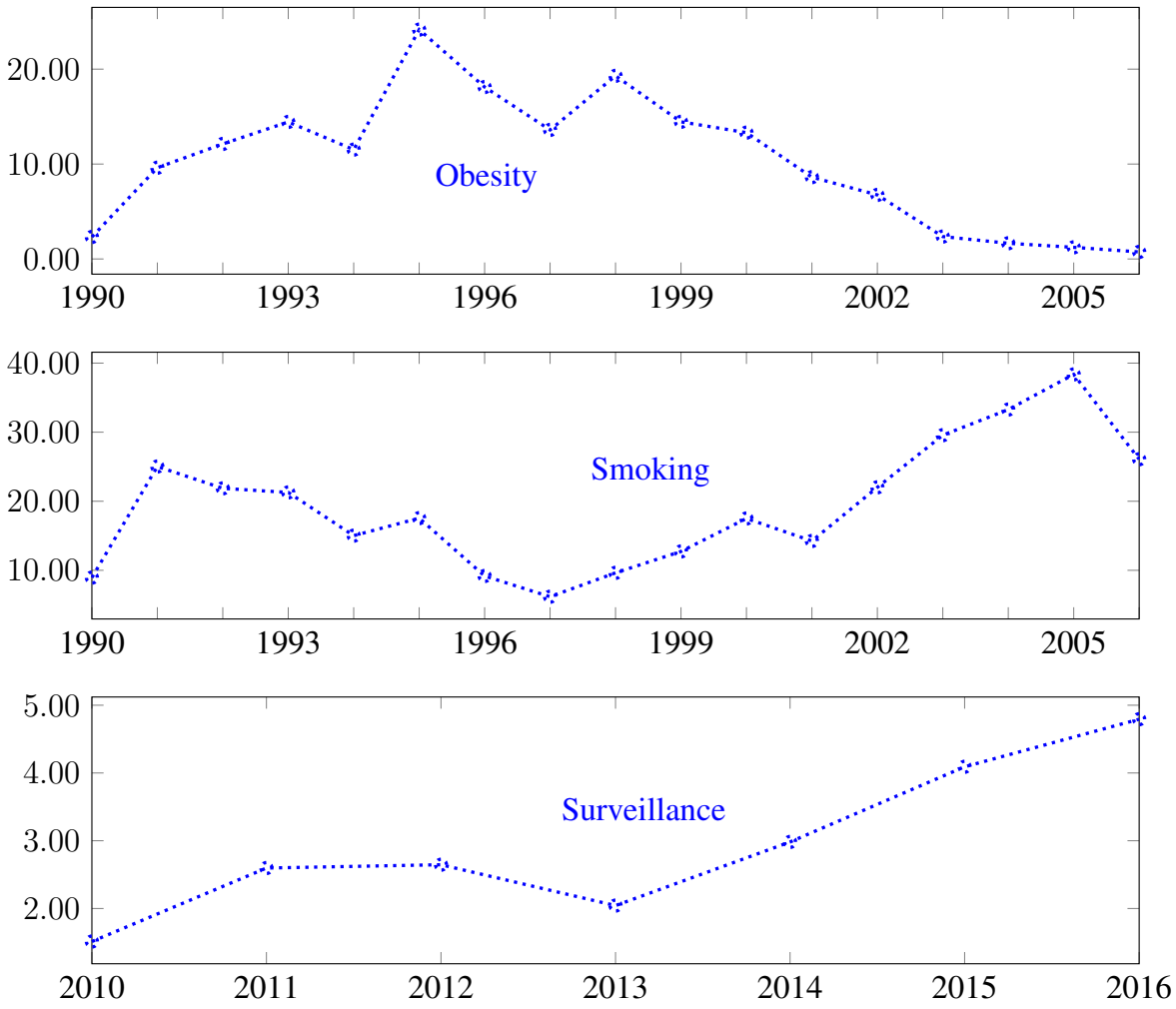


Fig. 4.2: The average number of adjectives per article, shown for our domains over their respective periods of interest. This measure serves as a measure of the subjectivity of news in a domain.

that the frequency with which adjectives occur in articles varies both by domain, and in different years within a domain.

Figure 4.1 depicts the average number of adjectives per article for each of our domains over the years in their respective periods of interest. We note that this count serves also as a measure of how subjective news publishing in a domain is, since adjectives are indicative of how events are framed.

Notice that in the domain *LGBT rights*, the peak in this measure immediately precedes a framing change from an earlier study Engel (2013).

Despite the fact that the volume of adjectives used per article vary dramatically (by up to 30%), we find that the variation in our annual cluster polarity between successive years is generally on the order of less than 1%. However, through a consistent trend lasting multiple years, our measure of annual polarity can change (increase or decrease) cumulatively by up to 5% (see figure 4.10 for an example). We refer to such a cumulative trend as a framing change.

We now consider the problem of fusing estimates from our five measures of annual cluster polarity. Consider the change in polarity of classes 1 , 3, 4, and 5 between 2005 and 2006 in figure 4.10, as against the change in class 2. Whereas classes 1, 3, 4, and 5 changed little, class 2 showed a substantial change.

In contrast, we note that in the period 2001 to 2013, a consistent trend was observable across all five classes, with substantial reductions in classes 2 and 3, and a notable corresponding inccrease in class 5. We exploit *correlations* between the changes in our five classes to identify framing changes.

Accordingly, we use Pearson correlation Benesty et al. (2009) between our classes as a measure of trend consistency. Let a given domain have m years in its period of interest. We generate all possible subsets of T_m, namely, T_{i-j}, where $i \leq j \leq m$, and T_{i-j} denotes the domain corpus from year i to year j.

Let $C = \{C_1, \ldots, C_5\}$ be the set of class vectors for this domain subset, where $C_1 = \begin{bmatrix} C_1^i \\ C_1^{i+1} \\ \cdots \\ C_1^j \end{bmatrix}$,

C_1^i is the value of class 1 for year i, and similarly for $C_2, \ldots, C_5$.

To measure the correlation of subset T_{i-j}, we compute its matrix of correlation coefficients

Mathworks (2019a) K. We reshape K into a vector of size $f \times 1$ where $f = i * j$, and evaluate its median, l. We find the maximum value of l, l_{max}, over all possible values of i and j. We denote the values of i and j corresponding to l_{max} as i_{max} and j_{max}. We return $T_{i_{max}-j_{max}}$ as our *period of maximum correlation (PMC)*.

We note that the smaller the duration of a PMC, the greater the possibility that our class vectors may have a high correlation in the period due to random chance. To compensate for this, we employ a threshold whereby a period is not considered as a candidate for the domain PMC unless it lasts at least k years. We uniformly employ a value of k=3 in this paper.

Sample correlations and our PMC for the domain *Immigration* are shown in Table 6.1.

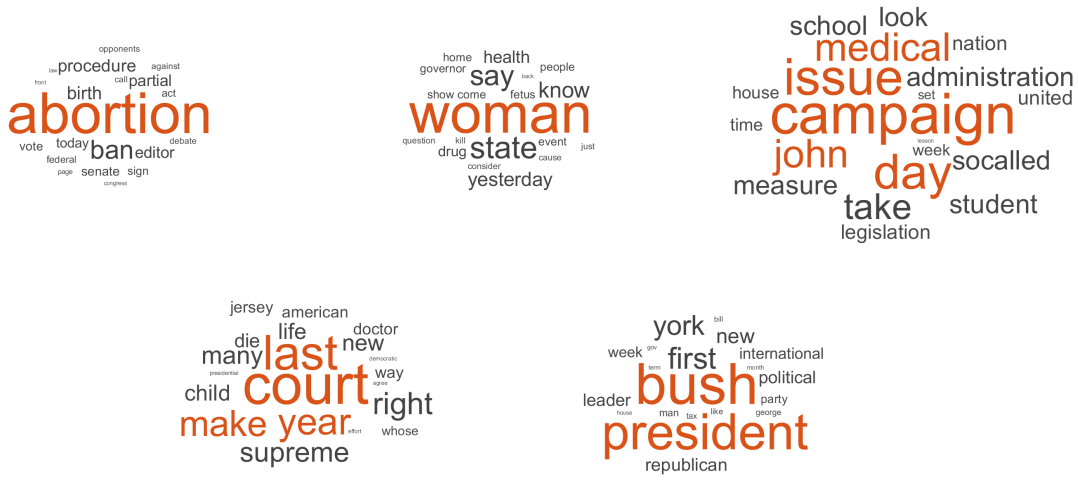


Fig. 4.3: Our estimated clusters for the domain *abortion*. Each cluster is said to represent a unique *frame*. The frame discussed in cluster 1 concerns a proposed ban on abortion. We analyze this cluster, and find that our estimated PMC (Figure 4.16) coincides with the period immediately preceding the Partial Birth Abortion Act of 2003.

4.3 Experiments and Results

We find that our periods of maximum correlation correlate substantially with framing changes described in earlier surveys Cummings and Proctor (2014); Engel (2013); Pew Research (2016);

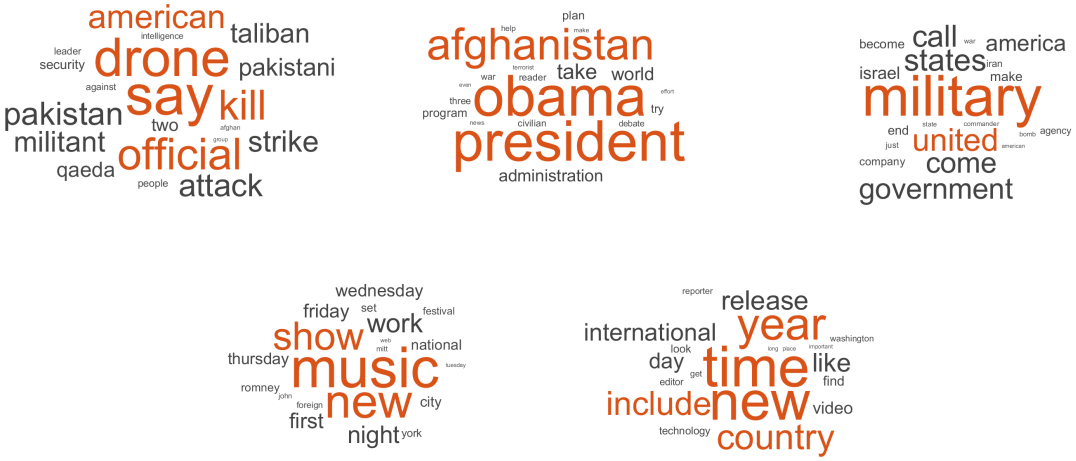


Fig. 4.4: Our estimated clusters for the domain *drones*. Each cluster is said to represent a unique *frame*. The frame discussed in cluster 1 concerns the use of drones against terrorist targets. Our analysis of this cluster returns a PMC of 2009 to 2011 (Figure 4.15). Our PMC immediately foreshadows the Federal Aviation Administration's Modernization and Reform Act of 2012.

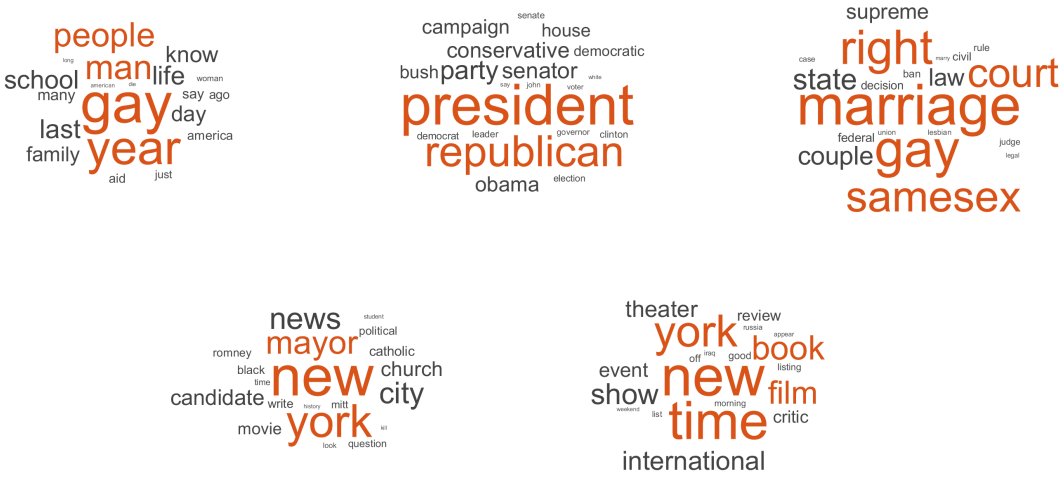


Fig. 4.5: Our estimated clusters for the domain *LGBT Rights*. Each cluster is said to represent a unique *frame*. The frame discussed in cluster 3 discusses the subject of same-sex marriage, and in particular, judicial activity on this topic. We analyze this cluster and estimate two PMCs of nearly identical correlation score (2006 to 2008 and 2013 to 2015 Figure 4.13). The PMC of 2013 to 2015 foreshadows and coincides exactly with the Supreme Court judgment of 2015 that legalized same-sex marriage in fifty states.

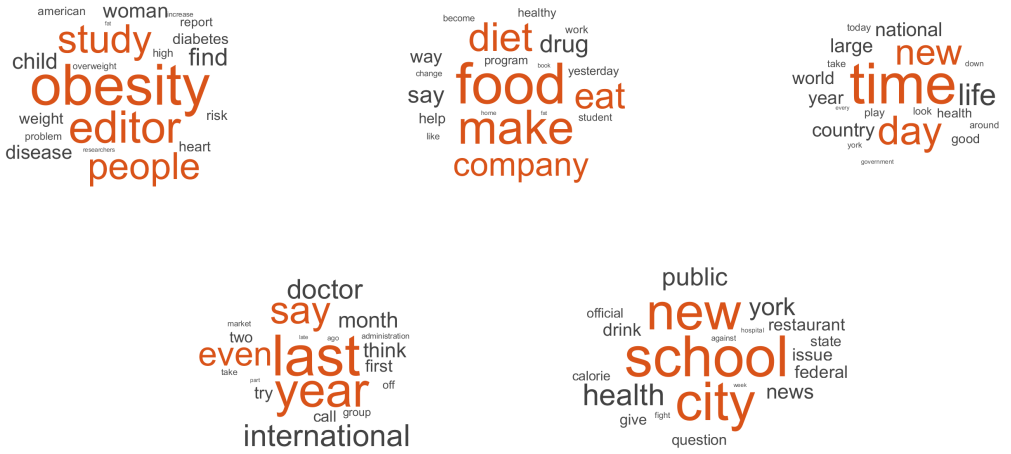


Fig. 4.6: Our estimated clusters for the domain *obesity*. Each cluster is said to represent a unique *frame*. We posit that cluster 2 represents societal causes of obesity (see the Obesity section). We analyze this cluster and estimate a PMC of 2005 to 2007 (Figure 4.12). Our PMC agrees with the findings of an earlier human survey Kim and Willis (2007).

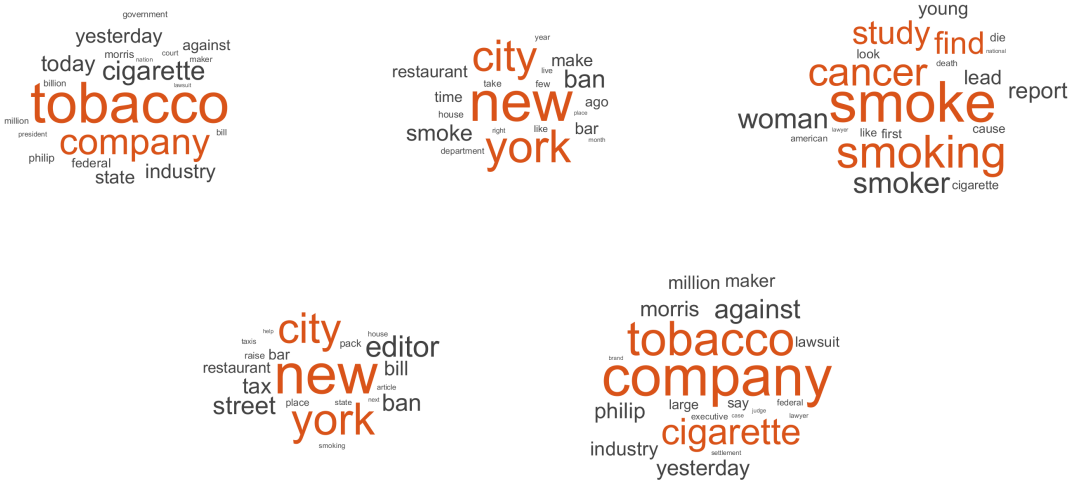


Fig. 4.7: Our estimated clusters for the domain *smoking*. Each cluster is said to represent a unique *frame*. The frame of cluster 3 discusses the health risks associated with smoking. We analyze this cluster and estimate a PMC of 2001 to 2003 (Figure 4.10). Our PMC coincides exactly with an earlier monograph from the The National Cancer Institute (NCI) that describes a progression towards tobacco control frames in American media between 2000 and 2003.

Fig. 4.8: Our estimated clusters for the domain *surveillance*. Each cluster is said to represent a unique *frame*. The frame of cluster 3 discusses the Snowden revelations of 2013. We analyze this cluster and estimate a PMC of 2013 to 2014 (Figure 4.11). Our PMC coincides exactly with the period following the Snowden revelations. Additionally, we note that the Columbia Journalism Review Vernon (2018) found that following the Snowden revelations, news coverage of Surveillance changed to a narrative focusing on individual rights and digital privacy.

Fig. 4.9: Our estimated clusters for the domain *Immigration*. Each cluster is said to represent a unique *frame*. The frame of cluster 2 discusses the waning of asylum, increased border refusals and the final 2002 white paper on "Secure Borders, Safe Haven". We analyze this cluster and estimate a PMC of 2000 to 2002 (Figure 4.14). Our PMC coincides exactly with the period immediately foreshadowing the government white paper.

Vanian (2015), and also foreshadow legislation.

Our computed class vectors are depicted in figures 4.10 to 4.16. We discuss each domain below.

4.3.1 Smoking

The National Cancer Institute (NCI) published a monograph discussing the influence of the news media on tobacco use National Cancer Institute (2019). On page 337, the monograph describes how, between 2000 and 2003, American news media progressed towards tobacco control frames. It states that 55% of articles in this period reported progress on tobacco control, whereas only 23% reported setbacks.

In contrast, the monograph finds (also on page 337) that between 1985 to 1996, tobacco control frames (11) were fairly well balanced with pro-tobacco frames (10). We extracted a dataset of over 2,000 articles from 1990 to 2007.

Our approach returns a PMC of 2001 to 2003 (see figure 4.10) for this domain, corresponding exactly with the period described in the monograph.

4.3.2 Surveillance

The Columbia Journalism Review Vernon (2018) found that following the Snowden revelations, news coverage of Surveillance in the US changed to a narrative focusing on individual rights and digital privacy. We compiled a dataset consisting of approximately 2,000 *surveillance* articles from the New York Times for the period 2010 to 2016.

Our class vectors for this domain are shown in figure 4.11. We obtain a PMC of 2013 to 2014 for this period, corresponding closely with the ground truth framing change.

The trends in our class vectors are indicative of the change. As can be seen from the figure, positivity (measured by class 5), drops markedly, together with a simultaneous increase in negativity (class 1) and neutrality (classes 2 and 3). Class 4 remains close to constant during this period and thus does not affect our hypothesis.

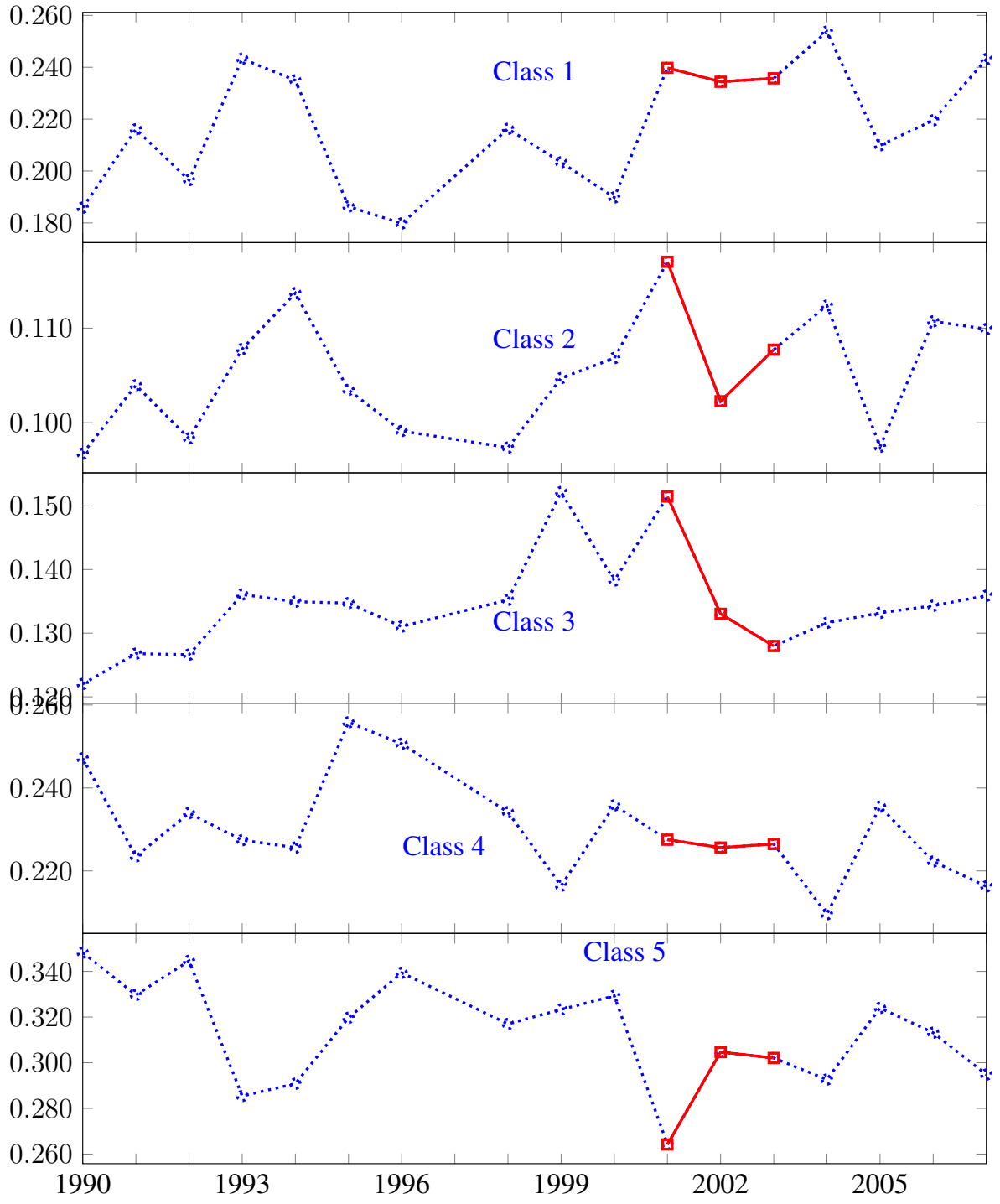


Fig. 4.10: Annual polarities for cluster 3 in Figure 4.7 from the domain *smoking* for the classes 1 to 5. The PMC is shown with solid lines in square markers, and coincides exactly with a framing change described in an earlier NCI monograph.

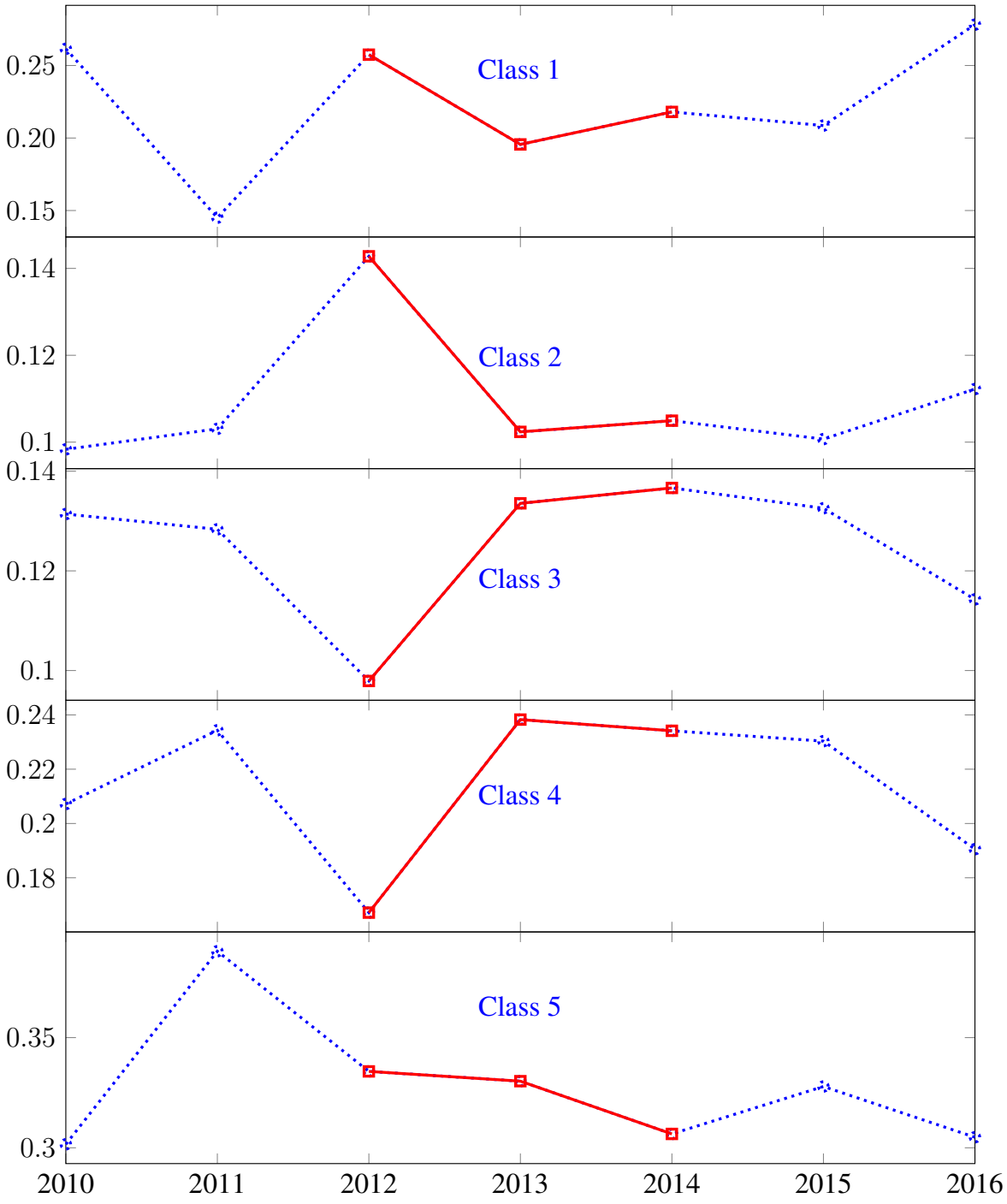


Fig. 4.11: Annual polarities for a representative cluster from the domain *surveillance* for the classes 1 to 5. The PMC is shown with solid lines in square markers.

4.3.3 Obesity

Kim and Willis (2007) found that the framing of *obesity* news underwent changes between the years 1997 and 2007. During this period, Kim and Willis found that the fraction of news articles attributing responsibility for obesity to societal causes increased significantly. Prior to this period, *obesity* was primarily framed as an issue of individual responsibility. For example, *obesity* news after the year 2000 has often criticized food chains for their excessive use of sugar in fast food, as shown in the NYT snippet in the Introduction and Contributions section (Section 4.1). We compiled a dataset of over 3,000 articles from the New York Times (since Kim and Willis restrict their study to Americans) from 1990 to 2009.

The clusters we estimate for this domain are shown in Figure 4.6. Cluster 2 addresses possible causes of obesity, with a particular focus on dietary habits. We posit that this cluster represents societal causes more than individual ones (since individual causes, as shown in the NYT snippet of the Introduction and Contributions section (Section 4.1) tend to discuss topics such as fitness and sedentary lifestyles, as opposed to food content). We observe that the PMC for this domain (2005 to 2007) is characterized by increased positivity, shown by classes 4 and 5, and decreased negativity (class 1). Our results for this domain thus agree with the findings of Kim and Willis (2007).

4.3.4 LGBT Rights

We compiled a dataset of over 3,000 articles from the period 1996 to 2015 in this domain. Figure 4.5 depicts our estimated clusters. Cluster 3 represents a frame that discusses the subject of same-sex marriage and its legality. We note that the Supreme Court ruled to legalize same-sex marriages in fifty states in the year 2015. Our class vectors for this domain are shown in figure 4.13. We obtained two PMCs with nearly identical correlation scores (0.999 for the period 2006 to 2008, and 0.989 for the period 2013 to 2015). Figure 4.13 highlights the period 2013 to 2015 foreshadowing the judicial action of 2015.

4.3.5 Immigration

We sudy the framing of *immigration* news in the United Kingdom. We obtained about 3,600 articles on the subject of Immigration from the Guardian API for the period 2000 to 2017. For

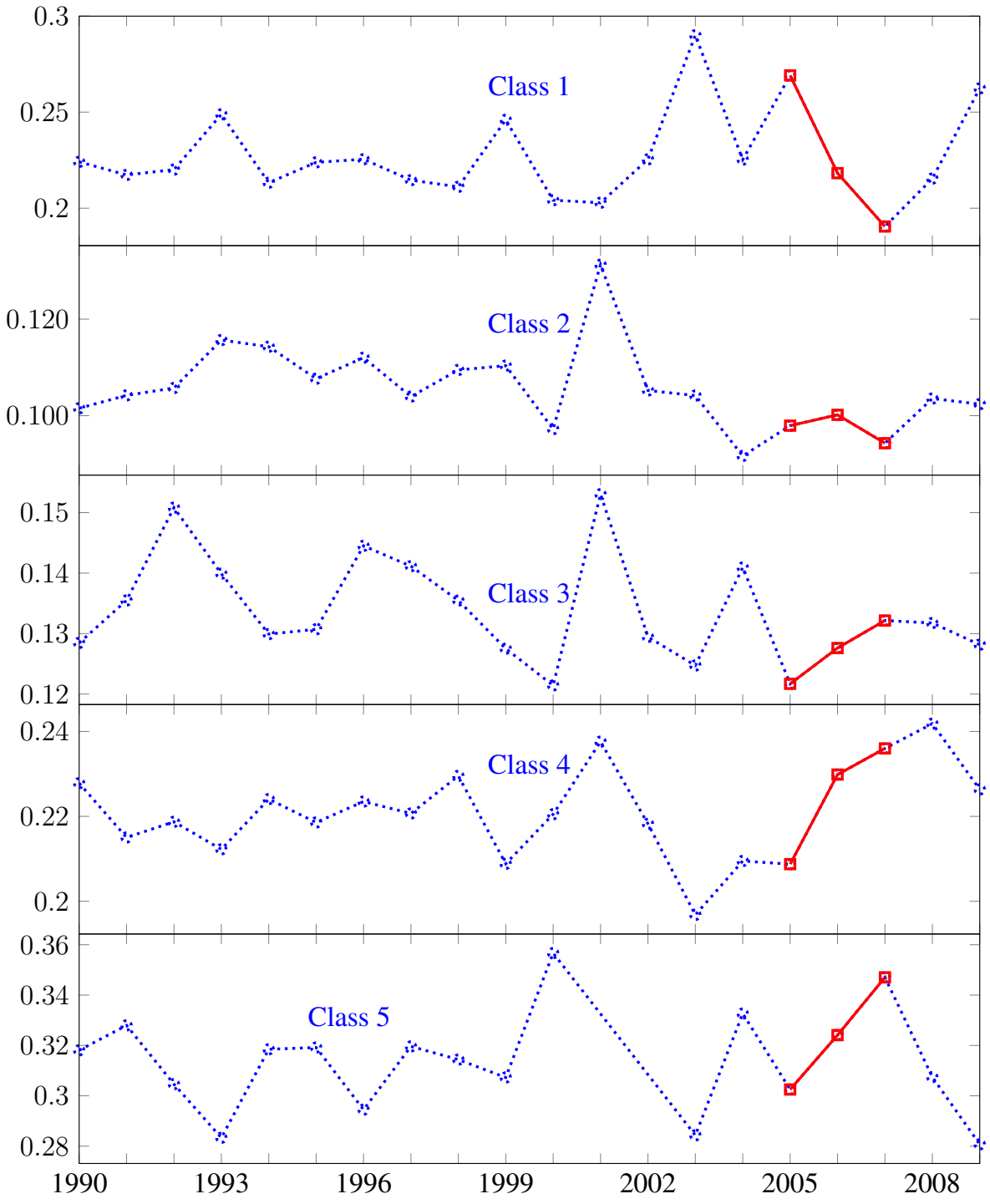


Fig. 4.12: Annual polarities for cluster 2 from Figure 4.6 from the domain *obesity* for the classes 1 to 5. The PMC is shown with solid lines in square markers. We posit that cluster 2 represents societal causes of obesity (see the Obesity section). We observe that the PMC for this cluster (2005 to 2007) agree with the findings of Kim and Willis (2007).

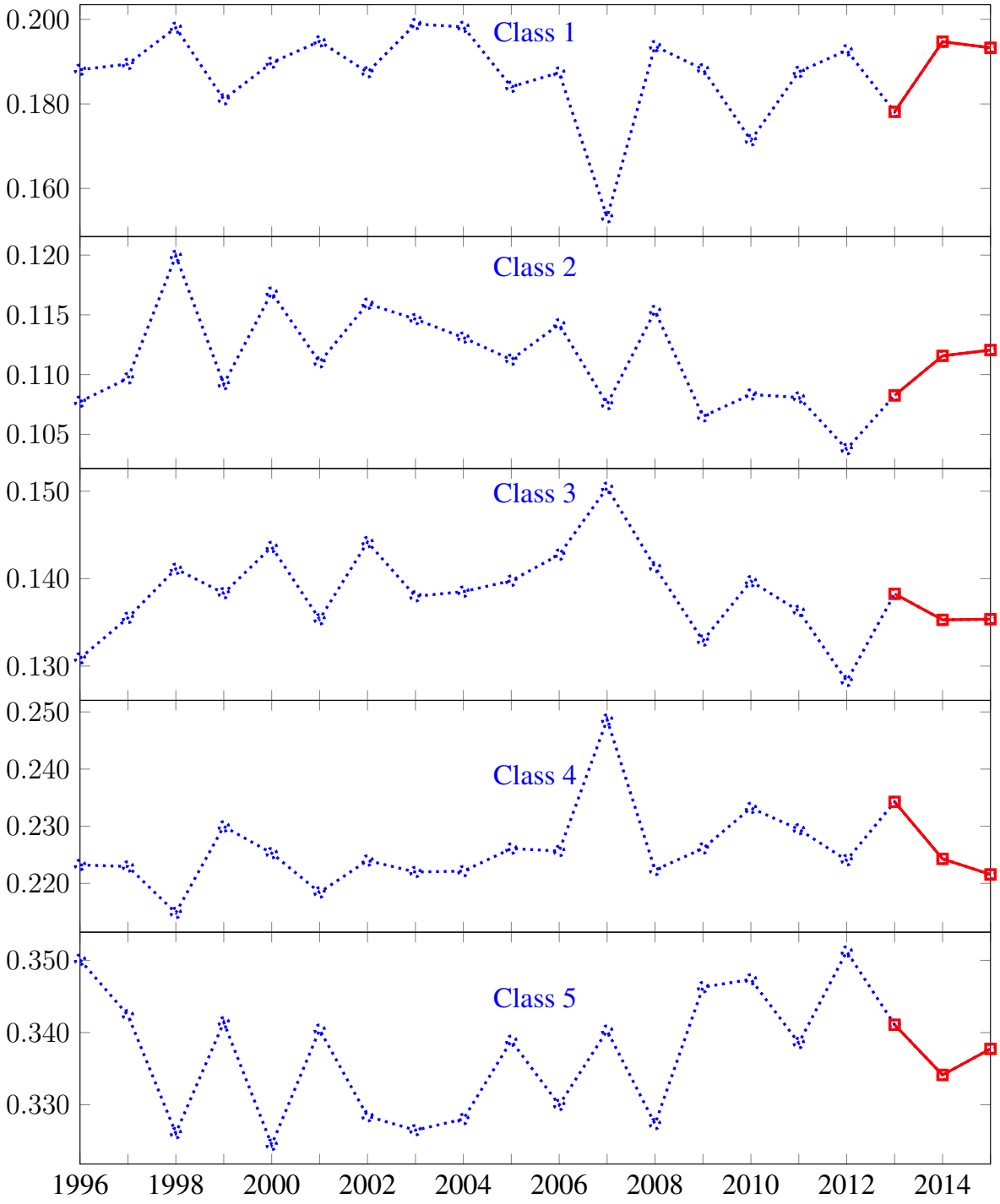


Fig. 4.13: Annual polarities for cluster 3 in Figure 4.5 from the domain *LGBT Rights* for the classes 1 to 5. We obtain two PMCS with nearly identical correlation scores, 2006 to 2008 and 2013 to 2015. The PMC of 2013 to 2015 is shown with solid lines in square markers, foreshadowing the judicial action of 2015.

this domain, we carried out our analysis on the article titles (rather than the full text). Since the Guardian returns full length articles, we found that this design choice allows us to produce a more focused domain corpus than the one generated by the full article text. We depict our estimated class vectors and PMC in figure 4.14.

We analyze the frame of cluster 2 in Figure 4.9. This cluster deals with the issue of asylum seekers to the United Kingdom. In the period beginning immediately before the year 2000,a new peak in asylum claims to the United Kingdom of 76,040 had been reached Wikipedia (2019). This event coincided with a high-profile terrorist act by a set of Afghan asylum seekers Wikipedia (2019).

These events resulted in increased border refusals and the final 2002 white paper on "Secure Borders, Safe Haven". We estimate a PMC of 2000 to 2002 (Figure 4.14). Our PMC coincides exactly with the period immediately foreshadowing the government white paper.

4.3.6 Drones

We obtained nearly 4,000 articles on this domain for the period 2003 to 2012. We obtain a PMC of 2009 to 2011 for this domain, as shown in Figure 4.15.

Our PMC immediately foreshadows the Federal Aviation Administration's Modernization and Reform Act of 2012.

The aforementioned two domains (*immigration* and *drones*) highlight the predictive utility of news framing. Whereas we did not find earlier surveys that coincide with our PMCs for these domains, we note that these PMCs foreshadowed substantial legislative activity. This observation suggests that PMCs estimated through real-time monitoring of domain news may yield predictive utility for legislative and commercial activity.

4.3.7 Abortion

The Partial-Birth Abortion Ban Act was enacted in 2003. We obtained 248 articles for the period 2000 to 2003, for this domain. We obtain a PMC of 2001 to 2003 for this domain, as shown in figure 4.16.

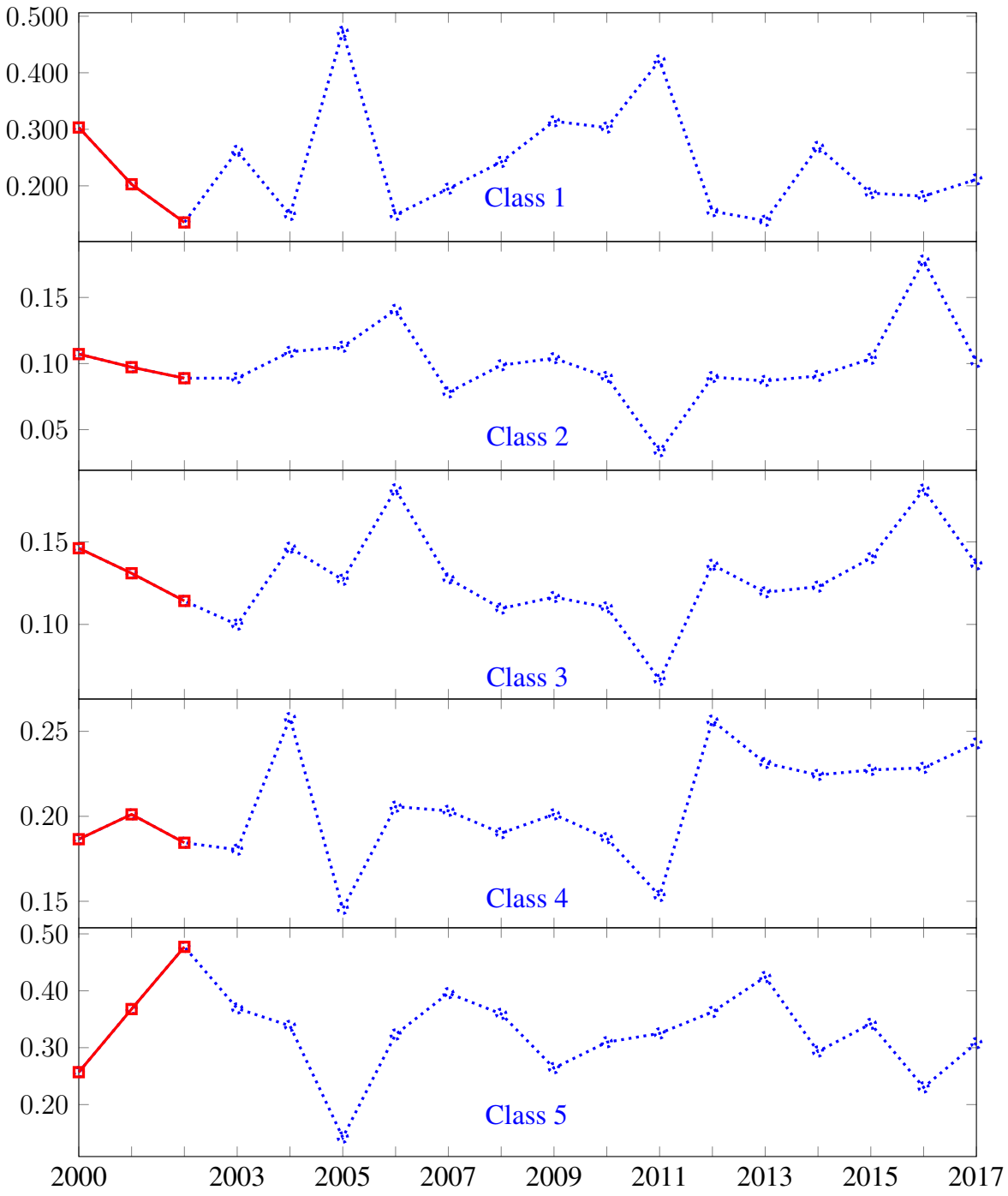


Fig. 4.14: Annual polarities for cluster 2 from Figure 4.9 from the domain *immigration* for the classes 1 to 5. The PMC is shown with solid lines in square markers, and foreshadows the "Secure Borders, Safe Haven" white paper of 2002.

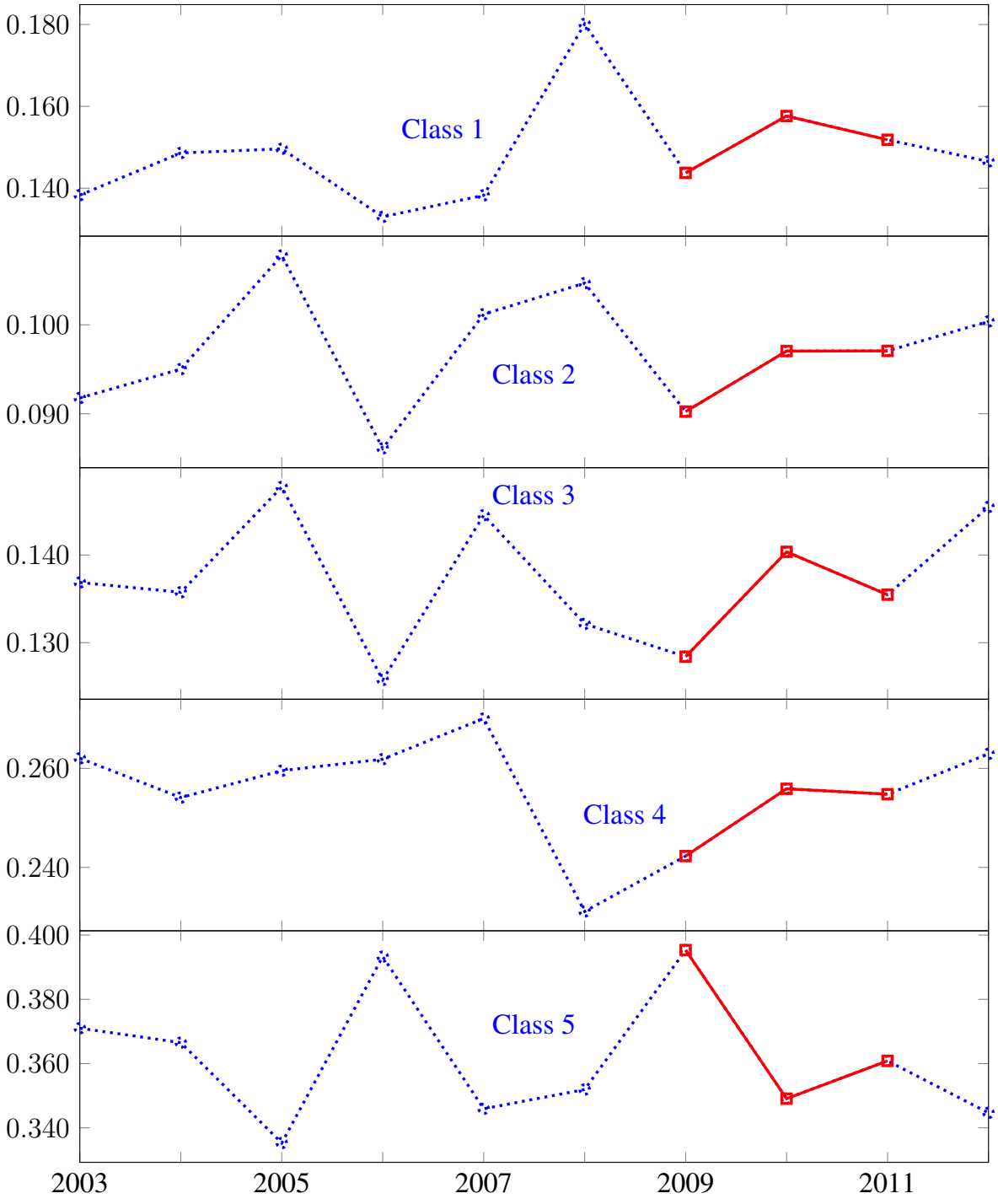


Fig. 4.15: Annual polarities for cluster 1 from Figure 4.4 from the domain *drones* for the classes 1 to 5. The PMC is shown with solid lines in square markers, and immediately foreshadows the Federal Aviation Administration's Modernization and Reform Act of 2012. This suggests the predictive utility of framing change detection for legislative activity.

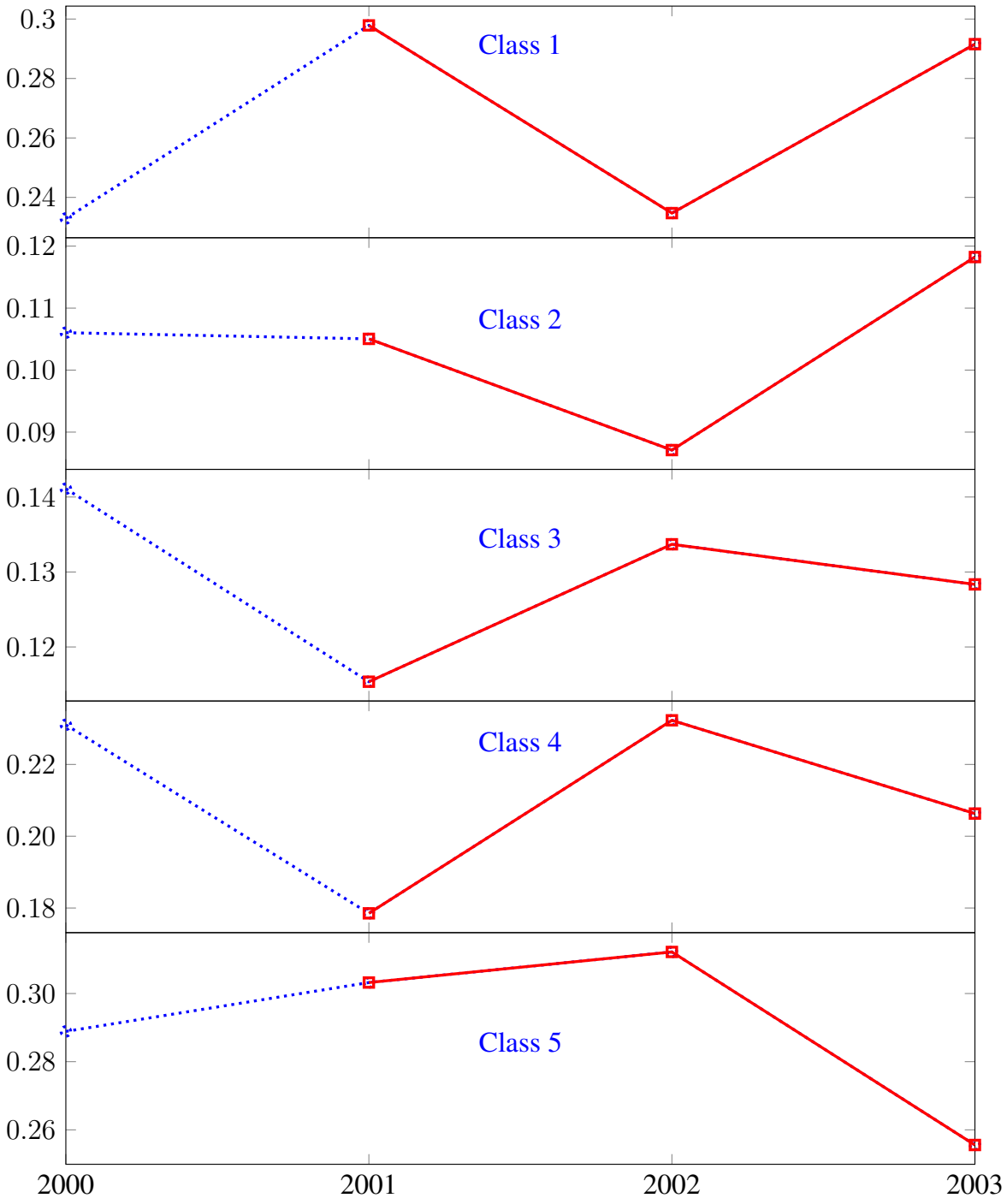


Fig. 4.16: Annual polarities for cluster 1 from Figure 4.3 from the domain *abortion* for the classes 1 to 5. The PMC is shown with solid lines in square markers (2001 to 2003) and foreshadows the partial birth abortion ban of 2003.

4.4 Conclusion

We highlight a problem of significant public and legislative importance, framing change detection. We contribute an unsupervised natural language based approach that detects framing change trends over several years in domain news publishing. We identify a key characteristic of such changes, namely, that during frame changes, the polarity of adjectives describing cooccurring nouns changes cumulatively over multiple years. Our approach agrees with and extends the results of earlier manual surveys. Whereas such surveys depend on human effort and are therefore limited in scope, our approach is fully automated and can simultaneously run over all news domains. We contribute the *Framing Changes Dataset*, a collection of over 12,000 news articles from seven *domains* in which framing has been shown to change by earlier surveys. We release the dataset with our paper. Our work suggests the predictive utility of automated news monitoring, as a means to foreshadow events of commercial and legislative import.

Chapter 5

Conclusions and Directions

5.1 Conclusions

To the best of our knowledge, this dissertation is the first attempt to model public reaction as a function of various news characteristics. Specifically, we identify patterns in news publishing that elicit macro-reactions from the public and legislature, and identify a pattern of varying likelihood for their emergence across different news domains. We provide computational methods to detect changes in news patterns that are likely to be influential.

We introduce the hyper-concentrated period of domain news publishing, which we show Granger-causes macro changes in public and legislative interest. We show that hyper-concentrated periods can be Granger-caused by changes in framing.

We estimate the characteristics of the distributions of news volume and Google Trend query volume change over domains. We infer a systematic selection bias in news and public interest from these distributions, that provides an estimate of how likely a macro-reaction may be in a given domain. We tie in this finding to the characteristic of prominence, and show that news volume, Google Trend query volume change, and prominence tend to exhibit covariance rather than being independent.

Finally, we contribute a fully unsupervised, data-driven approach to the automatic detection of framing changes. We demonstrate the approach by identifying changes that have been described in earlier human surveys. We show that framing changes can foreshadow substantial legislative activity.

Based on these observations, we conclude that our computational tools enable predictive utility for trends in public and legislative response to news patterns. We therefore claim that our contributions are sufficiently precise so as to enable such predictive utility.

Our findings have the following possible ramifications. Firstly, our work seems to suggest that news framing, a purely subjective factor of news, appears to influence legislator behavior. This suggests that news portals may represent a more influential entity in the spheres of politics and public opinion than previously established.

Secondly, the fact that news patterns such as those involving hyper-concentration and framing change may be computationally identified hints at the promise of predicting legislative and public response. This may enable the commercial consequences following events such as the HTML data leak of 2011 to be identified in advance, and possibly mitigated or avoided.

Our work is of substantial interest to audiences over a broad range of disciplines, such as Computer Scientists, Political Scientists, Governmental Agencies locally and abroad, and the general public. Our contributions span the areas of Computational Social Science, Group Communication, Natural Language Understanding, and Political Science.

We emphasize that this work relies on Granger causality as the primary means of establishing correlations between time series. Since Granger causality does not measure "causality" in the strict sense of the term, we consequently do not make strict causal claims.

Further, many exogenous and unobservable factors may influence the dependent variables considered in our study. Our analysis is limited to two data sources. Further, in the case of the New York Times, the data retrieved by the API is limited to a snippet of the full article.

We therefore state that whereas the results we present appear to show influence between news and legislation, we do not claim to have established a causal link. Many of our claims are thus qualified with the prefix "appear to influence".

Nonetheless, we believe that the correlations we establish are strong enough to merit future causal studies. In particular, candidate periods of framing change identified by our approach may be evaluated for causal influence by future work.

5.2 Future Work

Directions for future work are along the following major dimensions:

5.2.1 Machine Learning

With greater availability of news data, more sophisticated natural language models such as novel deep learning architectures may be conceived and applied to problems such as framing change detection, that model and estimate changes in several news aspects. Our initial exploration of framing change detection relies on news polarity, which is an important news characteristics. However, several other characteristics such as subjectivity, tone, and intent may be modeled with larger data.

5.2.2 Topic Modeling

Frame identification is another promising research direction. The present work uses a standard LDA approach to identify frames, which appears to isolate the main frames of interest in well-known domains. However, in smaller domains, it is useful to be able to identify nascent hyper-concentrations in sub-frames of an overall domain. Novel topic modeling approaches may be investigated to temporally identify frames originating within a domain. As an example, consider that within the domain *surveillance*, the term "surveillance+privacy," (referring to the privacy frame within the larger surveillance domain) became hyper-concentrated, foreshadowing the USA Freedom Act. The problem of identifying such frames is a promising direction.

5.2.3 Predictive Models

Whereas this work lays out a framework for modeling the interplay between different news characteristics, a challenging research direction lies in building predictive models that may potentially be applied in real-world scenarios, to estimate future states of dynamic systems such as that of news, or, perhaps even the more complicated system of human behavior.

REFERENCES

Scott Althaus and David Tewksbury. Agenda setting and the new news patterns of issue importance among readers of the paper and online versions of the New York Times. *Communication Research*, 29(2):180–207, 2002.

Lijphart Arend. Comparative politics and the comparative method. *American Political Science Review*, 65(3):682–693, September 1971.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th ELRA International Conference on Language Resources and Evaluation*, pages 2200–2204, Valletta, Malta, may 2010. European Language Resources Association. ISBN 2-9517408-6-7.

Gerard De Melo Mohit Bansal. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290, 2013.

Frank Baumgartner, Bryan Jones, and Peter Mortensen. Punctuated equilibrium theory: Explaining stability and change in public policymaking. *Theories of the Policy Process*, 8: 59–103, 2014.

Jacob Benesty, Jingdong Chen, Yiteng Huang, Cohen Yiteng, and Israel Cohen. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, pages 1–4, 2009.

Robert Benford and David Snow. Framing processes and social movements: An overview and assessment. *Annual Review of Sociology*, 26(1):611–639, 2000.

David Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 1532-4435. http://dl.acm.org/citation.cfm?id=944919.944937.

Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. Selection bias in news coverage: Learning it, fighting it. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, pages 535–543, Republic and Canton of Geneva, Switzerland, 2018. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-5640-4. doi: 10.1145/3184558.3188724. URL https://doi.org/10.1145/3184558.3188724.

Jeffrey Breen. Negative opinion lexicon, 2011. https://goo.gl/kQvlau.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, oct 2001. doi: 10.1023/A:1010933404324. http://dx.doi.org/10.1023/A:1010933404324.

Dallas Card, Amber Boydstun, Justin Justin Gross, Philip Resnik, and Noah Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint*

Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 438–444, 2015.

Dennis Chong and James Druckman. Framing theory. *Annual Reviews on Political Science*, 10:103–126, 2007.

Rob Crossley. Facebook ID leak hits millions of Zynga users, 2011. http://www.develop-online.net/news/facebook-id-leak-hits-millions-of-zynga-users/0107956.

Hugh Culbertson and Guido Stempel. The prominence and dominance of news sources in newspaper medical coverage. *Journalism and Mass Communication Quarterly*, 61(3):671, 1984.

Michael Cummings and Robert Proctor. The changing public image of smoking in the United States: 1964–2014. *Cancer Epidemiology and Prevention Biomarkers*, 23:32–36, 2014.

Claes de Vreese. News framing: Theory and typology. *Information Design Journal*, 13 (1): 51–62, 2005.

Daniel Drezner and Henry Farrell. Web of influence. *Foreign Policy*, 145:32–41, 2004. ISSN 00157228. http://www.jstor.org/stable/4152942.

George Edwards and Dan Wood. Who influences whom? The President, Congress, and the media. *American Political Science Review*, 93(2):327–344, 1999.

Stephen Engel. Frame spillover: Media framing and public opinion of a multifaceted LGBT rights agenda. *Law and Social Inquiry*, 38:403–441, 2013. ISSN 1747-4469.

Robert Entman. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58, 1993.

Caitlin Fitzsimmons. Facebook and Zynga sued over privacy, 2014. http://www.adweek.com/digital/facebook-zynga-sued/.

Katherine Flegal, Margaret Carroll, Brian Kit, and Cynthia Ogden. Prevalence of obesity and trends in the distribution of body mass index among us adults, 1999–2010. *Journal of the American Medical Association*, 307(5):491–497, 2012.

René D Flores. Taking the law into their own hands: Do local anti-immigrant ordinances increase gun sales? *Social Problems*, 62(3):363–390, 2015.

FTC. Children's Online Privacy Protection Rule, 1998. https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.

Guy Golan. Inter-media agenda setting and global news coverage. *Journalism Studies*, 7(2): 323–333, 2006.

Mark Gongloff. The state of the bush economy, 2004. https://money.cnn.com/2004/01/19/news/economy/election_sotu/index.htm.

Clive Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

Kris Gunnars. Ten causes of weight gain in america, 2015. https://www.healthline.com/nutrition/10-causes-of-weight-gain#section12.

Albert Gunther. The persuasive press inference effects of mass media on perceived public opinion. *Communication Research*, 25(5):486–504, 1998.

Hearl Harris. Information gain versus gain ratio: A study of split method biases. In *International Symposium on Artificial Intelligence and Mathematics*, Fort Lauderdale, 2002.

Christopher Hoadley, Heng Xu, Joey Lee, and Mary Beth Rosson. Privacy as information access and illusory control: The case of the Facebook news feed privacy outcry. *Electronic Commerce Research and Applications*, 9:50–60, 2010.

Shanto Iyengar and Donald Kinder. *News that Matters: Television and American Opinion*. University of Chicago Press, 2010.

Melissa Jacoby. Negotiating bankruptcy legislation through the news media. *Houston Law Review*, 41:1092–1144, 2004.

Daniel Kahneman. *Thinking Fast and Slow*. Straus Farrar and Giroux, New York, 2011.

Sei-Hill Kim and Anne Willis. Talking about obesity: News framing of who is responsible for causing and fixing the problem. *Journal of Health Communication*, 12(4):359–376, 2007.

Gary King, Benjamin Schneer, and Ariel White. How the news media activate public expression and influence national agendas. *Science*, 358(6364):776–780, 2017.

Spiro Kiousis. Explicating media salience: A factor analysis of New York Times issue coverage during the 2000 U.S. presidential election. *Journal of Communication*, 54:71–87, 2004.

Richard Landis and Gary Koch. The measurement of observer agreement for categorical data. *Journal of Biometrics*, pages 159–174, 1977.

Jey Han Lau. Pre-trained doc2vec models, 2017. https://github.com/jhlau/doc2vec#pre-trained-doc2vec-models.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, 2014a. International Machine Learning Society.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1188–1196, Beijing, China, 2014b.

Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event registry: Learning about world events from news. In *Proceedings of the 23rd World Wide Web Conference*, pages 107–110. ACM, 2014.

Suman Lee. International public relations as a predictor of prominence of us news coverage. *Public Relations Review*, 33(2):158–165, 2007.

Kalev Leetaru and Philip Schrodt. GDELT: Global data on events, location, and tone, 1979–2012. In *International Studies Association Annual Convention*, volume 2, pages 1–49. Citeseer, 2013.

Richard Lundman. The newsworthiness and selection bias in news about murder: Comparative and relative effects of novelty and race and gender typifications on newspaper coverage of homicide. In *Sociological Forum*, volume 18, pages 357–386. Springer, 2003.

Carolanne Mangles. Search engine statistics, January 2018. https://www.smartinsights.com/search-engine-marketing/search-engine-statistics/.

Mathworks. Correlation coefficients. https://www.mathworks.com/help/matlab/ref/corrcoef.html, 2019a.

Mathworks. The fit function, 2019b. https://www.mathworks.com/help/curvefit/fit.html.

John McCarthy, Clark McPhail, and Jackie Smith. Images of protest: Dimensions of selection bias in media coverage of washington demonstrations, 1982 and 1991. *American Sociological Review*, 61(3):478–499, 1996.

Diana Mutz and Joe Soss. Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opinion Quarterly*, 61:431–451, 1997.

National Cancer Institute. How the news media influence tobacco use. https://cancercontrol.cancer.gov/brp/tcrb/monographs/19/m19_9.pdf, 2019.

NYT. Developer APIs. http://developer.nytimes.com/, 2016.

Office of National Statistics. Internet users in the UK: 2017, 2017. https://tinyurl.com/yadnxcf6.

Pamela Oliver and Gregory Maney. Political processes and local newspaper coverage of protest events: From selection bias to triadic interactions. *American Journal of Sociology*, 106(2):463–505, 2000.

Pew Research. The state of privacy in post-Snowden America, 2016. http://www.pewresearch.org/fact-tank/2016/09/21/the-state-of-privacy-in-america/.

Sahand Rahbar. "The Evil of the Age": The influence of the New York Times on anti-abortion legislation in New York, 1865–1873. *Pennsylvania History Review*, 23(1):146–176, 2016.

Pew Research. 11% of Americans don't use the Internet. Who are they?, 2018. http://www.pewresearch.org/fact-tank/2018/03/05/some-americans-dont-use-the-internet-who-are-they/.

Karthik Sheshadri and Munindar P. Singh. The public and legislative impact of hyper-concentrated topic news. *Science Advances, forthcoming*, 2019.

Karthik Sheshadri, Nirav Ajmeri, and Jessica Staddon. No privacy news is good news: An analysis of New York Times and Guardian privacy news from 2010–2016. In *Proceedings of the 15th Privacy, Security and Trust Conference*, pages 159–167, Calgary, Alberta, Canada, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, 2013.

Richard Socher et. al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP)*, pages 1631–1642, Seattle, WA, Oct 2013. Association for Computational Linguistics.

Stuart Soroka. Good news and bad news: Asymmetric responses to economic information. *Journal of Politics*, 68(2):372–385, 2006.

Stuart Soroka, Lori Young, and Meital Balmas. Bad news or mad news? sentiment scoring of negativity, fear, and anger in news content. *The Annals of the American Academy of Political and Social Science*, 659(1):108–121, 2015.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Pearson Education India, New Delhi, 2006.

Colleen Taylor. After privacy uproar, Quora feeds will no longer show data on what other users have viewed, 2016. https://goo.gl/9wG65R.

The Guardian. Guardian Open Platform. http://open-platform.theguardian.com/, 2016. Accessed: 2016-3-3.

The Washington Post. What makes front page news, 2014. http://www.newseum.org/wp-content/uploads/2014/08/education_resources_frontpageposter.pdf.

Patrick Trasborg. The Google Trends API, 2018. https://www.npmjs.com/package/google-trends-api.

Louis Uchitelle. U.S. economy grows 4.2%; war spending provides push, 2004. https://www.nytimes.com/2004/04/30/business/us-economy-grows-4.2-war-spending-provides-push.html.

US Congress. Personal Data Protection and Breach Accountability Act, 2014. https://www.congress.gov/bill/113th-congress/senate-bill/1995.

U.S. Department of Education. Family Educational Rights and Privacy Act. https://tinyurl.com/ybohwmfm, 1974.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, Nov 2008.

Jonathan Vanian. Drone legislation. https://goo.gl/BZp7dJ, 2015.

Pete Vernon. Five years ago, Edward Snowden changed journalism. https://www.cjr.org/the_media_today/snowden-5-years.php, 2018.

Anthony Viera and Joanne Mills Garrett. Understanding inter-observer agreement: The Kappa statistic. *Family Medicine*, 37:360–363, 2005.

Wikipedia. The New York Times. https://en.wikipedia.org/wiki/The_New_York_Times, 2001.

Wikipedia. The Guardian. https://en.wikipedia.org/wiki/The_Guardian, 2002.

Wikipedia. Cosine similarity, 2017. https://en.wikipedia.org/wiki/Cosine_similarity.

Wikipedia. Star Wars: The Last Jedi, 2017. https://en.wikipedia.org/wiki/Star_Wars:_The_Last_Jedi.

Wikipedia. History of uk immigration control. https://en.wikipedia.org/wiki/History_of_UK_immigration_control, 2019.

Chapter 6

Appendix: Sample Correlations

Correlations for all subsets are shown in Table 6.1. The PMC is shown in bold.

Table 6.1: All correlations for cluster 2 of the domain *Immigration*.

Correlation	Start Index	End Index
1	1	2
1	2	3
1	3	4
1	4	5
1	5	6
1	6	7
1	7	8
1	8	9
1	9	10
1	10	11
1	11	12
1	12	13
1	13	14
1	14	15
1	15	16
1	16	17
1	17	18
0.99	**1**	**3**
0.98	10	12
0.98	12	14

Table 6.1 All correlations for cluster 2 of the domain *Immigration* (continued).

Correlation	Start Index	End Index
0.96	12	15
0.93	12	16
0.93	9	12
0.88	11	13
0.88	9	11
0.87	11	14
0.86	11	16
0.85	11	15
0.85	8	12
0.84	8	10
0.83	7	9
0.82	15	17
0.82	14	16
0.82	10	13
0.81	7	12
0.81	9	13
0.79	10	14
0.79	10	16
0.79	9	14
0.78	9	16
0.78	10	15
0.77	9	15
0.77	8	16
0.77	8	13
0.76	8	15
0.76	8	14
0.75	12	17
0.75	5	7
0.75	12	18
0.74	11	18
0.74	11	17
0.72	6	8
0.72	10	17
0.72	10	18

Table 6.1 All correlations for cluster 2 of the domain *Immigration* (continued).

Correlation	Start Index	End Index
0.72	7	11
0.72	13	15
0.71	7	10
0.70	8	11
0.69	2	4
0.68	7	13
0.68	4	6
0.67	3	5
0.67	16	18
0.66	7	14
0.66	9	17
0.66	9	18
0.66	14	17
0.65	8	17
0.65	7	16
0.65	8	18
0.64	7	15
0.63	6	11
0.62	6	10
0.62	6	12
0.62	6	9
0.61	6	14
0.61	6	13
0.61	5	13
0.61	5	12
0.60	5	8
0.60	5	14
0.60	7	17
0.60	6	16
0.60	4	12
0.59	7	18
0.59	4	13
0.59	5	16
0.59	2	5

Table 6.1 All correlations for cluster 2 of the domain *Immigration* (continued).

Correlation	Start Index	End Index
0.58	6	15
0.58	5	11
0.58	5	15
0.58	5	10
0.57	5	9
0.57	4	16
0.57	4	14
0.56	2	13
0.56	2	12
0.56	4	17
0.56	15	18
0.56	4	18
0.56	3	13
0.56	3	6
0.56	3	12
0.56	1	5
0.55	4	15
0.55	5	17
0.55	5	18
0.55	3	18
0.55	2	18
0.55	3	17
0.55	2	17
0.54	2	16
0.54	3	16
0.54	6	17
0.54	6	18
0.53	2	14
0.53	3	14
0.53	2	15
0.53	3	15
0.52	1	12
0.52	1	13
0.51	1	18

Table 6.1 All correlations for cluster 2 of the domain *Immigration* (continued).

Correlation	Start Index	End Index
0.51	1	17
0.50	13	17
0.50	4	7
0.49	2	6
0.49	1	16
0.49	1	6
0.48	1	14
0.48	1	15
0.47	13	18
0.46	13	16
0.44	4	8
0.43	1	4
0.42	4	9
0.40	4	10
0.40	4	11
0.39	14	18
0.36	3	7
0.36	2	11
0.36	1	7
0.36	3	11
0.35	3	8
0.36	2	7
0.35	2	8
0.35	2	9
0.34	1	8
0.34	2	10
0.34	3	9
0.34	3	10
0.34	1	11
0.34	1	9
0.32	1	10