## SOCIAL SCIENCES

# The public and legislative impact of hyperconcentrated topic news

**Karthik Sheshadri\* and Munindar P. Singh\***

News has been shown to influence public perception, affect technology development, and increase public expression. 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 hyperconcentrated news periods. Last, 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. We note, however, that our analysis does not disprove reverse causality and does not model other confounding factors.

## INTRODUCTION

The effect of news on public behavior has been the subject of considerable scientific interest. Previous work has established that news framing influences public perception (1, 2), affects technology development (3, 4), and contributes to setting agendas (5). Most recently, publishing from small news outlets has been shown to increase short-term public involvement in specific domains (6).

Our work enhances understanding by explicitly modeling the Granger causal (G-causal) (7) link between specific news characteristics, public opinion, and federal legislation. We note that G causality captures directionality in correlation between time series but does not correspond to "true" causality. In this work, we restrict ourselves to G causality and indicate every use of the term with the qualifier "G." First, we demonstrate a predictive relationship between news characteristics and federal legislation.

Second, we show that public and legislative reaction to news follows a punctuated equilibrium model (8). 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 these substantial reactions, namely, that they have high news volume occurring simultaneously with high similarity between articles. We term these periods hyperconcentrated news periods. We note that King et al.'s (6) approach artificially created these news conditions for short time periods and reader subsets.

Third, news reporting in general introduces subjective biases, referred to as framing. We adopt Entman's (9) formulation in this paper. Whereas news publishing is ordinarily event driven, we demonstrate that hyperconcentrated news periods, combining high article volume and similarity, can occur spontaneously, without event-based drivers, as an effect of news framing (see Fig. 1 for a compelling example). We find that hyperconcentrated 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. In addition, 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).

The G-causal flow we found is depicted in Fig. 3. We confirmed each link using a directional G causality test, which evaluates the influence of a G-causal time series on a G-caused one. Our choice of G causality over a structural model was deliberate, because we wished to infer rather than assume structure and direction. We note that previous research (10) agrees with this choice.

The details of the parameters we use are listed in table S2. To the best of our knowledge, we use the most stringent possible parameters to
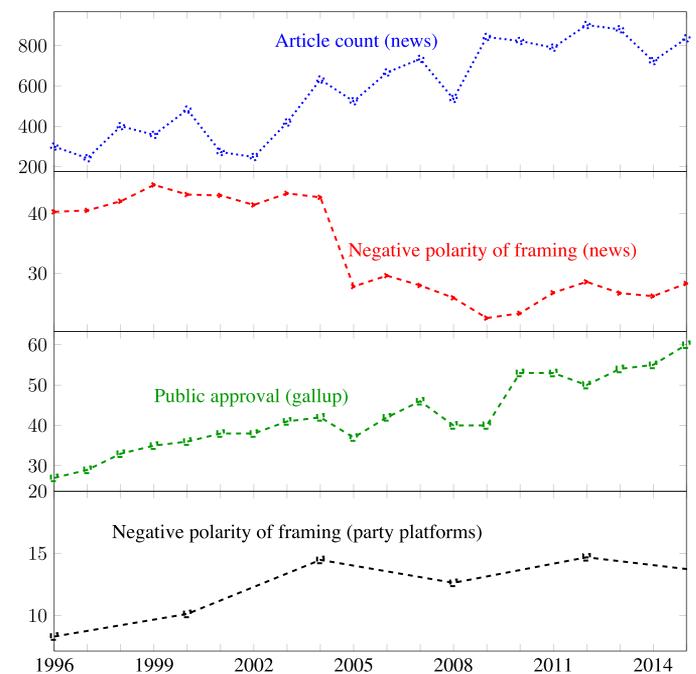


**Fig. 1. News framing as a G-causal precursor to public approval changes foreshadowing legislation in the domain LGBT Rights.** Public approval increases as negative framing declines. Note that the decline in framing polarity after 2004 coincides with a change in framing during 2003 described in an earlier survey (41).

Department of Computer Science, North Carolina State University, Raleigh, NC, USA.
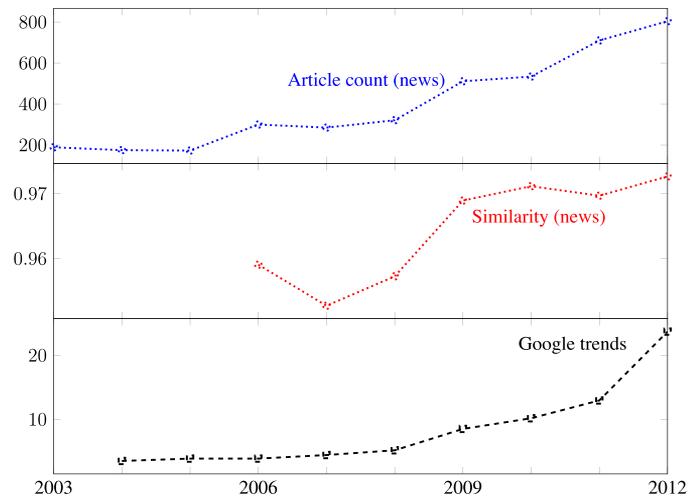*Corresponding author. Email: kshesha@ncsu.edu (K.S.); mpsingh@ncsu.edu (M.P.S.)

**Fig. 2. 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.
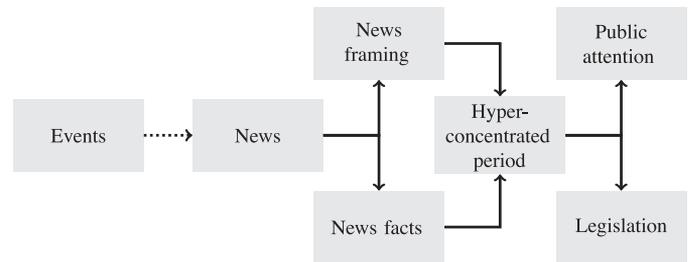


**Fig. 3. We posit the hyperconcentrated period of domain news, characterized by high article volume and similarity, which G-causes public attention changes and legislation.** Hyperconcentrated periods arise either due to news events or, independently of events, due to news framing. We observe and model every link in the figure, except the Events to News link, which is shown with a dotted arrow.
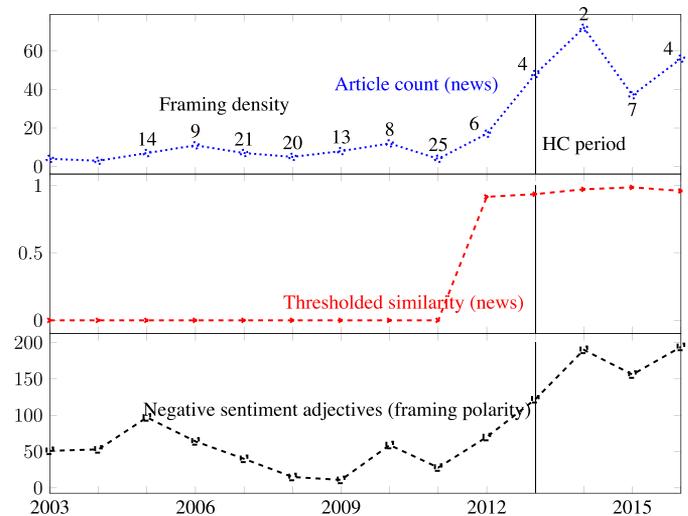


**Fig. 4. Framing changes may be characterized by low framing density and changes in framing polarity.** The figure shows news volume, median article similarity, and framing density in the domain Surveillance spike during a hyperconcentrated period, foreshadowing legislation.

evaluate our hypothesis. We note that if our parameters 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.

## Hyperconcentrated 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 the Supplementary Materials and present a full list in table S1). We posit the idea of a hyperconcentrated 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 G-causally significant changes in legislative activity are often foreshadowed by hyperconcentrated periods. Figure 4 illustrates a hyperconcentrated news period of the Surveillance domain.

We define the median similarity of a domain corpus of size $n$ as the median cosine similarity (11) in paragraph vector (12) 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 G-causally significant). This finding is unexpected, because 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 G-causal relationship between article volume and median corpus similarity is a novel finding of our work.

## Related work

Our approach is similar in spirit to that of King et al. (6) 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 hypercon-

centrated news period and show that hyperconcentrated news is a G-causal precursor to legislation. Our analysis applies to a larger population than the outlets used by King et al. (6), because our data sources (see the "Data sources" section) (13, 14) enjoy wide readership. We measure public perception annually rather than over a period of weeks, as King et al. (6) do. We distinguish between fact-based reporting and framing, and demonstrate that framing in itself is a G-causal predictor of public approval and legislation.

In addition, we note that King et al. (6) artificially created localized short-duration hyperconcentrated news periods in their work. The fact that these short-duration hyperconcentrated news periods did not G-cause legislation motivates the question of how long a hyperconcentrated news period must last to have such an influence. In our data, we found G causality occurring between hyperconcentrated news and federal legislation over periods lasting at least a year. We acknowledge that future work may find G causality over shorter periods.

Our conception of a hyperconcentrated news period is consistent with the idea of punctuated equilibria of media attention introduced by Baumgartner et al. (8) and the notion of an availability cascade posited by Kahneman (15). Last, we note that Jacoby (16) observes correlation
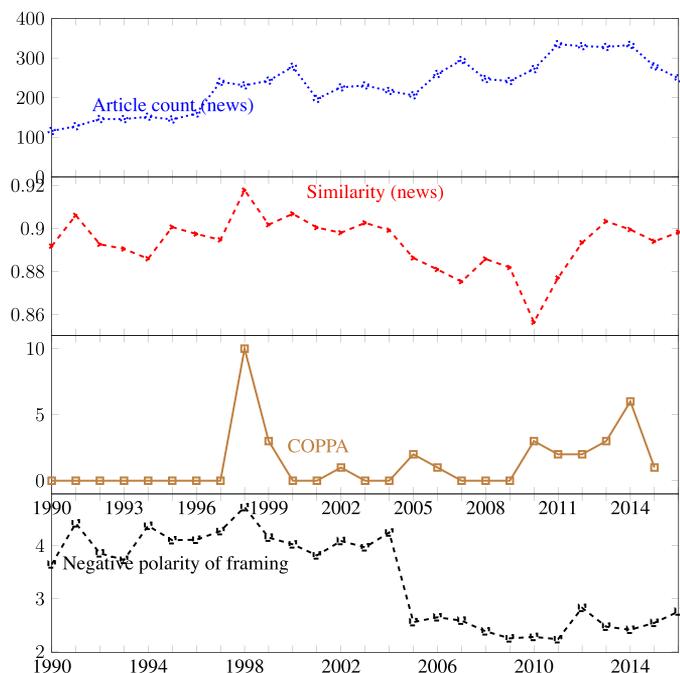
**Fig. 5. News characteristics and legislation for the domain Child Privacy.** Note that during the period 1996–1999, news volume and similarity sharply increase together, foreshadowing COPPA legislation. Notice further that news volume, similarity, and negative polarity of framing reach peaks in 1998, corresponding to the year that COPPA was promulgated.

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 macromutation (8) 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. (8) and Edwards and Wood (10), discuss G-causal effects between media coverage and Congress, we demonstrate that punctuated equilibria extend to sustained public attention and legislative reaction. Further, existing literature does not explicitly model the G-causal link between punctuated equilibria and legislation but restricts itself to measuring reactions by Congress and the president (10). In contrast, we establish G causality between hyperconcentrated news periods and federal legislation.

Our work is conceptually similar to the theory of punctuated equilibria developed by Baumgartner et al. (8). However, our data reveal the following insights that enhance understanding over existing work.

First, Baumgartner et al. posit that attention (the number of articles on a topic) and tone (the perspective adopted in the articles) comprise the two major dimensions of media coverage. Our work suggests a third major dimension, namely, similarity. We show that for multiple domains, median news similarity can have a G-causally significant correlation with legislation. In at least one of these cases, news volume (corresponding with Baumgartner et al.'s attention; see table S2) does not.

Baumgartner et al. (8) and Mazur (17) further posit that as media attention (whether positive or negative) increases, public acceptance decreases. Our results from the LGBT Rights domain (Fig. 1) may instead suggest a more nuanced relationship. Since, in this domain, media attention and public acceptance steadily increased over our period of interest, we posit that the polarity of news framing (see the "Framing Polarity" section) may also influence public acceptance.

In addition, Baumgartner et al. appear to model tone as a Bernoulli variable, considered to be either positive or negative at any given time. Our results in Figs. 1 and 4 instead suggest the utility of modeling tone as a continuous variable.

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 (18) 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" (18) 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 (18). The term framing is sometimes used to refer to tactics (18) 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 these fine-grained processes, and our use of the word frame does not refer to these analyses.

## MATERIALS AND METHODS
This section describes novel algorithms and methods introduced in our work. The Supplementary Materials provides additional details.

### Dataset collection
We described 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* (19) and *The Guardian* (20), 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 (2, 21–24). We noted that *The New York Times* has previously been shown to influence legislation (25), 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 (6, 26), 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 10 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 3 years (for the domain Abortion) and a maximum of 12 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 preexisting 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. First, we do not have advance knowledge of the frames in each domain. Second, 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), 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. 5 for an example, in which similarity increased in 1998, coinciding with COPPA legislation as described in the "Main findings" section). 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 S2. 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 $\kappa$ (27). Our median agreement was $\kappa = 0:67$, considered "substantial agreement" by Landis and Koch (28). 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 (13, 14), we assumed that sufficiently many relevant articles are included in our corpus.

### 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 adopted the idea of an entropic formulation of discriminative keywords, as proposed by Sheshadri *et al.* (26).

Below, a corpus $T$ is a set of news articles. Specifically, given two disjoint sets of news articles $T_1$ and $T_2$, we identified a set of $k$ $n$-grams that yield the largest Cross Entropy (29) 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 $S(x_i, T) = \{A \in T \mid x_i \in A\}$ be the set of articles in corpus $T$ in which the $n$-gram $x_i$ appears. We used a $|T| \times m$ term frequency 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.

$$\text{IG}(T, x_i) = H(T) - \frac{S(x_i, T)}{|T|} H(S(x_i)) \qquad (1)$$

Following Entman's (9) 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 1 January 2014 from those before them, but the unigram "surveillance" is common to articles from both periods and therefore may not. Because keywords from a particular news corpus distinguish it from others, they may be said to represent the "concentration" of news in that corpus.

### 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. To estimate similarity between two documents, we adopted doc2vec (12), a well-known tool that generates a vector representation (called a "paragraph vector") of a document. Specifically, we used a standard doc2vec model (30), trained on each domain corpus, to compute a vector for each document in our corpus. We defined the pairwise similarity of two documents as the cosine similarity of their respective document vectors (31).

Whereas we do not in general deny that high median similarities can occur in annual corpora with low news volume (see fig. S1), 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 paper.

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 S2). 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.

### 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, although 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. 4. We posit, as in Fig. 4, 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.

## Framing polarity
We are interested in measuring the net polarity of the adjectives and adverbs within a corpus. Because 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 (32), 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 (33). 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 (26, 34).

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 hyperconcentrated 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.

## 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.

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

## Main findings
To establish the G-causal effect of hyperconcentrated news on legislation, we considered 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 G-caused by hyperconcentrated news periods. We acknowledge, however, that there may be further examples beyond those identified by our search. Whereas we do not claim hyperconcentrated news periods to be a necessary condition for legislation, we conclude that the probability of legislation being Granger caused by a hyperconcentrated period is statistically significant.

We illustrate our approach and results in Fig. 5, using a compelling example from the domain of Child Privacy. We use the abbreviation "HC period" to refer to hyperconcentrated news periods in this and other figures. The primary laws governing children's privacy protection in the United States are COPPA (Children's Online Privacy Protection Act) (35) and FERPA (Family Educational Rights and Privacy Act) (36). 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. Because of 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 G-causal variables of interest in Fig. 5 are annual news volume (blue dotted line) and median pairwise article similarity (red dashed line). We represent the volume of COPPA legislative activity by a time series depicted with brown solid line in Fig. 5. We represent the primary year of COPPA legislation, 1998, using a value of 10. Other years are represented according to the number of relevant FTC press statements during the year. Our G causality tests are therefore conducted between pairs of independent and (hypothesized) dependent time series, such as between news volume (blue dotted line) and COPPA legislation (brown solid line) in Fig. 5. 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 hyperconcentrated period, COPPA legislation was promulgated through the period ending in 2000. Another hyperconcentrated period occurs before the revival of interest in COPPA, as seen in the large number of amendments in the period 2011–2013.

We tested the G-causal flow depicted in Fig. 3 over the set of domains obtained as described in the previous paragraph (using news volume and legislation as our time-series), yielding statistically significant results in each case (see table S2 for a full list). Our results motivate the predictive utility of news as a G-causal set of independent variables that influence legislation.

Google Trends (37) estimate public interest in a topic of interest by measuring related searches worldwide over chosen time periods. Because 89% of U.S. (38) and 82% of UK residents (39) use the Internet and 74% of Internet users use Google as their primary search (40), we posit that Google Trends are a representative measure of public attention. For one of our domains (see Fig. 2), we found significant G causality between article volume and Google Trends 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. 1, illustrates the G-causal influence of framing on public opinion. Note that the negativity of framing drops in 2004–2005, after which public approval begins to climb steadily. Further, we note that an earlier survey (41) 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. Figure 1 demonstrates an inverse relationship between framing negativity and public approval. We note that following this trend, major LGBT legislation legalizing same-sex marriage in 50 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 G-causes public approval. This finding is reinforced by the fact that event-based drivers cannot influence framing polarity, because only adjectives and adverbs, taken here to be artifacts of how a domain is framed, contribute to 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 1996–2015 (see section S3). To gain confidence in our findings, we address an alternative hypothesis of note, namely, that political framing G-causally influences news framing, and not vice versa. We do not, in general, deny that such a G-causal direction may exist—such an effect has been demonstrated in previous work using news data collected from print newspapers (42). However, we did not find that this effect is G-causally significant for our data over the domains we examine.

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 4 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 this domain in each platform. Figure 1 depicts the results. G causality for the (Political Framing, Public Approval) and (Political Framing, Legislation) tuples was insignificant for this example, in contrast to the (News Framing, Public Approval) and (News Volume, Legislation) tuples, consistent with our hypothesis. We refer the reader to table S1 for a full list comparing the G-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.

Figure 4 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. S3). Further, fig. S3 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. S3 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 G causation between framing density (as computed in Fig. 4) and legislation.

In fig. S3, 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. 4 to depict framing density, were an event-based driver of news and not

in themselves a framing change. However, the Columbia Journalism Review (43) 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 (8).

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 2 years later, in 2015. We show that polarity of negative framing in Surveillance increased following the Snowden revelations (Fig. 4) and remained high until 2015, corresponding exactly with our hyperconcentrated period, after which legislation was promulgated and framing polarity increased.

In addition, we note that in the domain Child Privacy (Fig. 5), framing polarity is at its highest in 1998, coinciding with the introduction of COPPA. Because news events cannot affect framing polarity (because framing polarity depends purely on adjectives and adverbs), and we show that both framing polarity and framing density appear to have distinctive patterns during framing changes (figs. S2 and S3), we conclude that news framing can G-cause legislation.

## Hyperconcentrated news versus political framing as a G cause of legislation

This section details the full results of our G causality study. We consider federal legislation promulgated from 1991 to 2016. For each law, we compiled a news dataset according to the procedure detailed in the "Domain dataset generation" section.

From this list, we manually identified domains for which we were able to obtain data, and for which our data suggested the presence of a hyperconcentrated news period. Table S1 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 G 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 G-causally significant result in each case. Our threshold for significance is $\alpha = 0.05$. For each domain, table S1 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. 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 found that political framing did not G-cause legislation at the $\alpha = 0.05$ level. For two of these domains, we obtained $P = 0.20$ 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 hyperconcentrated news.

Some domains remain unmentioned through the relevant period in both party platforms, such as Cyberbullying, Drones, and Child Privacy. For these domains, because the political parties do not mention the domain, we conclude that there was no measurable political framing of these domains (table S1). Therefore, these domains do not affect our hypothesis, given significant G-causal measures between hyperconcentrated news characteristics and federal legislation.

## Measuring domain framing

Figure S2 shows framing polarity, and fig. S3 depicts framing density for the framing change positives, as well as for a random control. To

generate the random control, we retrieved a sample of 991 articles from *The New York Times* API with a null query for each year between 1990 and 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 section S2). 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. S2), it drops between 2004 and 2005, corresponding to the framing change of late 2003, which was reported in (41). Note also that consistent with our hypothesis, the framing polarity of random news remains close to constant between the years 1990 and 2005.

To depict the framing polarity of Surveillance news in fig. S2 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 the Framing Polarity section), 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 (44) (as described in the "Smoking" subsection of section S2), 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 years 2004 and 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. S3) depicts a generally constant value for random news while 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.

## Comparative evaluation

Last, we evaluate the validity of our hypothesis using the comparative method (45). We conducted tests using both the most different research design and explain that the most similar research design cannot be used for our data. Full results are presented in the Supplementary Materials. We summarize our research design and findings here.

We evaluate our hypothesis using the most different research paradigm (45), 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. To estimate the "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 interdomain comparisons. However, we normalize our overall distance to a scale of zero to one. Our data contains 10 domains with hyperconcentrated periods. We compute all $\binom{10}{2} = 45$ distances and pick the top 10 to represent our most different domain pairs, shown in table S3. Since in each of these domains, federal legislation was enacted, and further since each domain contains a hyperconcentrated 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 (45) 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 hyperconcentrated news periods G-cause legislation, we would evaluate G causality of the domain's news patterns with legislation, both with and without the presence of a hyperconcentrated period.

We were unable to use this research design, since, for many of our domains, there was little or no legislative activity during non-hyperconcentrated periods. This 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 considered federal legislation in the periods for which *The New York Times* and *The Guardian* APIs provide data). We then analyzed domains for which we obtained credible coverage from our APIs, and for which our data indicated the presence of a hyperconcentrated news period. We acknowledge, however, that there may be additional domains with hyperconcentrated periods that our search omitted. We find eight cases in which hyperconcentrated news G-causes legislation, as shown in table S1. Further, our comparative analysis demonstrates through the most different (table S3) paradigm that our hypothesis remains valid despite wide variation in the domains.

## CONCLUSION

Our data support the conclusion that hyperconcentrated news periods in news, brought about by both driver events and framing changes, G-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.

## SUPPLEMENTARY MATERIALS

Section S1. G causality
Section S2. Measuring domain framing
Section S3. Legislation
Section S4. Results from the comparative method
Fig. S1. The figure visualizes six years from our Surveillance dataset.
Fig. S2. Framing polarity: Random versus LGBT, Surveillance, and Smoking news.
Fig. S3. Framing density for random news versus for framing change positives (Smoking, Surveillance, and LGBT Rights).

Fig. S4. News volume and similarity as predictors of legislation in Cyberbullying.
Table S1. Comparing the G-causal effect of hyperconcentrated news against that of political framing for legislation in our domains.
Table S2. Details of our G causality study.
Table S3. A comparative evaluation of our hypothesis using the most different research design.
Reference (46)

## REFERENCES AND NOTES

1. A. C. Gunther, The persuasive press inference: Effects of mass media on perceived public opinion. *Commun. Res.* **25**, 486–504 (1998).
2. D. C. Mutz, J. Soss, Reading public opinion: The influence of news coverage on perceptions of public sentiment. *Public Opin. Q.* **61**, 431–451 (1997).
3. C. M. Hoadley, H. Xu, J. J. Lee, M. B. Rosson, Privacy as information access and illusory control: The case of the Facebook news feed privacy outcry. *Electron. Commer. Res. Appl.* **9**, 50–60 (2010).
4. C. Taylor, After privacy uproar, Quora feeds will no longer show data on what other users have viewed (2016); https://goo.gl/9wG65R.
5. S. Iyengar, D. Kinder, *News That Matters: Television and American Opinion* (University of Chicago Press, 2010).
6. G. King, B. Schneer, A. White, How the news media activate public expression and influence national agendas. *Science* **358**, 776–780 (2017).
7. C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424–438 (1969).
8. F. R. Baumgartner, B. D. Jones, P. B. Mortensen, Punctuated equilibrium theory: Explaining stability and change in public policymaking, in *Theories of the Policy Process*, P. A. Sabatier, C. M. Weible, Eds. (Westview Press, 2014), pp. 59–103.
9. R. M. Entman, Framing: Toward clarification of a fractured paradigm. *J. Commun.* **43**, 51–58 (1993).
10. G. C. Edwards III, B. D. Wood, Who influences whom? The president, Congress, and the media. *Am. Polit. Sci. Rev.* **93**, 327–344 (1999).
11. P.-N. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining* (Pearson Education India, 2006).
12. Q. Le, T. Mikolov, Distributed representations of sentences and documents, in *Proceedings of the 31st International Conference on Machine Learning* (International Machine Learning Society, 2014), pp. 1188–1196.
13. Wikipedia, The New York Times (2001); https://en.wikipedia.org/wiki/The_New_York_Times.
14. Wikipedia, The Guardian (2002); https://en.wikipedia.org/wiki/The_Guardian.
15. D. Kahneman, *Thinking Fast and Slow* (Straus Farrar and Giroux, 2011).
16. M. B. Jacoby, Negotiating bankruptcy legislation through the news media. *Houst. Law Rev.* **41**, 1092–1144 (2004).
17. A. Mazur, Media coverage and public opinion on scientific controversies. *J. Commun.* **31**, 106–115 (1981).
18. R. Benford, D. Snow, Framing processes and social movements: An overview and assessment. *Annu. Rev. Sociol.* **26**, 611–639 (2000).
19. The New York Times, Developer APIs (2016); http://developer.nytimes.com/.
20. The Guardian, Guardian Open Platform (2016); http://open-platform.theguardian.com/ [accessed 3 March 2016].
21. S. L. Althaus, D. Tewksbury, Agenda setting and the "new" news: Patterns of issue importance among readers of the paper and online versions of the New York Times. *Commun. Res.* **29**, 180–207 (2002).
22. D. W. Drezner, H. Farrell, Web of influence. *Foreign Policy* **145**, 32–40 (2004).
23. G. Golan, Inter-media agenda setting and global news coverage. *J. Stud.* **7**, 323–333 (2006).
24. S. Kiousis, Explicating media salience: A factor analysis of *New York Times* issue coverage during the 2000 U.S. presidential election. *J. Commun.* **54**, 71–87 (2004).
25. S. K. Rahbar, "The Evil of the Age": The influence of the New York Times on anti-abortion legislation in New York, 1865–1873. *Penn. Hist. Rev.* **23**, 146–176 (2016).
26. K. Sheshadri, N. Ajmeri, J. 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 International Conference on Privacy, Security and Trust* (IEEE Computer Society, 2017), pp. 159–167.
27. A. Viera, J. M. Garrett, Understanding inter-observer agreement: The kappa statistic. *Fam. Med.* **37**, 360–363 (2005).
28. J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174 (1977).
29. Mathworks, Matlab: Fit binary decision tree for multiclass classification (2019).
30. J. H. Lau, Pre-trained doc2vec models (2017); https://tinyurl.com/yddn2pwb.
31. Wikipedia, Cosine similarity (2017); https://en.wikipedia.org/wiki/Cosine_similarity.
32. S. Baccianella, A. Esuli, F. 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* (European Language Resources Association, 2010), pp. 2200–2204.
33. J. Breen, Negative opinion lexicon (2011); https://goo.gl/kQvlau.
34. S. N. Soroka, Good news and bad news: Asymmetric responses to economic information. *J. Polit.* **68**, 372–385 (2006).
35. Federal Trade Commission, Children's online privacy protection rule (1998); www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule.
36. U.S. Department of Education, Family Educational Rights and Privacy Act (1974); https://tinyurl.com/com9wmfm.
37. P. Trasborg, The Google Trends API (2018); www.npmjs.com/package/google-trends-api.
38. M. Anderson, A. Perrin, J. Jiang, 11% of Americans don't use the internet. Who are they? (2018); www.pewresearch.org/fact-tank/2018/03/05/some-americans-dont-use-the-internet-who-are-they/.
39. Office for National Statistics, Internet users in the UK: 2017 (2017); https://tinyurl.com/yadnxcf6.
40. C. Mangles, *Search engine statistics* (2018); www.smartinsights.com/search-engine-marketing/search-engine-statistics/.
41. S. M. Engel, Frame spillover: Media framing and public opinion of a multifaceted LGBT rights agenda. *Law Soc. Inq.* **38**, 403–441 (2013).
42. R. D. Flores, Taking the law into their own hands: Do local anti-immigrant ordinances increase gun sales? *Soc. Probl.* **62**, 363–390 (2015).
43. P. Vernon, Five years ago, Edward Snowden changed journalism (2018); https://www.cjr.org/the media today/snowden-5-years.php.
44. National Cancer Institute, How the news media influence tobacco use (2019); https://cancercontrol.cancer.gov/brp/tcrb/monographs/19/m19_9.pdf.
45. A. Lijphart, Comparative politics and the comparative method. *Am. Polit. Sci. Rev.* **65**, 682–693 (1971).
46. J. Vanian, Drone legislation (2015); https://goo.gl/BZp7dJ.

**Citation:** K. Sheshadri, M. P. Singh, The public and legislative impact of hyperconcentrated topic news. *Sci. Adv.* **5**, eaat8296 (2019).

# Science Advances

## The public and legislative impact of hyperconcentrated topic news
Karthik Sheshadri and Munindar P. Singh