Quantifying Bias in Social and Mainstream Media

Yu-Ru Lin[1] James P. Bagrow[2] David Lazer[1]

[1]Northeastern University and Harvard University

[2]Northwestern University

Social media, such as blogs, are often seen as democratic entities that allow more voices to be heard than the conventional mainstream media as well as a balancing force against the arguably slanted elite media. A systematic comparison between social and mainstream media is necessary but challenging due to the scale and dynamic nature of modern communication. We propose empirical measures to quantify the extent and dynamics of social (blog) and mainstream (news) media bias. We focus on a particular form of bias—coverage quantity—as applied to stories about the 111th US Congress. We compare observed coverage of Members of Congress against a null model of unbiased coverage, testing for biases with respect to political party, popular front runners, regions of the country, and more. Our measures suggest distinct characteristics in news and blog media. A simple generative model, in agreement with data, reveals differences in the process of coverage selection between the two media.

DOI: 10.1145/2246063.2246068 http://doi.acm.org/10.1145/2246063.2246068

1. INTRODUCTION

The extent of media bias determines the information available to the public and can affect public opinion and decision-making. Social media, powered by the growth of the Internet and related technologies, is envisioned as a form of grassroots journalism that blurs the line between producers and consumers and changes how information and opinions are distributed. Indeed, social media can be used by underprivileged citizens, promising a profound impact and a healthy democracy.

Many believe that the mainstream media is slanted, but disagree about the *direction* of slant. The conventional belief about media bias has held for decades, but attempts at developing objective measurement have only recently begun. The study by Groseclose and Milyo [Groseclose and Milyo 2005] showed the presence of bias in mass media (cable and print news) and new media (Internet websites, etc.). On the other hand, researchers have observed an "echo chamber" effect within the new media – people select particular news to reinforce their existing beliefs and attitudes. Iyengar and Hahn [Iyengar and Hahn 2009] argued that such selective exposure is especially likely in the new media environment due to information overload. Computationally identifying bias from media content remains an

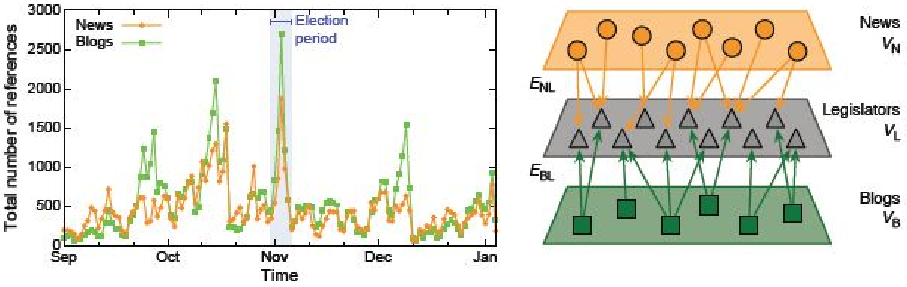

Fig. 1: [Left] The volume (total number of news articles or blog posts) over time. The highest peak corresponds to the mid-term election. [Right] The networked data model. There are three types of nodes: news outlets, blog outlets and legislators. An edge pointing toward a legislator represents each time an outlet references that legislator in an article or post.

emerging research topic, and requires insights from other language analysis studies such as sentiment analysis [Pang and Lee 2008] or partisan features in texts [Monroe et al. 2008; Gentzkow 2010].

Do social media exhibit more or less bias than mass media and, if so, to what extent? Identifying media bias is challenging for a number of reasons. First, bias is "in the eyes of the beholder" and hence not easy to observe, e.g., conservatives tend to believe that there is a liberal bias in the media while liberals tend to believe there is a conservative bias [Groseclose and Milyo 2005; Yano et al. 2010]. Second, the assessment of bias usually implies knowing what "fairness" would be, which may not be available or consistent across different viewpoints. Third, Internet-based communication promises easy, inexpensive, and instant information distribution, which not only increases the number of online media outlets, but also the amount and frequency of information and opinions delivered through these outlets. The scale and dynamic nature of today's communication should be accounted for.

The article presents our research [Lin et al. 2011] that attempts to overcome the challenges in the preceding discussion. We propose empirical measures to quantify the extent and dynamics of social (blog) and mainstream (news) media bias. Our major contribution is that we propose empirical measures to quantify the extent and dynamics of "bias" in mainstream and social media (hereafter referred to as *News* and *Blogs*, respectively). Our measurements are not normative judgment, but examine bias by looking at the attributes of those being mentioned, against a null model of "unbiased" coverage. We focus on the number of times a member of the 111th US congress was *referenced*, and study the distribution and dynamics of the references within a large set of media outlets. We demonstrate bias measures for slants in favor of specific political parties, popular front-runners, or certain geographical regions. Our measures suggest distinct characteristics in news and blog media. A simple generative model, in agreement with data, reveals differences in the process of coverage selection between the two media.

2. QUANTIFYING BIAS

Data Collection Our data is based on RSS feeds aggregated by OpenCongress[1][2]. OpenCongress is a non-profit, non-partisan public resource website that brings together official government data with timely information about what is happening in Congress. We continuously monitor and collect the OpenCongress RSS feeds for each individual member of Congress[3]. This paper examines News and Blogs coverage about the 111th US Congress, both Senators and Representatives. The dataset spans from September 1 to January 4, covering the 2010 mid-term election on November 2.

Figure 1 [left] shows the volume (total number of news articles or blog posts) over time in this dataset. The central peak corresponds to the mid-term election. In total, there are 57,221 news articles and 66,830 blog posts being collected in the four-month period.

Reference Networks We study the structure of the two media by constructing a modal network containing different types of nodes and edges. The network structure is illustrated in Fig. 1 [right]. More specifically, we have:

Nodes. There are three sets of nodes: a news set, denoted by V_{N}, that contains 5,149 news outlets, a blog set V_{B} of 19,693 blogs[4], and a legislator set V_{L} that covers 530 lawmakers.

Edges. Each edge e_{ik} records when media outlet i publishes an article referencing legislator k. We extract 64,222 such edges in 46,501 news articles, denoted as edge set E_{NL}, and 91,837 edges in 62,301 blog posts, denoted as E_{BL}. Edges are associated with timestamps and texts.

Node attributes. For legislators, we record attributes such as party, district, etc., based on the legislators' profiles and external data sources.

While we focus on "reference" or citation edges, this networked model can also include other types of edges, e.g. hyperlinks between outlets, voting preferences among legislators, etc.

Bias measure Consider a news or blog outlet's biased coverage of two political parties. We quantify bias of an outlet i by a *slant score* θ_{ik} which is defined as

$$\theta_{ik} = \log(\text{odds-ratio}) = \log\left(\frac{n_{ik}/(n_i - n_{ik})}{p_k/(1-p_k)}\right), \tag{1}$$

where n_{ik} be number of times an outlet i references legislators in group k, n_i is the total references of i, and p_k is the *baseline probability* that i refers to k. The advantage of having such a baseline probability is that "fairness" become configurable, e.g., one can consider fairness as a 50-50 chance to reference either party (i.e. $p_{\mathrm{D}} = p_{\mathrm{R}} = 0.5$), or define $p_{\mathrm{D}} = 0.6$ since roughly 60% of the Congress are Democrats. In this two-party case,

[1] `www.opencongress.org`

[2] OpenCongress uses Daylife (`www.daylife.com`) and Technorati (`technorati.com`) to aggregate articles from these feeds. The possible selection biases in these filtering processes are not considered in this paper.

[3] An example news/blog coverage feed can be found at `http://www.opencongress.org/people/news_blogs/300075_Lisa_Murkowski`

[4] We also have a small number of blogs hosted by mass media news outlets, e.g. CNN (blog). This paper does not include analysis of such blogs.

we take $\theta_i \equiv \theta_{ik}$, with $k = \mathrm{D}$, and $\theta_i > 0$ means outlet i is more likely to be D-slanted and 0 simply means no bias w.r.t that baseline. To characterize the overall bias within a media, we derive a media-wide *collective slant score*, Θ, which is defined as $\Theta \equiv \theta^*$, where θ^* is the asymptotically unbiased estimator [DerSimonian and Laird 1986] for θ based on a *random effect* model.

To study how media bias may change over time, we calculate the slant scores using references made during running windows. We measure $\Theta(t, w)$ as a function of time t and window length w. Figure 2 [left] shows the temporal slant scores for the two media during the four-month period, based on a $w =$ 2-week running window. The slant of both media changes slightly after the mid-term election: Compared with their pre-election slants, News become slightly more R-slanted when referencing Senators and Blogs are more R-slanted when referencing Representatives. Overall, the media, especially Blogs, become more R-slanted after election. This is reasonable due to the Republican victories.

These results raise an important question: do the majority of outlets become more R-slanted after the election, or do R-slanted outlets become more active while D-slanted outlets become quieter? To examine what caused the slant change we plot in Fig. 2 [top-right] the change in slant score $\Delta\theta_i = \theta_i(t_2) - \theta_i(t_1)$, where $t_1 \in$ [Sep. 1, Oct. 30] and $t_2 \in$ [Nov. 7, Jan. 4], for each outlet against its slant score before the election. (Point size indicates the amount of references observed after the election.) We use a linear regression to quantify the slant change. Surprisingly, we see media outlets shifted slightly toward the other side after the election regardless of their original slants, but overall the originally D-slanted outlets become more R-slanted (as shown in Fig. 2 [top-right]).

We extend such dichotomous-outcome measures to multi-outcome bias measures such as front-runner slant. Using these measures to examine newly collected data, we have observed distinct characteristics of how News and Blogs cover the US congress. Our analysis of party and ideological bias indicates that Blogs are not significantly less slanted than News. However, their slant orientations are more sensitive to exogenous factors such as national elections. In addition, blogs' interests are less concentrated on particular front-runners or regions than news outlets.

3. MODELING THE REFERENCE-GENERATING PROCESS

To better understand the distinctive slant structures between the two media, we propose to use a simple "wealth allotment" model [Bagrow et al. 2008] to explain how legislators gain attention (references) from different media. The model is as follows. Initially ($t = 0$), we assume a single reference to some legislator k' such that $n_k(0) = \delta(k, k')$, for all k. At each time step the media (News or Blogs) selects a random legislator to reference in an article. With probability q, however, the media rejects that legislator and instead references a legislator with probability proportional to his or her current coverage. That is, at each time step t, $n_k(t+1) = n_k(t) + 1$ occurs with probability $p_k(t)$:

$$p_k(t) = \begin{cases} 1/\left|V_{\mathrm{L}}\right| & \text{with prob. } 1-q \ , \\ n_k(t)/\sum_{k'} n_{k'}(t) & \text{with prob. } q. \end{cases} \tag{2}$$

This captures the intuitive "rich-get-richer" notion of fame, while the parameter q tunes its relative strength. Those legislators lucky (or newsworthy) enough to be referenced early

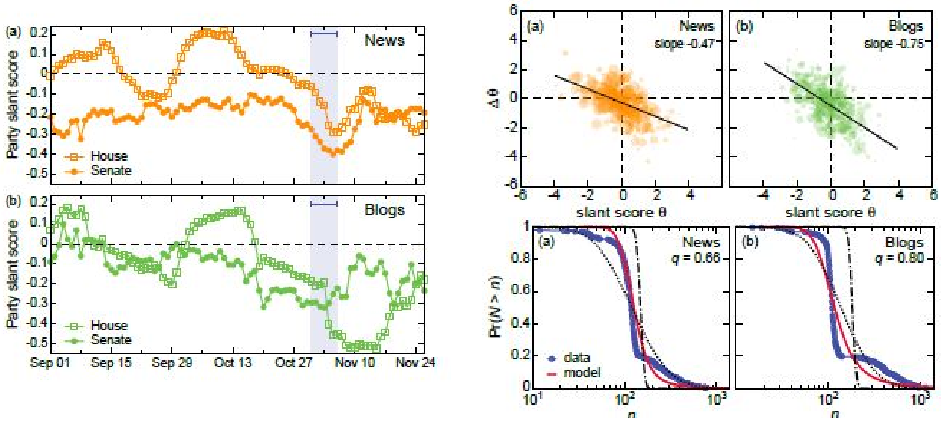

Fig. 2: [Left] Slant score as a function of time. Overall, the media, especially Blogs, become more R-slanted after the 2010 election. [Top-right] Media outlets are slightly shifting towards the other side after election. The majority of news outlets become slightly more R-slanted. For blogs, originally D-slanted blogs become more R-slanted. Each point represents a media outlet. [Bottom-right] The generative model for the distribution of references n per legislator. The larger value of q for Blogs indicates that they are more driven by the rich-get-richer mechanism than News (both distributions are heavy-tailed). Dashed lines indicated fitted poisson and log-normal distributions, for comparison.

on are likely to become heavily referenced, since they have more opportunities to receive references, especially as q increases.

Figure 2 [bottom-right] compares the observed $P(n)$ with that generated using the model process. We observe good qualitative agreement, better than fitted poisson or log-normal distributions, although there is a slight tendency to overestimate popular legislators and underestimate unpopular legislators. The empirical distributions also exhibit a slight bimodality, perhaps due to the 2010 election, that is not captured by the model. The larger value of q for Blogs than for News provides evidence that Blogs collectively are more driven by a rich-get-richer selection process than News, although this may not hold at the individual outlet level.

This observation does not contradict our measures of bias – compared with news media, blogs are weaker adherents to particular parties, front-runners or regions but are more susceptible to the network and exogenous factors.

4. DISCUSSION AND FUTURE DIRECTIONS

In this paper, we describe system-wide bias measures to quantify bias in mainstream and social media, based on the number of times media outlets reference to the members of the 111th US Congress. In addition to empirical measurements, we also present a generative model to explore how each media's global distribution of the number of references per legislator evolves over time. We observe that social media are indeed more social, i.e. more affected by network and exogenous factors, resulting in a more heavily-skewed and uneven distribution of popularity.

We plan to continue work on long-term tracking of slant dynamics in the two media, modeling individual outlets' biases, and leveraging content analysis and deterministic learning methods. We believe the study sheds light on how political communication is carried on in different forms of media and how the targeting audience may be identified after adjusting likely biases.

ACKNOWLEDGMENTS

We thank J. Chu-Shore, F. Simini and J. Menche for many useful discussions, and gratefully acknowledge support from MURI grant # G00003585. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reect the views of the funding source.

REFERENCES

BAGROW, J. P., SUN, J., AND BEN-AVRAHAM, D. 2008. Phase transition in the rich-get-richer mechanism due to finite-size effects. *J. Phys. A 41,* 18, 185001.

DERSIMONIAN, R. AND LAIRD, N. 1986. Meta-analysis in clinical trials. *Controlled clinical trials 7,* 3, 177–188.

GENTZKOW, M. 2010. What drives media slant? evidence from us daily newspapers. *Econometrica 78,* 1, 35–71.

GROSECLOSE, T. AND MILYO, J. 2005. A measure of media bias. *The Quarterly J. of Economics 120,* 4, 1191–1237.

IYENGAR, S. AND HAHN, K. 2009. Red media, blue media: Evidence of ideological selectivity in media use. *J. of Communication 59,* 1, 19–39.

LIN, Y.-R., BAGROW, J. P., AND LAZER, D. 2011. More voices than ever? quantifying media bias in networks. In *ICWSM*. The AAAI Press.

MONROE, B., COLARESI, M., AND QUINN, K. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis 16,* 4, 372.

PANG, B. AND LEE, L. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval 2,* 1-2, 1–135.

YANO, T., RESNIK, P., AND SMITH, N. 2010. Shedding (a thousand points of) light on biased language. In *NAACL Workshop on Creating Speech and Language Data With Amazons Mechanical Turk.*