Social Media News Communities: Gatekeeping, Coverage, and Statement Bias

Diego Saez-Trumper
Universitat Pompeu Fabra
Barcelona, Spain
dsaez-trumper@acm.org

Carlos Castillo
QCRI
Doha, Qatar
chato@acm.org

Mounia Lalmas
Yahoo! Labs
Barcelona, Spain
mounia@acm.org

ABSTRACT

We examine biases in online news sources and social media communities around them. To that end, we introduce unsupervised methods considering three types of biases: selection or "gatekeeping" bias, coverage bias, and statement bias, characterizing each one through a series of metrics. Our results, obtained by analyzing 80 international news sources during a two-week period, show that biases are subtle but observable, and follow geographical boundaries more closely than political ones. We also demonstrate how these biases are to some extent amplified by social media.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing

General Terms

Measurement, Human Factors

Keywords

Online News, Framing, News Bias

1. INTRODUCTION

What is published by the news media depends on numerous factors, an important one being the newsworthiness of a story, but also factors such a space constraint, timeliness, and how close a story is to readers in a geographical and cultural sense [7]. Since it is impossible to report everything, selectivity is inevitable. Nonetheless, reputable news media are expected to be objective in which stories they report and how they report them; their role is to inform people about what is happening either locally, nationally or worldwide.

It is however known that *media bias* exists. For instance, Fox News has been formally accused of misrepresenting facts in an effort to appeal to conservative viewers.[1] Exposure to biases in news reporting has numerous consequences. It has been shown to have the capability to foster intolerance as well as ideological segregation, and even antagonisms in major political and social issues [8]. Bias can also affect voting behavior, depending on the degree and direction of it, and on voters' reliance on media [6, 10]. Being aware, tracking, and overcoming bias in news reporting is important for a fair society, as media indeed has the power to shape a democratic society.

Twitter is a major component of the online news ecosystem. Once passive, users consuming news online filter news and discuss what media publish on Twitter. Social media can play a major role in terms of overcoming biases, since social media users can freely (in principle) report on current events showing other angles of a story, and also help us understand the presence of bias in the news online ecosystem. In this paper, we focus on the latter.

Bias can happen in a number of ways [19]: which stories are selected, *selection* bias, how much attention is given to a story, *coverage bias*, and how a story is reported, *statement bias*. We analyze data from dozens of international news media organizations to answer *how can we quantify biases in online news?* Naturally, we do not expect that media bias can be reduced to a single quantity or metric. Therefore, for each type of bias, we introduce a set of metrics that capture different aspects of it. Our main contributions are the following:

- We introduce unsupervised methods to characterize biases in online news media and in their communities in social media.
- We demonstrate multiple metrics that capture geographical and political biases in a large sample of international news media.
- We describe how, in some cases, biases in social media are amplified with respect to traditional news sources.

2. DEFINITIONS

This section introduces the concepts we use in this paper. We start with three definitions.

News article. An online news article or simply "article" is any document with a publicly-accessible URL, posted on one of the news websites we follow.

[1] `http://www.guardian.co.uk/media/2004/jun/15/broadcasting.ofcom`

CIKM'13, October 27–November 01 2013, San Francisco, CA, USA

ACM 978-1-4503-2263-8/13/10 ... $15.00.
http://dx.doi.org/10.1145/2505515.2505623.

News story. An online news story or simply "story" is a collection of several articles that are strongly related to a seminal event [15], e.g. "Death of former UK PM Thatcher".

Entity. We focus on people mentioned in the news, i.e. named entities of type person appearing in the content of news articles or Twitter messages. This typically includes politicians, athletes, and artists, among many others.

2.1 Bias

Bias is offering a partial perspective on facts [18]. The degree to which bias is present on a text is often subject of considerable debate. We consider three types of bias, following the work by [5]:

Selection bias or gatekeeping. In partisan politics, the preference for selecting stories from one party. We observe selection bias by determining which media/community covers a certain story or person.

Coverage bias. In partisan politics, the preference for giving a larger amount of coverage (time/space) to stories about one party. We observe coverage bias by looking at the amount of attention each story or person is given.

Statement bias. In partisan politics, the preference for expressing more favorable (or more disfavorable) statements for one party. We observe statement bias by looking at the sentiments in statements mentioning different people.

2.2 Social media news communities

All of the news sources we study have a community of social media users who follows and reposts their stories in social media. There are at least two ways in which these communities can be understood. On one hand, social media communities can be considered a type of "fan" of the news source. On the other hand, some community members can also be considered as part of the interpretive community of news [24], as they are becoming more accustomed to participate actively on the news process (e.g., discovering new stories or placing them into context).

Both interpretations agree that some users will be more active than others and motivate our next definition. We define the *active social media community* of an online news source (or "community" in the rest of this paper), as the set of users who are *regularly* exposed to articles from that source (most days of the week for instance), are *interested* in sharing those articles, and are *active* in social media. In the next section we operationalize this definition.

3. DATA PROCESSING

We assembled two collections, one containing news articles and the other containing Twitter messages ("tweets").[2]

3.1 Collecting news articles

Our data collection covers a large fraction of the English-speaking audience of online news. Alexa[3] maintains a list of the most visited sites on the Web; from this list, we picked the top 100 websites under the category "news". We added to this list prominent international news sources listed on Wikipedia.[4] We discarded news aggregators (e.g., Yahoo! News) and websites that do not belong to traditional news organizations (e.g., Reddit, The Onion, and PR Web).

Next we determined the RSS feed for each news source, when available, and their corporate Twitter account(s). A news website may have more than one corporate Twitter account. For the cases where they correspond to different sections of the site (e.g., `@BBCWorld` and `@BBCBusiness`), we considered each account as a separate news source. For the cases where one account links to a subset of the news posted from a second account (e.g., `@AJELive` and `@AJEnglish`), we merged them.

During a period of two weeks in April 2013, we checked each source every 30 minutes for new articles. We considered URLs appearing in the RSS feed (when present), or posted through a corporate Twitter account. After downloading, we removed the ancillary elements of each page (common headers, footers, navigational elements, etc.) using a service[5] that applies an heuristic based on tag-to-text ratio. Finally, named entities where extracted using Open Calais.[6]

3.2 Aggregating articles into stories

News articles can be aggregated into stories that discuss a common event or topic. In order to create news stories, we measured the cosine similarity of pairs of articles using TF.IDF weighting. We use the same measure of text similarity in other tasks throughout the paper. Two articles having a similarity larger than $\theta = 0.4$ (set empirically on a hold-out set) were considered equivalent in terms of content. We built a graph containing all articles, joining by an edge all articles having a similarity larger than the threshold. Each story corresponds to one connected component of this graph, similarly to [25]. These sub-graphs were post-processed to increase precision, ensuring that all articles on each group were closely related to each other. The post-process consisted in recursively removing all vertices with degree less than 2.

3.3 Determining social media communities

The active social media community of a news source should include people who are likely to read the news source almost every day, and who frequently share on Twitter articles from that source. For each article from a news source, we collected the usernames of all Twitter users who posted that URL[7] on Twitter in the first 12 hours after the article's publication. A recent study [11] showed that almost all shares of news articles happen during this period. To avoid automatic accounts (bots) we blacklisted all users posting more than 10 articles from a single source within a day.[8] Twitter's API allows to obtain up to 1,500 tweets for each URL, which was enough for all the articles in our observation period.

We define the community of a news source on a given day, as the set of all the people who have tweeted at least K_1

[2]Both collections are available upon request for research purposes.

[3]`http://www.alexa.com/topsites/category/Top/News`

[4]`http://en.wikipedia.org/wiki/International_broadcasting` `http://en.wikipedia.org/wiki/International_news_channels`

[5]`http://viewtext.org/` with ratio 0.7.

[6]`http://www.opencalais.com/`

[7]Shortened URLs (e.g. `bit.ly` ones) were expanded, and all URLs were normalized by removing unnecessary and/or tracking-related parameters.

[8]We actually observed that many of these accounts were later removed/deactivated by Twitter.

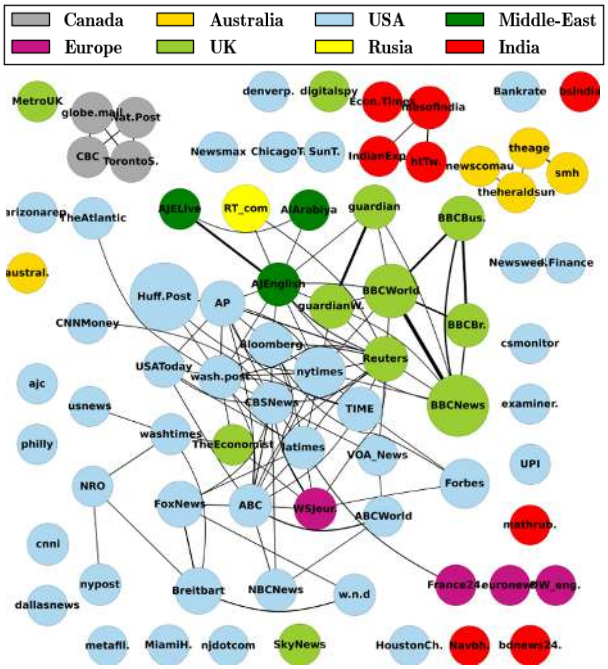


Figure 1: Depiction of the overlap of the communities between news sources. Edges connect two sources if the Jaccard coefficient of their respective community members is greater than 0.03 (Best seen in color).

articles from that news source in the past K_2 days. While we expect communities to be dynamic, we do not expect them to change completely from day to day, hence we tune the two parameters K_1 and K_2 to provide a certain degree of stability. After experimentation, we set $K_1 = 3$, $K_2 = 3$, which produces communities that change by roughly 10% of their members every day on average.

Previous works e.g. [1] have considered the followers of a news source on Twitter as its community. While our definition is different, community sizes according to both definitions are correlated ($r^2 = 0.74$, computed in log-log scale). The few exceptions are news media sources with content that does not change on a daily basis, such as Newsweek.

The number of common followers in Twitter between US-based news media has been found to be correlated with political leaning [1]. We perform the same measurement in international news media and with our definition of communities, and found this to be correlated more with geographical factors than political leaning. The resulting graph, thresholded at communities having a Jaccard coefficient greater to 0.03, is shown in Figure 1. We observe clear clusters for UK-, USA-, India-, and Australian-based media. Community overlaps vary widely, with 90% of the news sources having between 2% and 34% of their community shared with at least one other source. The exclusivity of the community of a news source is to a large extent independent from the community size ($r^2 = 0.14$).

The complete list of data sources, including details on their number of articles, their number of followers, community sizes, and example stories, are included in the *Supplementary Material*.

4. SELECTION BIASES

Selection bias is also known as *gatekeeping*. Printed media has space constraints, whereas radio and television broadcast have temporal constraints. These force editors to routinely take decisions about which (out of potentially hundreds or thousands of news stories) to cover. The Web allows for more latitude, but selectivity is still present.

4.1 Prolificacy and exclusivity

To place selection biases in context, we first study the quantity of stories in online media, and the extent to which those stories are exclusive to one specific news source.

We first compare news sources and their communities in terms of their overall prolificacy. The number of articles each source publishes during a 2-week period is typically in the low hundreds but can reach up to a few thousand in certain cases. We find that the number of different stories a source publishes is correlated ($r^2 = 0.83$) with the number of articles (distinct URLs) that are published, i.e. the more articles a news source posts, the more likely they are to cover a story. Looking at communities, larger communities tend to post[9] more stories ($r^2 = 0.73$), and communities post about 2-3 times more stories than the news sources they follow.

Next we observe to what extent the content posted by a news source is unique. A sizable fraction of the English content in news media is produced by agencies. According to [17], "only four organizations do extensive international reporting (Reuters, AP, AFP, BBC), a few others do some international reporting (CNN, MSN, New York Times, Guardian) and most do no original international reporting." For each article i, we compute its exclusivity E_i as $\mathrm{E}_i = 1 - \max(\mathrm{sim}(i, j))$, where sim is the cosine similarity with TF.IDF weighting and the maximum is taken across all articles $j \neq i$. We find that Associated Press (AP) and other agencies have the most content that is not exclusive. The Economist, Newsweek, and other magazines have the largest amount of original stories; they tend to carry a smaller number of stories, but most of their content is exclusive. Exclusivity seems to be weakly correlated (negatively) with the number of stories each media covers ($r^2 = -0.4$), suggesting that in news online being prolific does not necessarily require having more original content.

4.2 Selection bias and prominence

A major factor affecting the selection of stories is their relative importance [7]. We measure the *prominence* of each story, which corresponds to the fraction of news sources that has at least one article about the story. As in [4], this number ranges from a maximum of 1.0 if the story is in the N news sources in a sample, to a minimum of $1/N$ if it is only in 1 of them.

Different news sources may have different policies for the selection of stories. For instance one news media may want to cover only the top stories of the day, while another may want to include a number of minor/niche ones. Indeed, we observe in practice a wide range of prominence distributions across stories published by online news media. In general magazine-type of media such as The Economist or Newsweek tends to focus on stories of high prominence, and covering

[9]A community is said to post a story whenever at least one of its members posts on Twitter the URL of one article belonging to a story.

less stories is correlated with having larger average prominence ($r^2 = -0.51$).

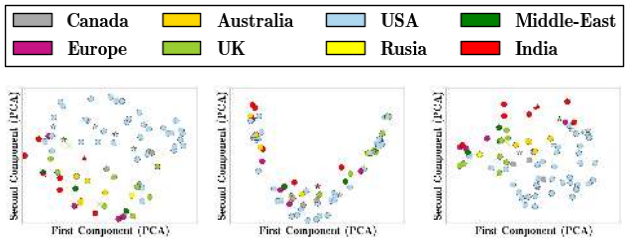
Social media, in principle, should allow a broader selection of stories, including niche ones. In general the prominence of stories posted by social media communities is significantly smaller than the prominence of the news media source they follow. In particular, stories of large prominence are not posted as often by social media users. Two factors may contribute to this. First, given that social media users do not need to appeal to a broad audience, they may have a stronger preference for niche content. Second, saturation effects have been observed in Twitter; [20] demonstrated that the probability that a user posts something initially increases with the number of exposures to it, but then drops. Both factors may contribute to observing more stories of smaller prominence in social media.

Statistics for each data source in terms of exclusivity, as well as details on our analysis of prolificacy and examples of prominence distribution for some media can be found in the *Supplementary Material*.

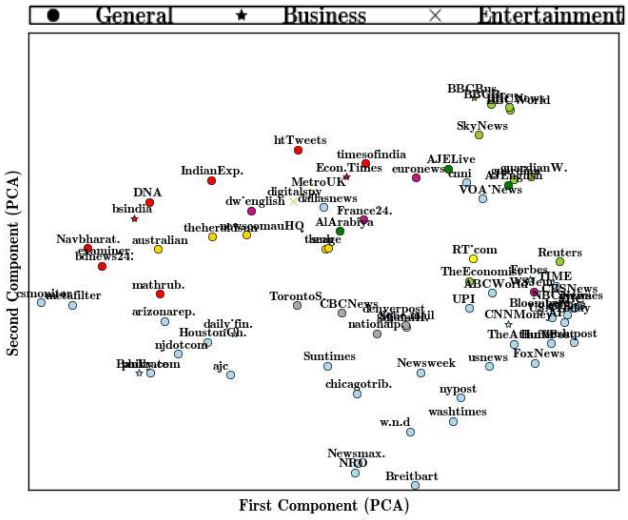
4.3 Selection bias and geography

Next we compare news media according to the overlap in the stories they post, measured by computing for every pair of media the Jaccard coefficient of their sets of stories. In order to visualize this similarity matrix, we project it in two dimensions using Principal Component Analysis (PCA).

The result is shown in Figures 2(a) for news sources and 2(b) for social media communities. For online news sources, there is a clear separation between US-based media and the rest. In social media communities, there is much more mixing of different regions. This means that US-based news media tends to agree in their selection of stories, while their communities in social media are interested in a more diverse range of issues. However, as shown next in Section 5.2, they both tend to be more geographically homogeneous when looking at the amount of attention they devote to stories.



(a) Selection bias in sources (b) Selection bias in communities (c) Coverage bias in sources (#words)



(d) Coverage bias in communities (#tweets)

Figure 2: Similarity according to selection bias in (a) and (b). Similarity according to coverage bias in (c) and (d) (Best seen in color.)

5. COVERAGE BIAS

Coverage bias is a preference for giving more airtime, space, or attention to some issues in contrast to others [5]. While two news media may publish articles about the same story, it might be the case that one gives the story much more attention than the other.

5.1 Measuring coverage bias

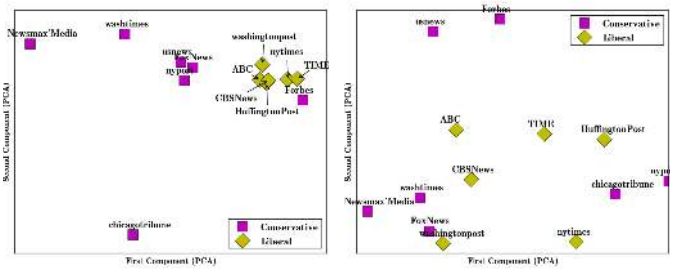
There are several ways in which the distribution of attention given to stories can be quantified. In each news source, we can look at the *length* of the articles covering the story, counting words and adding across multiple articles of the same story, when necessary. In social media, we count the number of tweets containing links to articles on a given story. Communities tend to be influenced by the news media source they follow. A story prominently displayed and promoted by traditional news sources should obtain a larger number of social media reactions from its community. In addition to observing the distribution of coverage of stories, we can also quantify coverage biases in the treatment of different people, by measuring the distribution of number of mentions per person across different media. These distributions are compared by using the Jensen-Shannon (JS) divergence between them for each pair of news sources, and for each pair of social media communities. Coverage bias by story words and by people mentions are somewhat correlated ($r = 0.68$). Interestingly, news media coverage as measured by the length of the stories is correlated with the one observed in the distribution of social media reactions across all communities ($r = 0.84$), but not so correlated with the distribution of tweets in each media's community ($r = 0.40$). This is in agreement with classical results by [13]; the importance given by people to different issues tends to be more correlated with media as a whole than with the specific media source(s) each person follows.

More details on the correlations between selection and coverage bias can be found in the *Supplementary Material*.

5.2 Coverage bias, geography, and politics

We measure the extent of coverage bias with respect to geographical regions and partisan politics.

The distribution of the coverage of different stories, as measured in terms of number of words, is strongly correlated with geographical regions, as shown in Figure 2(c). The same happens if we look at the distribution of tweets given to different stories by communities, depicted in Figure 2(d). In both cases, the geographical biases are more evident than when measured using selection-bias metrics. This means that, as expected [7], news media tends to write articles about the country/region where they are based, and



(a) Coverage bias (tweets) and partisan politics

(b) Coverage bias (stories) and partisan politics

Figure 3: Coverage bias and politics.

that those articles tend to be longer and more frequently tweeted by their social media communities. The same geographical biases have been observed online with respect to search queries [23], indicating that users are interested in what is happening around them, and what is happening to those around them.

We observe that the amount of attention social media communities dedicate to stories follows geographical regions more closely than their selection of stories (compare Figures 2(d) and 2(b)). The relative diversity in terms of selection can be explained by considering the broad range of stories that each community covers, and the low prominence of those stories. The similarities in terms of coverage are aligned with previous findings [21], in the sense that while anyone in social media can propose new ideas (news stories in our case), only few ideas succeed in getting enough attention. Therefore, although communities talk about a broad range of news, they spend most of their time in a few of them, behaving similarly to traditional news sources.

Next, we use a list of USA-based news sources classified by political party from [2], considering six conservatives (Chicago Tribune, Fox, Forbes, NY Post, Newsmax, U.S. News and Washintong Times) and five liberals (ABC, CBS, NY Times, Huffington and Washington Post). We found a strong correlation between political leaning and measures of coverage of stories (this was not the case when using measures of selection of stories). This is depicted in Figure 3 where we project in 2 dimensions the matrix of JS-divergence between the coverage of media sources. We can see that the distribution of tweets per story follows partisan lines closely, whereas the distribution of story lengths also exhibits the same bias, but not as clearly as in social media. The fact that these biases are stronger in social media than in traditional online news may be due to traditional media attempting to strive for an ideal of objectivity that social media users may not aspire to. This agrees with results in Section 6 where we show that in social media the language seems to be more strongly opinionated.

6. STATEMENT BIAS

Statement bias has been used to describe a tendency towards using more favorable statements to refer to one political party at the expense of another [5]. In this section, we study such statements with respect to people in the news, by using sentiment analysis to determine the emotional valence (positive or negative) of expressions in which people are mentioned. Sentiment analysis has shown to be a valuable tool to study emotions expressed in text, e.g. [16] looked at the relationship between tweet sentiments and polls in order to examine how the sentiments expressed in Twitter can be used as political or economic indicators.

Table 1: People selected in our sample for statement bias, including number of articles mentioned each person and the boundaries of the 1st (more negative) and 4th (more positive) quantiles of valence scores for mentions of them.

Name	Number of articles	Sentiment Quartiles Media 1st	Media 4th	Community 1st	Community 4th
Barack Obama	3,241	5.20	6.26	4.82	6.41
Margaret Thatcher	986	5.15	6.47	4.00	6.41
Kim Jong-un	984	5.60	6.62	4.53	6.45
John Kerry	850	5.10	6.32	4.89	6.39
David Cameron	377	5.82	6.25	4.31	6.48
Julia Gillard	303	5.28	6.62	4.93	6.83
Vladimir Putin	291	5.13	6.31	4.89	6.62
Ban Ki-moon	271	4.79	5.33	4.32	5.52
Bashar al-Assad	260	4.69	6.06	4.49	6.00
Hugo Chavez	211	4.86	6.32	4.55	6.39

We sort all named entities of type *person* in our dataset by decreasing number of mentions. The top of this list is dominated by politicians of international relevance, so we focus on a group of 10 present and former heads of state (plus the Secretary of State of the USA and the Secretary General of the United Nations, both prominently mentioned in our sample). We merge entities referring to the same person, e.g. "Obama" and "Barack Obama". The list is shown in Table 1.

We analyze the sentiments used in relationship to persons in our list, using the dictionary provided by the Affective Norms for English Words (ANEW) [3]. We use the *valence* dimension, which assign to each word a number from 1 (if it evokes sadness, dissatisfaction and despair) to 9 (if it evokes happiness, satisfaction, hope). The average sentiment for a person on a news source is the micro-average of sentiments in all the statement on all the articles mentioning that person. The average sentiment for a person on a news community is the average of sentiments in all the tweets posted by members of that community mentioning the person (according to an exact string match of the last name, which due to their prominence and as verified in our sample, is almost invariably a reference to the correct person).

Statistics about the distribution of the valence of sentiments are shown in Table 1. In general the lower quantile (more negative sentiments) is significantly lower in social media communities when compared to news sources, across all the persons included in our sample. Other anecdotal examples can be found in the *Supplementary Material*.

7. RELATED WORK

Reputable news reports are expected to be objective; their role is to inform people about what is happening in the world. It is however known that bias exists in the way news is reported by the media e.g. [9]. Most attempts to detect bias are still done on a small scale, with news manually examined and coded. In this paper, we automatically process a large sample of articles published in several in-

ternational English-speaking online news sites, avoiding the manual coding of articles.

Recently, [12], examined the presence of coverage bias in mainstream news and blogs. They show that overall only a slight slant in terms of party and political leaning could be observed. However, this changes during important political events, such as a mid-term election. We also observe some slight political slants, but the geographical bias seems more prevalent.

In this paper, we focus on three types of bias, selection (gatekeeping), coverage, and statement bias, all studied before. For instance, [5] found no substantial bias on US magazines, but a small coverage bias was detected on US television. With respect to coverage bias, [22] studied Dutch and German television and observed that top leaders such as chancellors or prime ministers get a substantially larger number of mentions than the second most mentioned politicians. Finally, for statement bias, [14] found large differences in the sentiment polarity with which US candidates were treated.

8. CONCLUSIONS

We studied the presence of bias in online news and the social media communities that surround them. Our results support the following high-level conclusions.

In international news media, selection and coverage biases seem more correlated with geographical variables than political leaning. In other words, online news sources in a given geographical region tend to select the same stories, and write articles of similar relative length. Social media follows the same pattern, with the communities of media in a region showing a similar proportion of tweets to stories.

Political bias is evident in social media, in terms of the distribution of tweets different stories receive. This distribution is more closely related among communities of news media having the same political leaning (at least in the US for which we could obtain political leaning information). Political bias is also observable in terms of the distribution of length of articles on different stories in traditional media, but to a smaller extent than in social media. Statement bias is also evident in social media. In a sample of statements referring to world leaders, we find that the language used in social media is more opinionated, and often more negative, than the one used in traditional news media.

In terms of editorial policies regarding the prominence or importance of stories covered, we observe that magazine-type of news (which in general covers less articles and stories) tends to select stories of high prominence and produce exclusive content. More importantly, we observe that social media tend to be much more focused in niche content than traditional news media. In particular, very prominent stories seem to receive much less attention in social media than in traditional news sources.

Reproducibility. The dataset used in this paper is available upon request for research purposes.

Acknowledgements. Diego Saez-Trumper was supported by the HIPERGRAPH project (TIN2009-14560-C03-01) from the Spanish Economy and Competitiveness Ministry. This work was carried out as part of Diego Saez-Trumper internship at QCIR. The authors with to thank colleagues in Al Jazeera English for valuable discussions, and Noora Al-Emadi for her help with the processing of data.

References

[1] J. An, M. Cha, P. K. Gummadi, and J. Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *ICWSM*, 2011.

[2] J. An, D. Quercia, and J. Crowcroft. Beyond selective exposure in social media: Concrete evidence for partisan sharing. 2013.

[3] M. Bradley and P. Lang. Affective norms for English words (ANEW). Technical Report C-1, The Center for Research in Psychophysiology, U. of Florida, 1999.

[4] C. Castillo, G. De Francisci Morales, M. Mendoza, and N. Khan. Says who? Automatic text-based content analysis of television news. In *MNLP*, San Francisco, CA, USA, 2013.

[5] D. D'Alessio and M. Allen. Media bias in presidential elections: a meta-analysis. *J.of Communication*, 50(4):133–156, 2000.

[6] S. DellaVigna and E. Kaplan. The fox news effect: Media bias and voting. *The Quarterly J.of Econ.*, 122(3):1187–1234, 2007.

[7] J. Galtung and M. H. Ruge. The structure of foreign news. *J.of Peace Research*, 2(1):64–91, 1965.

[8] C. J. Glynn, S. Herbs, G. J. OKeefe, , and R. Y. Shapiro. *Public Opinion*. Boulder CO: Westview Press, 1999.

[9] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly J.of Econ.*, 120(4):1191–1237, 2005.

[10] B. G. Knight and C.-F. Chiang. Media bias and influence: Evidence from newspaper endorsements. Working Paper 14445, National Bureau of Economic Research, October 2008.

[11] J. Lehmann, C. Castillo, M. Lalmas, and E. Zuckerman. Finding news curators in twitter. In *SNOW*, May 2013.

[12] Y.-R. Lin, J. P. Bagrow, and D. Lazer. More voices than ever? quantifying media bias in networks. In *ICWSM*, 2011.

[13] M. E. Mccombs and D. L. Shaw. The Agenda-Setting function of mass media. *The Public Opinion Quarterly*, 36(2):176–187, 1972.

[14] J. S. Morris and P. L. Francia. From network news to cable commentary: The evolution of television coverage of the party conventions. In *State of the Parties Conf.* U. of Akron, 2005.

[15] R. Nallapati, A. Feng, F. Peng, and J. Allan. Event threading within news topics. In *CIKM*, N. York, NY, USA, 2004.

[16] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 2010.

[17] C. Paterson. News agency dominance in international news on the internet. *Papers in International and Global Communication*, (01/06), May 2006.

[18] S. Reese and P. Shoemaker. Mediating the message: Theories of influence on mass media content. *Journalism Quarterly*, 74:2, 1996.

[19] S. M. Rivolta. *Strategic Maneuvering and Media Bias in Political News Magazine Opinion Articles*. PhD thesis, U. of Amsterdam, 2011.

[20] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proc.of WWW*, 2011.

[21] D. Saez-Trumper, G. Comarela, V. Almeida, R. Baeza-Yates, and F. Benevenuto. Finding trendsetters in information networks. In *Proc. of KDD*, 2012.

[22] K. Schoenbach, J. De Ridder, and E. Lauf. Politicians on TV news: Getting attention in dutch and german election campaigns. *European J.of Political Research*, 39(4):519–531, 2001.

[23] E. Yom-Tov and F. Diaz. Out of sight, not out of mind: on the effect of social and physical detachment on information need. In *SIGIR*, 2011.

[24] B. Zelizer. Journalists as interpretive communities. *Critical Studies in Mass Communication*, 10(3):219–237, 1993.

[25] Y. Zhai and M. Shah. Tracking news stories across different sources. In *ACM Multimedia*, 2005.

Supplementary Material – Social Media News Communities: Gatekeeping, Coverage, and Statement Bias

This document contains supplementary materials to the paper "Social Media News Communities: Gatekeeping, Coverage, and Statement Bias."

1 Data processing

Table 1 includes the list of data sources used in the paper, as described in Section 3. For each data source, we have included its name, Twitter accounts, the number of articles we collected, and the size of its community, along with the Country/Genre.

Table 1: News sources including name, Twitter accounts, number of articles, size of community, country where the media is based and genre.

Name	Twitter accounts	Arts.	Comm.	Country/Genre	Name	Twitter accounts	Arts.	Comm.	Country/Genre
ABC	@ABC	565	478	USA/Gen.	Houston Chron.	@HoustonChron	326	40	USA/Gen.
ABC	@ABCWorldNews	140	69	USA/Gen.	Hindustan Times	@htTweets	583	118	India/Gen.
Atlanta J.-C.	@ajc	166	53	USA/Gen.	Huffington Post	@HuffingtonPost	2,545	4,406	USA/Gen.
Al Jazeera	@AJELive	34	189	Qatar/Gen.	Indian Express	@IndianExpress	122	57	India/Gen.
Al Jazeera	@AJEnglish	667	645	Qatar/Gen.	LA Times	@latimes	1,151	391	India/Gen.
Al Arabiya	@AlArabiya	284	151	Dubai/Gen.	Mathrubhumi	@mathrubhumi	63	7	USA/Gen.
AP	@AP	454	649	USA/Gen.	Metafilter	@metafilter	92	6	India/Gen.
Arizona Rep.	@arizonarepublic	124	35	USA/Gen.	Metro	@MetroUK	485	75	USA/Gen.
The Australian	@australian	127	10	Australia/Gen.	Miami Herald	@MiamiHerald	285	194	UK/Gen.
Bankrate	@Bankrate	228	18	USA/Bus.	National Post	@nationalpost	1,240	184	USA/Gen.
BBC	@BBCBreaking	44	606	UK/Gen.	Navbharat Times	@NavbharatTimes	479	10	Canada/Gen.
BBC	@BBCBus.	160	1,045	UK/Bus.	NBC	@NBCNews	638	324	USA/Gen.
BBC	@BBCNews	735	3,512	UK/Gen.	News Austrlia	@newscomauHQ	395	77	Australia/Gen.
BBC	@BBCWorld	363	2,150	UK/Gen.	Newsmax	@Newsmax_Media	320	49	USA/Gen.
bdnews24	@bdnews24com	724	146	Bangladesh/Gen.	New Jersey	@njdotcom	253	68	USA/Gen.
Bloomberg	@BloombergNews	822	724	USA/Bus.	National Review	@NRO	174	87	USA/Gen.
Breitbart News	@BreitbartNews	748	1,111	USA/Gen.	NY Post	@nypost	1,075	270	USA/Gen.
Business Std.	@bsindia	234	9	India/Bus.	NY Times	@nytimes	801	1,638	USA/Gen.
CBC	@CBCNews	246	148	Canada/Gen.	Philly.com	@phillydotcom	268	22	USA/Gen.
CBS	@CBSNews	599	435	USA/Gen.	Reuters	@Reuters	824	775	UK/Gen.
Chicago Trib.	@chicagotribune	224	117	USA/Gen.	Russia Today	@RT_com	881	1,203	Rusia/Gen.
CNN	@cnni	575	235	USA/Gen.	Sky News	@SkyNews	410	340	UK/Gen.
CNN	@CNNMoney	380	90	USA/Bus.	Sydney Morn. Hrld.	@smh	439	171	Australia/Gen.
Chr. Sci. Monit.	@csmonitor	685	6	USA/Gen.	Sunday Times	@Suntimes	356	80	USA/Gen.
Daily Finance	@daily_finance	4,531	56	USA/Bus.	The Age	@theage	558	119	USA/Gen.
Dallas News	@dallasnews	233	45	USA/Gen.	The Atlantic	@TheAtlantic	584	346	USA/Gen.
Denver Post	@denverpost	760	96	USA/Gen.	The Economist	@TheEconomist	153	116	UK/Gen.
Digital Spy	@digitalspy	540	112	UK/Ent.	Herald Sun	@theheraldsun	1,170	20	Australia/Gen.
DNA	@dna	376	25	India/Gen.	TIME	@TIME	455	614	USA/Gen.
Deutsche Welle	@dw_english	521	16	Germany/Gen.	Times of India	@timesofindia	3,753	404	India/Gen.
Economic Times	@EconomicTimes	1,782	69	India/Bus.	Toronto Star	@TorontoStar	717	157	Canada/Gen.
EuroNews	@euronews	170	31	France/Gen.	UPI	@UPI	1,165	98	USA/Gen.
Examiner	@examinercom	95	28	USA/Ent.	USA Today	@USATODAY	408	196	USA/Gen.
Forbes	@Forbes	2,062	1,484	USA/Bus.	U.S News	@usnews	797	65	USA/Gen.
FOX	@FoxNews	599	1,156	USA/Gen.	Voices of America	@VOA_News	520	81	USA/Gen.
France 24	@France24_en	249	101	France/Gen.	Washington Post	@washingtonpost	534	482	USA/Gen.
Global Mail	@globeandmail	2,348	194	Canada/Gen.	Washington Times	@washtimes	488	184	USA/Gen.
Guardian	@guardian	650	1,089	UK/Gen.	WND News	@worldnetdaily	287	268	USA/Gen.
Guardian World	@guardianworld	945	749	UK/Gen.	Wall St. Journ.	@WSJeurope	4	312	Europe/Gen.

Examples of stories found using the method in Section 3.2 are listed on Table 2.

Table 2: Examples of stories found in our dataset.

Story (date 2013)	Arts.	Sources	Tweets	Comms.
Former UK PM Margaret Thatcher dies (Apr 9th)	309	67	49K	51
Tensions in the Korean peninsula (Apr 11th)	402	62	53K	48
Presidental elections in Venezuela (Apr 13th)	18	16	1.2K	52

The correlation of $r^2 = 0.74$ described in Section 3.3 is supported by Figure 1.

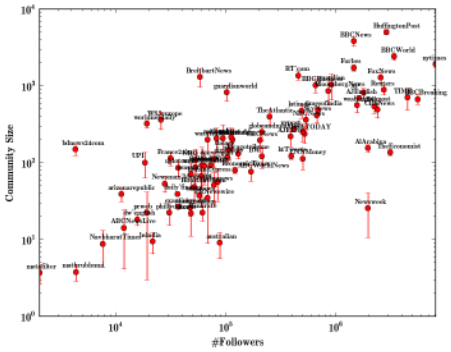


Figure 1: Number of followers of the corporate account of each news source versus size of the (active) community of each source. These quantities are correlated ($r^2 = 0.74$), but communities are 3 orders of magnitude smaller than the number of followers.

2 Selection biases

The correlations described in Section 4.1 are a sub-set of those found in Table 3.

Table 3: Correlation between quantities of articles and stories and community sizes. The table shows the Person correlation between the logarithm of the metrics.

	MA	MS	TF	TS	TP	Median
MA. Articles in media	-					454
MS. Stories in media	0.83	-				98
TF. Followers of media	0.36	0.49	-			105 K
TS. Stories in comm.	0.29	0.48	0.80	-		263
TP. People in comm.	0.23	0.38	0.74	0.73	-	117

The fraction of exclusive stories per each data source is depicted in Figure 2

The distribution of prominence of stories, as described in Section 4.2, is depicted for a set of example news media in Figure 3.

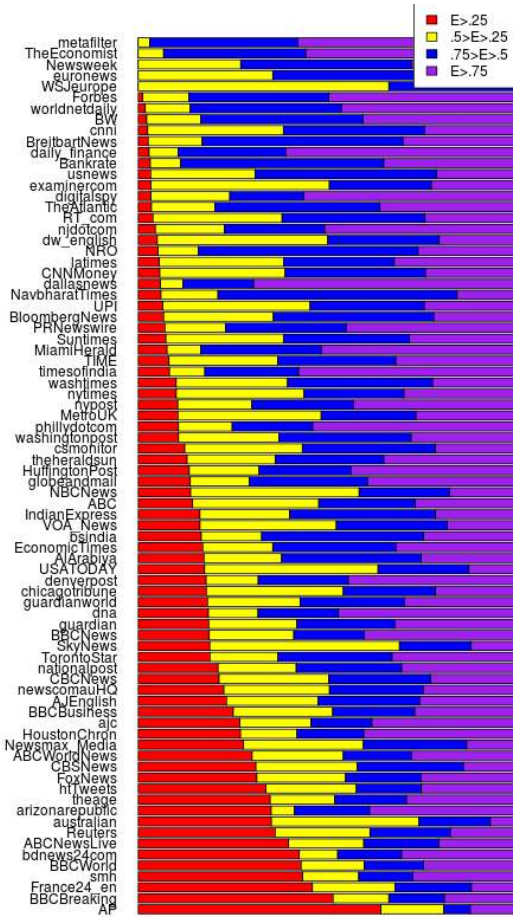


Figure 2: Exclusivity of stories. The fraction of exclusive articles varies widely across online news sources. $E > 0.75$ are articles that are basically unique to one source, while $E < 0.25$ appear in two or more sources with minor differences.

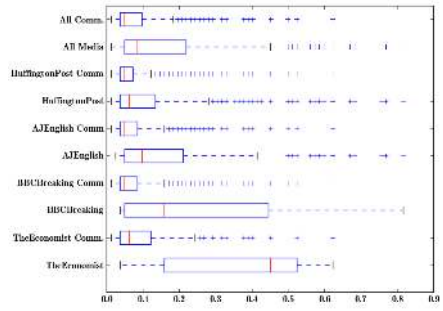


Figure 3: Prominence of stories. Social media news communities share more often niche content, and in general do not share stories having high prominence. This is in sharp contrast with online news media sources, which prefer stories having in general higher prominence than the social media ones.

3 Coverage bias

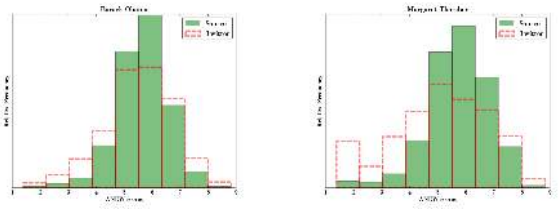
Table 4 presents more correlations between selection and coverage biases, as described in Section 5.1.

Table 4: Correlation between selection and coverage biases as presented by news media sources. Correlations above 0.8 between different biases are shown in boldface.

	SS	SP	CS	CP	CT	CT'
SS. Selection bias by stories	1	-	-	-	-	-
SP. Selection bias by people	0.66	1	-	-	-	-
CS. Coverage bias by story words	**0.81**	0.63	1	-	-	-
CP. Coverage bias by people mentions	0.68	**0.94**	0.68	1	-	-
CT. Coverage bias by tweets (all)	0.65	0.53	**0.84**	0.60	1	-
CT'. Coverage bias by tweets (community)	0.30	0.22	0.40	0.28	0.37	1

4 Statement bias

Figure 4 supplements observations on Section 6 with anecdotal observations about the death of Margaret Thatcher on April 8th, 2013: social media users were described as "dancing on the grave" of the former UK Prime Minister, while traditional news media was much more circumspect.[1]



(a) Barack Obama (b) Margaret Thatcher

Figure 4: Distribution of valence scores in mentions in media sources and social media communities, for two politicians in our sample. Mentions in social media exhibit a wider range of expression and a tendency towards more negative sentiments.

[1] http://www.independent.co.uk/voices/comment/margaret-thatchers-death-newspapers-pay-respect-while-social-media-dances-on-her-grave-8565679.html