Can Trending News Stories Create Coverage Bias? On the Impact of High Content Churn in Online News Media

Abhijnan Chakraborty*† Saptarshi Ghosh† Niloy Ganguly* Krishna P. Gummadi†

†Max Planck Institute for Software Systems, Germany
*Indian Institute of Technology Kharagpur, India

ABSTRACT

Online news media sites are presently the primary sources of news for a large number of users world-wide. The content published by media sites have high temporal churn as they tend to focus on recommending recent / breaking news. Analyzing data from a popular online *mass media* site **nytimes.com** as well as a popular online *social media* site **twitter.com**, we show that the resulting churn in popular content can lead to temporal coverage biases in the stories that are consumed by a user, based on the time of the day when the user accesses the media sites.

1. INTRODUCTION

Online news media sites, be they mass media sites like New York Times (`nytimes.com`) or CNN (`cnn.com`) or social media sites like Facebook or Twitter, are emerging as the primary (and frequently *only*) sources of news for a large and rapidly growing fraction of people world-wide. The number of users receiving news via traditional offline methods, e.g., via print newspapers and weeklies, are in steep decline [14]. A recent survey by the Pew Research Center [5] found that around 48% of American Internet users got politics news on social media sites like Facebook, almost as many as those that got such news from local television channels.

Due to the round-the-clock (24/7) nature of online news and the need to keep their audience coming back to their site[1], online news sites today are emphasizing *recent* news stories over *relevant or important* news stories. A user visiting the sites is often recommended news stories that became popular only recently, often within the last hour. Social media sites like Facebook and Twitter update "Trending Topics" every 10 to 15 minutes, while mass media sites like nytimes.com update the "Top Stories" displayed prominently at the top of their home pages constantly throughout the day. Such constant churn in top recommended content incentivizes users to visit the sites repeatedly and helps disseminate breaking news stories rapidly. However, excessive emphasis on recency in online news media risks overloading users with unnecessary information [11] and raises the common refrain *why is this news?* [17].

In this paper, we investigate a previously overlooked concern with the constant churn in popular content published by online media sites. Specifically, when users browse the media sites at different times of the day, they might receive very different top news stories. A user who habitually browses *nytimes.com* at 9 AM every day might see a very different topical coverage of news-stories, as compared to another user who browses the site at 9 PM every day. For instance, the breaking stories at 9 AM might be predominantly covering Business / Finance news, while the breaking stories at 9 PM might be largely covering Sports news. A recent survey by the American Press Institute [1] found that a majority 63% of users prefer to read news at some specific point of time in a day - either in the morning, in the evening, in the afternoon or right before bed.

Over extended periods of time, such differences in the browsing pattens of individual users might lead to substantial biases in *topical coverage* of news stories consumed by individual users. Using the terminology introduced in [8], we refer to the distribution of topics (e.g., Sports, Business) of news stories consumed by a user as the user's *information diet*. So the central question we investigate in this paper is: *Do temporal recommendations in online news media lead to biased topical coverage in their users' information diets?*

To address the above question, we analyze extensive data gathered from two popular online news media sites: (i) Twitter, a social media site, and (ii) New York Times (nytimes.com), a mass media site. Our analysis reveals high churn in top recommended content in both sites. We also find considerable diurnal variation in coverage of news stories related to specific topics. For example, news stories related to Economy or Science are recommended predominantly at certain times of the day and not at other times. We further show that such churn in top news stories can induce a significant bias in the topical coverage of a user's diet, depending on the diurnal browsing pattern of the user.

In summary, our study highlights the potential for coverage bias in today's online news media sites. Through this paper, we seek to raise awareness about the problem and call on the research and news media community to explore practical solutions to the problem.

2. BACKGROUND AND RELATED WORK

Coverage Bias of News Media: Several works in media studies [3, 7] analyzing the news articles on the issues like political bias, fairness and accuracy of the presented facts. Moreover, several media watchdog groups like FAIR (`fair.org`) monitor the news media organisations for any bias or misinformation introduced in their stories. However, in this paper, we show how temporal recommendation introduces unintended bias in the content stream – which otherwise goes unnoticed if analyzed on individual stories.

[1] Similar to most online websites, many online news media sites are also predominantly funded by their users watching advertisements on their sites.

Recency vs. Relevancy Debate: Present approaches on designing content recommendation systems are putting increasing emphasis on the recency and realtimeness of content. For example, [9] presents the design of a time-aware content recommendation system. [10] proposes a framework to detect breaking news, trending events from online social media in realtime. Due to this focus on recency, content recommendation systems are considering more temporal parameters to rank documents. As a result, there is a growing concern over the relevance (or long-term importance) of the content recommended by such systems and many users view such recommended contents as potentially waste-of-time information [11, 15]. Although this debate on recency versus relevance is going on for some time, to the best of our knowledge, we are the first to point out that the media systems' emphasis on recency are creating a filtering effect on users, who are potentially missing out on certain types of information depending on their browsing habits.

Diurnal patterns in user browsing behavior: In this paper, we argue that browsing patterns of users impacts their consumption of news. Intuitively, we would expect different users around the world (in different timezones) to access media websites at different times of the day. Prior studies have observed strong diurnal patterns in accessing messages and applications on Facebook [6], and in the content generation of blog posts, bookmarks, as well as answers in Q&A websites [4]. A study by Benevenuto *et al.* [2] analyzed browsing patterns of tens of thousands of users, using click-stream data from a social network aggregator site. They observed that most of the users accessed the sites only a few times and during certain periods of a day.

Personalization and Filter Bubbles: Researchers have investigated the issue of 'Filter Bubbles' arising out of personalized content recommendation to suit user interests [13]. According to [13], as users get recommendations based on their profiles (locations, past click behaviors, search histories), they gradually become separated from the type of information that diverts from their past behavior and eventually get isolated in their own cultural or ideological bubbles.

Interestingly, in our study, we show that filtering effects can arise even in the absence of personalization. All the content recommendation systems we investigate in this work broadcast the same recommended contents to all users; therefore, there is no personalization involved. Rather the filtering effects arise inadvertently due to the complex correlations between user browsing times and high churn in the popular content recommendations.

3. COVERAGE BIAS IN USER DIETS

As stated in Section 1, we want to investigate whether temporal recommendations in online news media sites lead to biased topical coverage in the users' information diets. We attempt to address this question in the context of two content recommendation systems deployed in two popular online news media – (i) 'Trending Topics' in the online social media site *Twitter* (`www.twitter.com`), and (ii) 'Top Stories' in the online mass media site *New York Times* (henceforth referred to as NYTimes) (`www.nytimes.com`).

3.1 Datasets Gathered

Twitter Trending Topics: Twitter periodically publishes 10 trending topics ('trends' in short) to help its users find the currently most popular topics of discussion in Twitter. According to the official Twitter blog [12], trends are the keywords whose usage at a certain time instant dramatically increases compared to its earlier usage. Twitter users can specifically choose to receive either worldwide trends or trends specific to a particular geographical region. We collected US trends using the Twitter API[2] at 15-minute intervals, during January – July, 2014 (7 months). Overall, we collected 21, 180 distinct trends during this period.

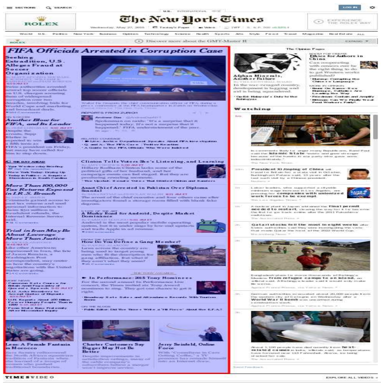

SECTIONS
SEARCH
LOG IN
ROLEX
The New York Times
FIFA Officials Arrested in Corruption Case
Seeking Extradition, U.S. Alleges Fraud at Soccer Organization
Another Blow for FIFA, and Its Leader
Clinton Tells Voters She's Listening, and Learning
More Than 100,000 Tax Returns Exposed in I.R.S. Breach
Axact Chief Arrested in Pakistan Over Diploma Scandal
A Murky Road for Android, Despite Market Dominance
Trial in Iran May Be About Leverage More Than Justice
How Do You Define a Gang Member?
In Performance: 2015 Tony Nominees
Lens: A Female Fantasia in Morocco
Charter Customers Say Bigger May Not Be Better
Jerry Seinfeld, Online Force
Afghan Minerals, Another Failure
Ethics for Authors in China
Watching
TIMES VIDEO
EXPLORE ALL VIDEOS

Figure 1: Snapshot of the NYTimes homepage. The 'Top Stories' section is highlighted in bluish grey.

NYTimes Top Stories: To enable the users to take a quick look at the most important news-stories at a particular time, NYTimes provides around 20 articles in the 'Top Stories' section in the website. As shown in the Figure 1, this section consumes a very prominent screen space in the NYTimes home page. Hence, it can be safely assumed that most visitors of NYTimes read the articles listed in the 'Top Stories' section. Using the NYTimes developer API[3], we collected the top stories at 5-minute intervals, over 40 days during April – May, 2015. Overall, we collected 3, 050 distinct top stories.

3.2 Users' News Consumption Habits

We begin by studying news consumption habits of users. A recent survey by the American Press Institute [1] found that while 33% of users consumed news throughput the day (hence, differences in the topical composition of news-stories is less likely to affect them), a majority 63% of the users prefer to read news only during a specific period of a day – either in the morning, or in the afternoon, or in the evening, or right before bed. These users can be exposed to different topical coverages if the composition of news at these times are different.

We further probed users' reading habits on the NYTimes site. There is no easy way of knowing which news-stories a user has read on the site, hence we considered the posting of a comment on a news-article by a user as a proxy for his reading the article. Overall, we collected 885,421 comments posted by 133,189 distinct users during April to August, 2015, and analyzed the commenting behavior of the users.

We divided the users into four groups based on their commenting frequencies, as shown in Table 1. It can be seen that more than 90% of the users – labelled as the *incidental* and *occasional* news consumers – commented only during one or two hours of a day on average. We also looked into

[2] `dev.twitter.com/rest/reference/get/trends/place`
[3] `developer.nytimes.com/docs/top_stories_api/`

Group	Nos. of comments posted	Nos. of users	Nos. of hours active daily
Incidental Users	Less than 10	118,652 (89%)	95% users: 1 hour, 5% users: 2 hours
Occasional Users	10 to 50	11,747 (9%)	90% users: 1 hour, 10% users: 2 hours
Regular Users	50 to 100	1,625 (1%)	78% users: 1 hour, 21% users: 2 hours
News Addicts	More than 100	1,165 (1%)	20% users: 1 hour, 65% users: 2 hours, 15 % users: $\geq$ 3 hours

Table 1: Different groups of users of NYTimes, based on their commenting behavior during April–August 2015.

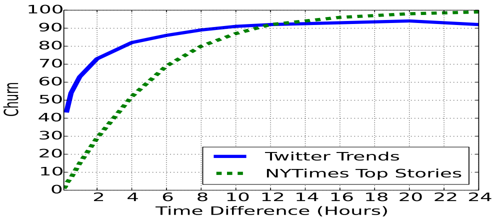


Figure 2: Churn in Twitter Trends (solid blue curve) and NYTimes Top Stories (dashed green curve).

the specific hours at which they post the comments, and found that *most of these users post comments only during 1–3 specific hours of a day*. We also noticed another interesting aspect – out of all the news-stories on which users commented, around 75% were top stories, and this fraction is similar across all groups.

Thus, both our observations and those by the American Press Institute [1] indicate that a large majority of users are incidental / occasional news consumers who access news media only during one or two specific hours of a day, and read mostly the top news-stories. Hence, the problem we study – differences in the topical composition of top news-stories at different hours of a day – is very likely to affect these users (though, it might not affect the few news addicts).

3.3 Churn-Rate of Trending News Stories

Our analysis above shows that a majority of users visit media sites only once or twice in a day. The next step is to understand *the rate at which the popularity of content is changing within such systems*. If the popular content is fairly static, the churn-rate (i.e., the rate at which the set of trending news-stories is changing) will be very low. As a result, users will receive similar information regardless of the time they are visiting the media site. However, if the churn-rate is high, then the users visiting the site at different times of the day will consume very different sets of content.

To compute the churn-rate in a news media (Twitter or NYTimes), we measure the average fraction of *non-overlapping* stories between every pair of trending stories' sets separated by time t in our dataset, where t varies from 15 minutes to 24 hours. Figure 2 shows the churn in both Twitter Trends and NYTimes Top Stories.

For Twitter, on an average, around 45% of the trends change within a gap of only 15 minutes. If we compare two sets of trends at a time difference of 2 hours, the average churn is around 75%. With larger time difference, the churn increases further and remains above 90%.

The churn rate of NYTimes Top Stories is less than the churn rate of Twitter trends; however, as much as 70% of the top stories gets replaced within a six-hour time difference.

Content	Topic Categories
Twitter Trends	Arts-crafts, Automotive, Business-finance, Career, Education-books, Entertainment, Environment, Fashion-style, Food-drink, Health-fitness, Hobbies-tourism, Paranormal, Politics-law, Religion-spiritualism, Science, Society, Sports, Technology
NYTimes Top Stories	Africa, Americas, Arts, Asia Pacific, Baseball, Books, Business Day, DealBook, Economy, Education, Europe, Food, Health, International Business, Media, Middle East, Multimedia/Photos, N.Y./Region, NYT Now, Obituaries, Opinion, Politics, Pro Basketball, Real Estate, Science, Style, Technology, Television, The Upshot, Travel, U.S., Your Money

Table 2: Topical categories for Twitter trends and NYTimes top stories.

These results clearly show that the online media systems' emphasis on 'recency' has made the systems so dynamic that all the content have very fleeting priority – even the set of top (most popular) content in these systems is not remaining the same for more than $1-2$ hours. As a result, two users visiting the media sites at only 6 hours time differences, would receive $70\% - 85\%$ different sets of popular contents. Therefore, the timing at which a particular user is visiting the media becomes immensely important and affects the composition of information received by the user. In the next section, we show how we can use the notion of 'information diet' to characterize the difference in information consumption by the users.

3.4 Temporal Bias in Topical News Coverage

To understand the temporal bias within a particular topic, we need to infer topics of the trending news-stories in Twitter and NYTimes respectively - the methodology of which is elaborated next.

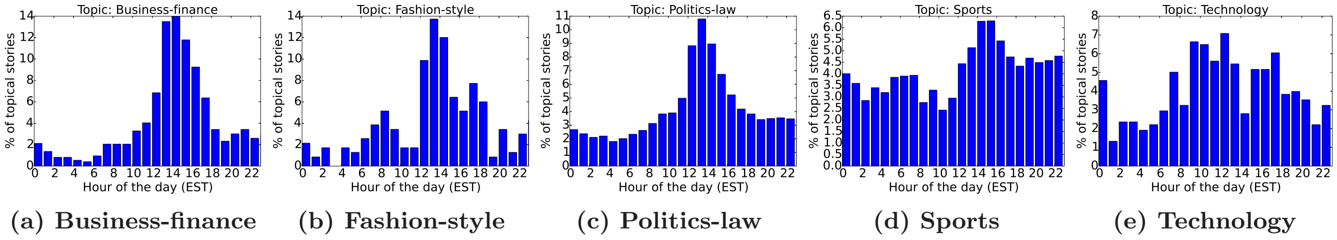
3.4.1 Inferring topics for trending news stories

Inferring the topics of Twitter trends is challenging due to the limited information contained in very short tweets (at most 140 characters) as well as the frequent use of informal language. Several prior studies have attempted to infer topics for trends / keywords in Twitter [8, 16]. In this paper, we use the topic inference methodology recently developed in [8], which infers the topic of a given keyword from the topical expertise of the authors of the tweets containing the keyword. The basic intuition is that if a certain keyword is being posted by many users who are interested in (or experts on) a common topic, then the keyword is very likely to be relevant to the topic. Table 2 (1st row) shows the different topics that are inferred by this methodology for any given trend.

For NYTimes articles, the authors (or editors) assign a particular section (or a subsection) to each article. We consider this section / subsection of an article as its topic. Table 2 (2nd row) lists all major topic categories for NYTimes top stories.

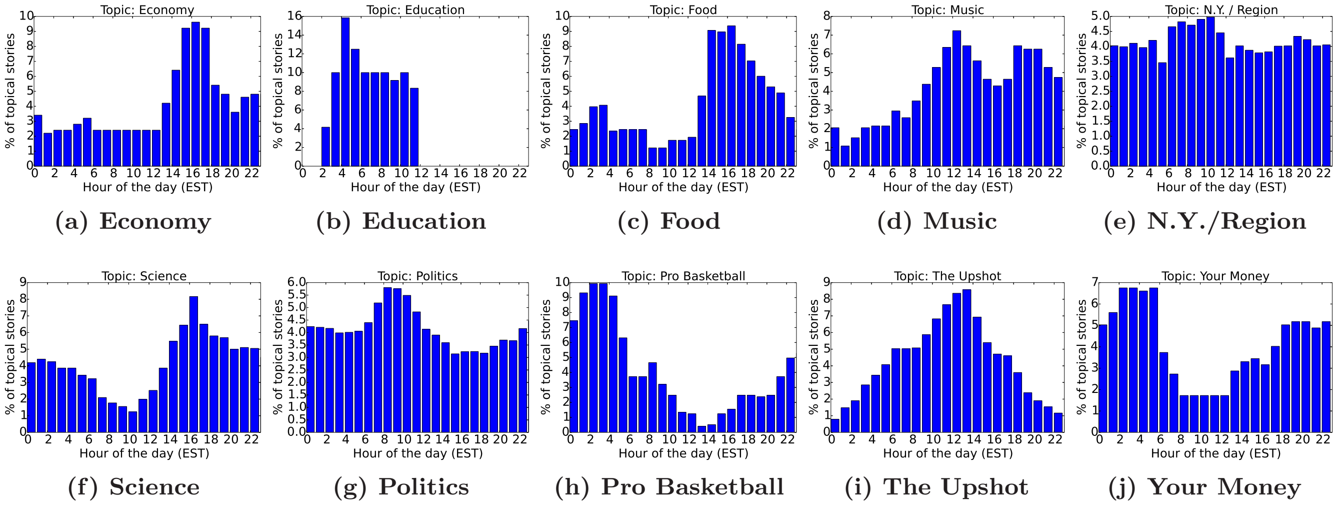
3.4.2 Diurnal variation in topical news coverage

We first check for any temporal bias in the coverage of news-stories on different topics. For this, we consider all news-stories on a particular topic, and compute how these news-



(a) Business-finance (b) Fashion-style (c) Politics-law (d) Sports (e) Technology

Figure 3: Coverage distribution for topics on Twitter Trends



(a) Economy (b) Education (c) Food (d) Music (e) N.Y./Region

(f) Science (g) Politics (h) Pro Basketball (i) The Upshot (j) Your Money

Figure 4: Coverage distribution for topics on NYTimes Top Stories

stories are recommended at different hours of a day. This distribution will be nearly uniform for a topic if that particular topic gets uniform coverage throughout the day.

Figure 3 and Figure 4 show the hourly distribution for different topics, respectively for Twitter trends and NYTimes top stories. We see that, except a few topics which are covered adequately throughout the day (like 'Sports' in Twitter trends and 'Politics' or 'N.Y./Region' in NYTimes), there are huge diurnal variations for most of the topics.

Specially for niche topics like 'Education' in NYTimes (Figure 4(b)), there are several time periods in a day where there is no article on the topic among the top stories. As a result, if some user is interested in a niche topic, she needs to browse the top stories at certain hours to have a higher chance of getting stories on her topic of interest. Similarly, if a user is browsing the site at specific hours everyday, she might be missing the niche topics which do not get recommended at these hours.

3.5 Topical Coverage Bias in User Diets

As different topics get non-uniform diurnal coverage, the pertinent question to ask is *how users' overall information consumption are getting affected by the diurnal variations in coverage*. We use the notion of 'information diet' [8] to address this question. A user's information diet is computed as the topical composition of all content consumed by her. Therefore, if a large fraction of the news-stories consumed by a user is related to a certain topic, her information diet will become biased towards that topic.

To characterize the topical coverage bias, we consider different users who browse the media sites regularly during different hours of a day, e.g., a user who browses a site between 9 A.M. and 10 A.M. every day. We assume for simplicity that a user browsing a media system at a particular hour will read all the content recommended (as top stories, or trending topics) during that one-hour period. Figure 5 and Figure 6 show how different users visiting Twitter and NYTimes at different times will cover different topics in different proportions, and hence have different information diets. The hours shown in the figures are chosen to represent the most likely hours for users to visit media sites according to the American Press Institute Survey [1]. Since a particular topic's contribution to the information diet is different at different hours, users' topical coverage will be dependent on their browsing habits. For example, as shown in Figure 6, a user browsing NYTimes during 5 P.M. – 7 P.M. every day will receive more Science related news, whereas a user browsing during 11 A.M. – 1 P.M. every day will receive more stories from 'The Upshot' section.

We next investigate whether the diurnal variation in the topical coverage affects the users' diets more for certain topics than others. To quantify this, we compute the standard deviation of the topics' contributions to the hourly information diet, and then normalize the deviation by the mean contribution value. Figure 7 and Figure 8 show the result for the Twitter trends and NYTimes top stories. We see that the variation in user's diet, depending on her browsing time, is relatively less for more popular topics (e.g., Politics-law or Entertainment in Twitter, and Politics or U.S. in NYTimes)

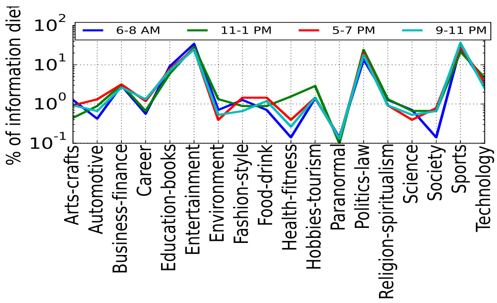


Figure 5: Difference in information diet for some topics in Twitter Trends at different hours.

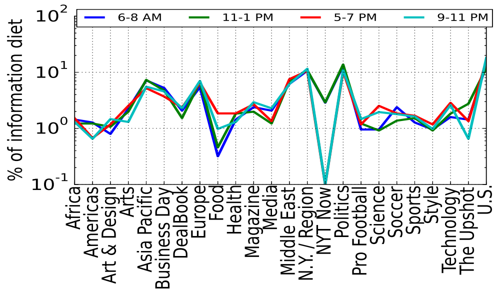


Figure 6: Difference in information diet for some topics in NYTimes top stories at different hours.

since these topics get recommended uniformly throughout the day. However, the variation is substantially higher for the niche topics (e.g. Health-fitness or Environment in Twitter, and Travel or Food in NYTimes) than the popular ones. Therefore, for users who are specifically interested in these niche topics, the temporal variation in the coverage of news-stories on these topics can lead to significant imbalances in their information diets.

4. CONCLUDING DISCUSSION

In this work, we showed that the topical composition of the information consumed by a user is effected by the user's browsing patterns on online news media sites. This problem - which, to our knowledge, has not been reported in any prior work – is a result of the high temporal churn in the popular content published by the media sites. Possible solutions to solve this imbalance can be twofold - (i) The users will need to devise a strategy of accessing the media sites such that their desired information diet is maintained; however, this is very difficult for individual users to attain. (ii) The media systems will need to tune their published content according to not only the popularity / recency of news-stories, but also the user's browsing habit and preferred information diet. In our future work, we plan to investigate the pros and cons of both of the above approaches.

Acknowledgement: The authors thank the anonymous reviewers whose suggestions helped to improve the paper. A. Chakraborty was supported by Google India PhD Fellowship and Prime Minister's Fellowship for Doctoral Research.

5. REFERENCES

[1] How Americans get their news. http://www.americanpressinstitute.org/publications/reports/survey-research/how-americans-get-news/.

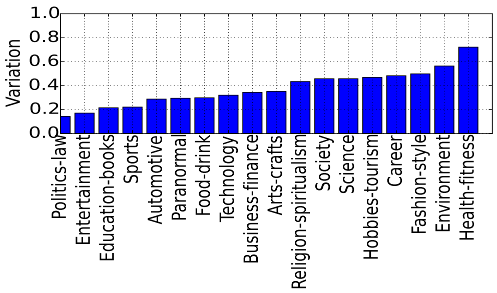


Figure 7: Variation in the contribution in information diet for different topics in Twitter trends.

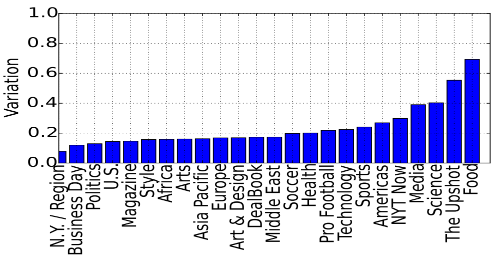


Figure 8: Variation in the contribution in information diet for different topics in NYTimes top stories.

[2] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida. Characterizing User Behavior in Online Social Networks. In *ACM IMC*, 2009.

[3] C. Budak, S. Goel, and J. Rao. Jasmine: A real-time local-event detection system based on geolocation information propagated to microblogs. In *ICWSM*, 2015.

[4] F. Duarte, B. Mattos, A. Bestavros, V. Almeida, and J. Almeida. Traffic characteristics and communication patterns in blogosphere. In *ICWSM*, 2007.

[5] Social Media, Political News and Ideology | Pew Research Center. http://www.journalism.org/2014/10/21/section-2-social-media-political-news-and-ideology/.

[6] S. Golder, D. Wilkinson, and B. Huberman. Rhythms of social interaction: messaging within a massive online network. In *ICCT*, 2007.

[7] T. Groseclose and J. Milyo. A measure of media bias. *The Quarterly Journal of Economics*, 2005.

[8] J. Kulshrestha, M. B. Zafar, L. E. Noboa, K. P. Gummadi, and S. Ghosh. Characterizing Information Diets of Social Media Users. In *ICWSM*, 2015.

[9] H. Liang, Y. Xu, D. Tjondronegoro, and P. Christen. Time-aware topic recommendation based on micro-blogs. In *CIKM*, 2012.

[10] M. Mathioudakis and N. Koudas. Twittermonitor: Trend detection over the twitter stream. In *SIGMOD*, 2010.

[11] Stop Overdosing on Celebrity Gossip, The News, and Low Quality Information. http://jamesclear.com/brain-food.

[12] To Trend or Not to Trend... | Twitter Blogs. https://blog.twitter.com/2010/trend-or-not-trend.

[13] E. Pariser. *The filter bubble : what the Internet is hiding from you*. Penguin Press, 2011.

[14] Newspapers: By the Numbers | State of the Media. http://tinyurl.com/news-by-numbers.

[15] Recency vs Relevancy | thoughts: archive. http://www.thoughtsarchive.com/recency-vs-relevancy/.

[16] K. Rudra, A. Chakraborty, M. Sethi, S. Das, N. Ganguly, and S. Ghosh. #FewThingsAboutIdioms: Understanding Idioms and Its Users in the Twitter Online Social Network. In *PAKDD*, 2015.

[17] Why is this news? https://twitter.com/search?q=%22why%20is%20this%20news%3F%22.