Received: 30 March 2019 | Revised: 11 September 2019 | Accepted: 13 September 2019

DOI: 10.1002/widm.1342

FOCUS ARTICLE

WIREs DATA MINING AND KNOWLEDGE DISCOVERY WILEY

On the dynamics of user engagement in news comment media

Lihong He[1] | **Chao Han**[1] | **Arjun Mukherjee**[2] | **Zoran Obradovic**[1] | **Eduard Dragut**[1]

[1]Department of Computer and Information Sciences, Temple University, Philadelphia, Pennsylvania

[2]Department of Computer Science, University of Houston, Houston, Texas

Correspondence
Eduard Dragut, Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122.
Email: edragut@temple.edu

Funding information
U.S. NSF, BigData, Grant/Award Numbers: 1838145, 1838147; EAGER, Grant/Award Number: 1842183; SES, Grant/Award Number: 1659998

Abstract

Many news outlets allow users to contribute comments on topics about daily world events. News articles are the seeds that spring users' interest to contribute content, that is, comments. A news outlet may allow users to contribute comments on all their articles or a selected number of them. The topic of an article may lead to an apathetic user commenting activity (several tens of comments) or to a spontaneous fervent one (several thousands of comments). This environment creates a social dynamic that is little studied. The social dynamics around articles have the potential to reveal interesting facets of the user population at a news outlet. In this paper, we report the salient findings about these social media from 15 months worth of data collected from 17 news outlets comprising of over 38,000 news articles and about 21 million user comments. Analysis of the data reveals interesting insights such as there is an uneven relationship between news outlets and their user populations across outlets. Such observations and others have not been revealed, to our knowledge. We believe our analysis in this paper can contribute to news predictive analytics (e.g., user reaction to a news article or predicting the volume of comments posted to an article).

This article is categorized under:

Internet > Society and Culture

Ensemble Methods > Web Mining

Fundamental Concepts of Data and Knowledge > Human Centricity and User Interaction

KEYWORDS

online social dynamic, social media, user interest

1 | INTRODUCTION

A decade after the advent of the Web 2.0 era, the Internet has become a game-changer in governance and social life. Since 2008, social media enriched websites have stormed into the sociopolitical scene as viable communication tools. The social media enriched websites offer citizens exceptional opportunities to actively engage in political and democratic processes (Fernandes, Giurcanu, Bowers, & Neely, 2010). During the 2008 elections, the majority of U.S. adults went online to keep themselves informed and involved with the election. Almost 40% of them talked about politics online (Somasundaran & Wiebe, 2010), and more than 1,000 Facebook groups were created offering an influential forum for political expression (Woolley, Limperos, & Oliver, 2010). Moreover, the social media enriched websites were instrumental in organizing revolts during the "Arab Spring" (Cottle, 2011; Ghannam, 2011), and mobilizing individuals to build social movements and political parties (Bennett & Segerberg, 2011; Segerberg & Bennett, 2011).

While millions of people take to the large social media enriched websites (L-SMWs), such as YouTube, Reddit, and Instagram, many people express their opinions on various daily topics on smaller-scale social media enriched websites (S-SMWs), oftentimes organized around news outlets and blog environments. Individually, S-SMWs may not amass massive populations, but they are still large, complex and together represent a significant portion of our society. For example, the user populations at Washington Post, The Guardian, and Daily Mail have more than 120,000 users each (Table 1). S-SMWs create the conditions for the emergence of a *civic long tail*: numerous small-scale masses, loosely connected, actively engaged in conversations on daily topics.

Compared to their larger counterparts, S-SMWs:

- contain focused conversations seeded by topics of news articles (most of the 889 comments on "The war against Columbus Day" at Washington Post are on this topic),
- tend to be more homogenous in their viewpoints about various daily life topics (e.g., New York Times endorses stricter gun control),
- are more polarized as users selectively view only materials aligned with their world view (e.g., Fox News opposes universal healthcare vs. New York Times endorses universal health care),
- are organized around a core group of users (the Pareto principle holds: less than 20% of the users in a news outlet produce more than 80% of the overall comment volume).

In this paper, we present an analysis of the relationship between a news outlet and its user population who comment there. One needs to recognize a distinguishing dynamic between the two:

> the user population can be assumed to be in a dormant state, which lasts until the outlet publishes a news article. This action wakes up the user population, which starts to comment on the topic of the article. Eventually, the interest of the users dies out. The cycle is repeated with the publication of new news articles.

In this study, we analyze a number of interesting features of the relationship between news outlets and their user populations:

TABLE 1 Summary statistics of our dataset

News outlet	Articles	Aw.C	Users	Comments	Avg.UA	Avg.CA	Avg.CU	%VolTop20%U (%)
The Guardian	8,032	3,656	**182,373**	2,976,831	49.9	814.2	16.3	86.40
Washington Post	**13,157**	**11,300**	141,565	4,532,532	12.5	401.1	32.0	91.70
Daily Mail	13,023	7,417	136,868	2,218,041	18.5	299.0	16.2	82.27
Fox News	5,831	3,271	126,299	**9,074,984**	38.6	**2,774.4**	**71.9**	95.88
New York Times	8,055	1,483	77,157	854,255	52.0	576.0	11.1	84.15
Market Watch	2,338	1,983	26,313	182,248	13.3	91.9	6.9	82.73
Wall Street Journal	7,796	5,499	19,380	913,080	3.5	166.0	47.1	89.97
CNN	6,095	436	15,221	326,087	34.9	747.9	21.4	91.66
BBC	5,774	105	14,958	99,844	**142.5**	950.9	6.7	79.20
Star Tribune	1,574	823	5,167	95,174	6.3	115.6	18.4	42.54
AlJazeera	708	684	4,700	44,931	6.9	65.7	9.6	83.18
Seattle Times	1,548	1,013	4,512	60,390	4.5	59.6	13.4	83.08
New York Daily News	5,209	367	1,893	10,208	5.2	27.8	5.4	73.17
VentureBeat	507	104	1,303	1,478	12.5	14.2	1.1	29.43
New York Post	3,503	102	873	1,577	8.6	15.5	1.8	50.22
TIME	4,457	568	665	16,840	1.2	29.6	25.3	**96.29**
Las Vegas Sun	213	148	616	4,516	4.2	30.5	7.3	63.94

Note: Outlets are ranked by user volume. The values given in bold are the largest one in each column.

Abbreviations: Aw.C, Articles with Comments; Avg.UA, Per Article avg. number of users; Avg.CA, Per Article avg. number of comments; Avg.CU, per user avg. number of comments; %VolTop20%U, % comment volume by the top 20% users.

- the relationship between the volume of articles on a story published by a news outlet and the volume of comments generated by its users.
- the distribution of time interval between the publication of an article and its first comment, as well as the distribution of time interval between the first and last comment.
- the user commenting activity pattern at different time granularities.
- the breadth and heterogeneity of user interest across different outlets.

We find that there is a strong mutual relationship between the article volume published by a news outlet and the comment volume generated by its user population. We also find that the user commenting activity at news outlets has a very different pattern than that of Twitter or Facebook by considering the user activity by times of day. Besides, we show that the similarity among news outlets along the users' breadths of interest (range of news stories) does not necessarily imply a similarity of heterogeneity of users' interest between the same outlets. In other words, the user interest may be distributed quite differently at outlets with similar breadths of user interest. At some outlets, users are interested in similar stories but they may show different focus levels.

We believe that the insight gained from our work may be used to develop new or to improve existing social media mining tools. Consider the work on predicting the volume of user comments a news article receives (Artzi, Pantel, & Gamon, 2012; Backstrom, Kleinberg, Lee, & Danescu-Niculescu-Mizil, 2013; Tsagkias, Weerkamp, & De Rijke, 2009; Yano & Smith, 2010). These works employ a variety of article-specific features, such as publication time and bag of words extracted from title and article body. But, to our knowledge, none of these works utilize user-specific features, which we believe may significantly hinder their performance. This statement is supported by the observation that between 20% and 50% of the comments do not respond to the journalistic news article, but rather to a previously posted user comment (Ruiz et al., 2011; Singer, 2009). Therefore, article-specific features cannot account for those comments. Our work in this paper is a step toward understanding the relationship between a news outlet and its user population. This in turn will suggest new features (e.g., user comment arrival time or rate), which may help improve the accuracy of forecasting models for social media at news outlets.

Our analysis is conducted on a dataset of 500 GB collected over 15 months, October 2015 to January 2017, and it includes both the articles and the comments gathered from 17 news outlets. To our knowledge there is no other work that matches the scale of our study.

2 | RELATED WORK

The human interaction in social networks has seen increased interest since the advancement of Web 2.0. One research topic in this area is the analysis of user activities (Barabasi, 2005; Oliveira & Barabási, 2005; Vázquez et al., 2006). A range of probability distributions is used to characterize user activities, including Poisson (Katti & Rao, 1968) and power law distributions (Clauset, Shalizi, & Newman, 2009). A key finding in these works is that the intervals between cause and action follow the power law distributions. Moreover, some studies (Iwata, Shah, & Ghahramani, 2013) show that there is a strong influence between users when the conversion focuses on a specific item/topic.

User activities play an important role in understanding topic trends, both in microblog and social media service portals. Most of the studies focus on the user behavior on Twitter, like users' repeated involvement (C. Wang & Huberman, 2012), the evolution of Twitter users and user behavior (Y. Liu, Kliman-Silver, & Mislove, 2014; Martins, Magalhães, & Callan, 2016), and the retweet activity (Kobayashi & Lambiotte, 2016). In the aspects of blog (J. Wang, Yu, Yu, Liu, & Meng, 2012) mentions different types of diversionary comments under political blog posts. The user activities in Reddit are also studied (Ferraz Costa, Yamaguchi, Juci Machado Traina, Traina Jr., & Faloutsos, 2015). Many other works study the user interaction at Digg and YouTube, such as the relaxation response (Crane & Sornette, 2008), the correlation between the trends at early and later duration (Szabo & Huberman, 2010), the relationship between sentiment of comment and comment rating (Siersdorfer, Chelaru, Pedro, Altingovde, & Nejdl, 2014).

As suggested in (Aker et al., 2016), social media receive thousands of comments every day. News outlets have become significant online platforms for users to share opinions. Mining the content produced by user in social media is a fruitful research, particularly for sentiment analysis (B. Liu, 2015) and social sciences (e.g., communications) (Artzi et al., 2012; Backstrom et al., 2013; Yano & Smith, 2010). N. Diakopoulos and Naaman (2011a) examine the relationships between news comment topicality, temporality, sentiment, and quality. Dos Rieis et al. (2015) analyze the sentiment of comments and headlines of news article. Tatar, Antoniadis, De Amorim, and Fdida (2014) study the duration between the publication time of an article and its last comment. Rizos, Papadopoulos, and Kompatsiaris (2016) predict the popularity of news article by mining the

online comments. Tan, Friggeri, and Adamic (2016) study the entire process of news propagation: from the information source to news article to user comment, the authors utilize sentiment to predict how far the article is from the information source. The study of online news comments attracts lots of attention from researchers, such as the conversational relevance (N. A. Diakopoulos, 2015), anonymity (Sachar & Diakopoulos, 2016), comment quality (Park, Sachar, Diakopoulos, & Elmqvist, 2016; N. Diakopoulos & Naaman, 2011b; Berry & Taylor, 2017), the structure of the discussions (Aragón, Gómez, & Kaltenbrunner, 2017), and the difference between male and female commenters (Pierson, 2015). A number of studies conclude that the content of comments combined with the content of news articles leads to superior results than when the content of news articles alone is used in mining tasks (Llewellyn, Grover, & Oberlander, 2016; Ma, Sun, Yuan, & Cong, 2012; Tan et al., 2016), such as topic clustering of news articles.

Our work distinguishes from the existing body of work in several ways. First, there is no study of the interplay between news outlets and their user populations, and how this may be quantified. Second, there is very limited study of social media formed by news outlets in general. We analyze the user commenting activity in all 17 news outlets and posit that the pattern of user activity at news outlets is quite different than that of Twitter and Facebook. Finally, we propose to analyze the dynamics of the outlet-story-article-user ecosystem through modeling user interest with breadth and heterogeneity. We are not aware of any work that characterizes the user interest in the social media coalesced at news outlets.

3 | DATA

We develop a user-configurable news article comment crawler, which collects the news articles along with their user comments. Each article is monitored for 6 months. Our analysis is based on the data crawled from 17 news outlets, which includes a mixture of major (e.g., BBC, Fox News, Wall Street Journal) and regional news outlets (e.g., Las Vegas Sun and Seattle Times).

3.1 | Ecosystem

We identify five main entities in social media at news outlets that define a unique (hierarchical) ecosystem: news outlets, news stories, news articles, comments, and users. A news outlet publishes news articles on various news stories. Users then read the news articles (of their interests) and provide comments. This leads to a natural hierarchical organization of these entities. Figure 1 illustrates such an ecosystem, where CNN releases a number of news articles about "iPhone" and "Iran" (stories), and the news articles are the catalysts for users to engage in commenting on news stories.

3.2 | Data summary

The data are summarized in Table 1. The dataset contains 21,413,016 comments from 38,959 news articles and 759,863 users from 17 news outlets. This data corresponds to 1,942 news stories that appear in the Top Stories in Google News (GNews) from October 2015 to January 2017. Washington Post (U.S.) with 11,300 provides the largest volume of articles with comments, Fox News (U.S.) provides the largest volume of comments, over 9 million, and The Guardian has the largest user population in our dataset. Fox News has the largest average user comment volume both per article, 2,774.4, and per user, 71.9. These are highlighted in the table.

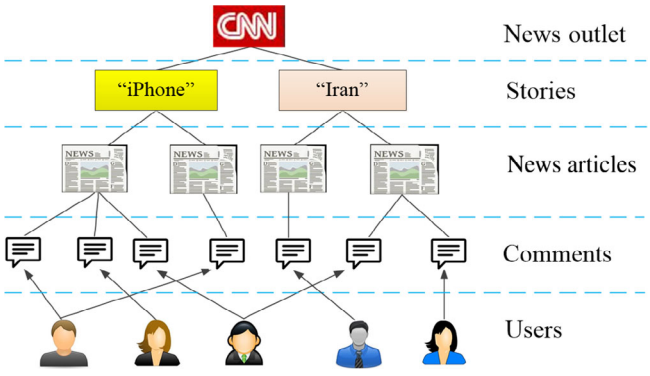


FIGURE 1 The hierarchical diagram of the social media ecosystem at a news outlet

3.3 | A bird's-eye view of users activity

In addition to these global views, we have some detailed observations about user activities.

First, we observe that some news outlets harbor very active users, for example, Fox News and Washington Post. The user population at Fox News produces twice or even four times as many comments as other user populations of comparable sizes (e.g., Washington Post and Daily Mail).

Second, our initial expectation was that major news outlets host user populations of comparable sizes. But, this was soon contradicted: for example, the number of users at BBC and CNN are quite small (at least 10 times smaller) compared to that at Guardian and Fox News. The reasons may be that (a) BBC and CNN only expose a small fraction of their news articles to user comments; some news outlets require to log in before one is able to post comments (which may discourage user activity); and (b) we may not reach a good fraction of the news articles with user comments (although we manually checked hundreds of articles from these outlets and empirically confirmed that few were made available to user discussion, for reasons unknown to us).

Third, outlets have a broad range of user loyalty (see column Avg.UA in Table 1): they range from user population constituted of occasional wanderers (e.g., TIME has on average one user per article) to frequent followers (e.g., BBC has 142 distinct users per article).

Fourth, despite the uneven user sizes, major news outlets have on average similar comment number per article (see column Avg.CA in Table 1): CNN (748) and BBC (951) compared to The Guardian (814). Fox News is the outlier with 2,774 comments per article.

Fifth, most of the content generated by a user community is from a small fraction of its constituency (as shown in the column %VolTop20%U in Table 1), for example, 95% of over 9 million comments at Fox News is generated by 20% of its users. This observation is farther illustrated for six major news outlets in Figure 2. This finding echoes the observations noticed in L-SMWs, such as Twitter where a great portion of all information consumed is generated by a small number of elite users (Hu et al., 2012; Wu, Hofman, Mason, & Watts, 2011).

4 | DYNAMICS OF THE ECOSYSTEM

In this section, we analyze the dynamics among the five entities in the ecosystem as shown in Figure 1. We realize that a news outlet is not a self-sustained social environment like Twitter, where messages are exchanged almost constantly with varied rates. Users require an impulse from their outlet to commence commenting. The impulse is a news article published by the outlet. Unless is further stimulated by the outlet, the user activity will go dormant. We study this interaction between news outlet and user here, which is represented by article and comment, respectively. We start with a case study, where we highlight certain patterns, and then propose a method to capture the relationship between them.

4.1 | A case study

We get the top-10 most popular distinct stories based on the news article volume in each story: $S1$ ("Donald Trump"), $S2$ ("Hillary Clinton"), $S3$ ("Bernie Sanders"), $S4$ ("Syria"), $S5$ ("iPhone"), $S6$ ("Prince"), $S7$ ("NFL"), $S8$ ("Orlando"), $S9$ ("Brussels"), and $S10$ ("Iran"). Recall that a news story follows a (dis)appear news cycle. For example, "Donald Trump" appears 138 times during our crawl. $S1$ is the story "Donald Trump" from January 6 to May 10 in 2016.

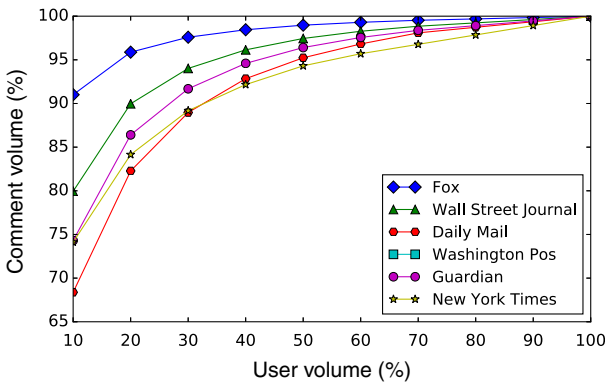


FIGURE 2 Comment volume from user population. Users are ranked based on the number of comments they post

We drill down on these stories in Figure 3, where we only consider the articles and comments from the outlets in our study. For each news story, we show the distribution of article volume across different outlets in the first graph, and the distribution of comment volume by outlets in the second one. Notice that Washington Post dominates the percentage of articles published for all but three stories, *S*5, *S*6, and *S*9. However, Fox News dominates the user-generated comments for all except two stories, *S*6 and *S*7.

An interesting dynamic is observed if we carefully analyze these two graphs. Intuition seems to suggest that the more news articles an outlet publishes on a story the more user comments will receive on it compared to the other outlets. Our data analysis shows that this is not true across the board. There is no universal and direct correlation between the volume of articles and that of comments working for all news outlets. On one hand, one notices in the two figures that at some news outlets a relatively small number of news articles on a topic can trigger an avalanche of comments. The clear example in our study is Fox News. For each of the news stories (except for *S*6 and *S*7), it has less than 10% of the total volume of articles, but its users produces between 20% and 70% of the entire volume of comments. On the other hand, a prodigious number of news articles at an outlet on a story may deter the user activity on that story in the long run. For instance, Washington Post despite publishing the larger number of articles per story in most cases, does not draw as large volume of comments as the other news outlets with comparable or even smaller user populations (see Table 1, column Users), for example, Fox News on stories *S*1, *S*2, and *S*3, or BBC on *S*4.

4.2 | Interplay between outlets and their users

An overall picture emerges between the outlets and their users from the above case study. Ideally, the relation between an outlet and its users should be of mutual enforcement: on one hand, an outlet should continue to publish on a news story as long as its users maintain a steady interest; on the other hand, the user should show interest in a story, thus demanding more news articles, by actively commenting (response) on the news articles (impulse). As it is apparent from our case study this relationship is not always even: some outlets publish more news articles than their users show interest in (e.g., Washington Post), while others publish less than its user demands (e.g., Fox News).

The question is whether we can characterize this relationship in a concise and intuitive way. If we pair the comment volume percentage and article volume percentage variables over all news stories, the deviation of their regression line b from the *ideal regression line* (at 45°)l provides a good indicator of the relation between outlet and its users. Using the cartoon example in Figure 4, one observes that as long as b is close to l, the news outlet and its users are in an ideal relation. If their regression line is farther to the left of l, then the user dominates their relation by generating large volume of comments per small number of articles. We say that user is *hyper-engaged* in their relation. If b is farther to the right of l, then the outlet dominates their relation by publishing more comments then its users are willing to comment on. In this case, user is *hypo-engaged*.

We can use a couple of measures to quantify this relationship. The most intuitive is to use the trigonometric function tan (α), where α is the angle of the regression line. If tan(α) is close to 1, user is "ideally engaged" with the news outlet. If it is close to 0, it indicates a dominance by the outlet. And, if it is much larger than 1, then user dominates their relation.

We provide the plot of news article and comment percentages in six news outlets in Figure 5. The Pearson correlation coefficient of each plot is shown in Table 2. According to this parameter, it is obvious that there is a correlation between the article

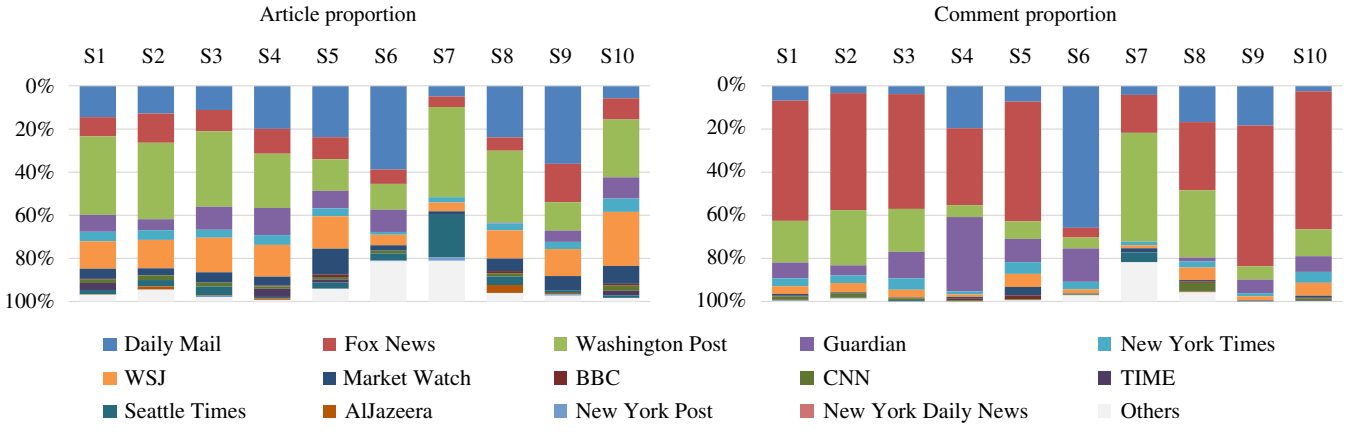
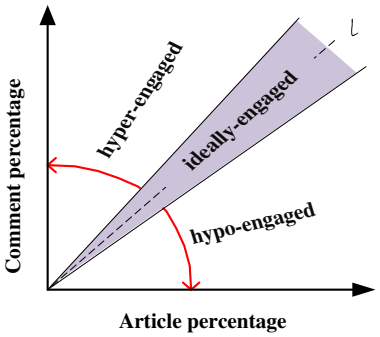


FIGURE 3 Proportion of each outlet for top 10 distinct stories based on total article volume

FIGURE 4 The modeling of interplay between outlets and users



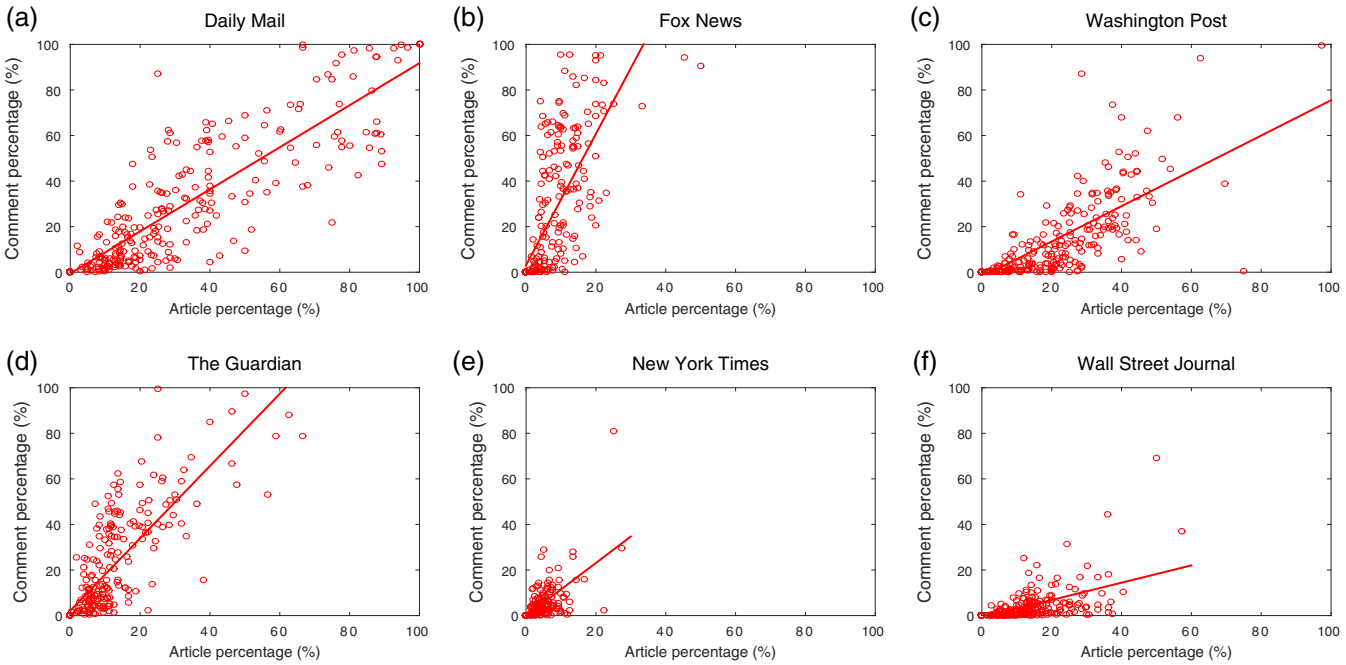


FIGURE 5 Interplay between news outlets and users. (a) Daily Mail, (b) Fox News, (c) Washington Post, (d) The Guardian, (e) New York Times, and (f) Wall Street Journal

TABLE 2 Parameter of correlation, linear regression line between article, and comment

News outlet	PCC	Angle	Tan	Relation
Daily Mail	0.89	42.7∘	0.92	Ideal
Fox News	0.77	70.9∘	2.88	Hyper
Washington Post	0.78	37.8∘	0.77	Hypo
The Guardian	0.83	57.8∘	1.59	Hyper
New York Times	0.69	49.3∘	1.16	Ideal
Wall Street Journal	0.6	20.8∘	0.38	Hypo

and comment percentages in each news outlet. After fitting the plot by a linear regression line, we know that the user commenting activity in Fox News is much more active than others, while the users activity in Wall Street Journal is the least active in the six outlets. These plots explain better why the average comments number per article of Fox News in Table 1 is much larger than other outlets, while the number of Wall Street Journal is the smallest. Because an active commenting population will lead to a high comment volume in a news article in general.

5 | DYNAMICS OF ECOSYSTEM VIA USER COMMENTING ACTIVITY

In this section, we focus on the analysis of users commenting activity at news outlets, which mainly involves three entities of the ecosystem in Figure 1: articles, users, and comments. We analyze:

- User reaction: the time difference between the first comment posted for an article and the article's publication.
- Duration of user comments: the time difference between the last and first comment for a news article.
- The intensity of user commenting activity at different time granularities.
- User stickiness: user retention rate of commenting activity over time.
- The dual popularity of a story in GNews with the user engagement across news outlets.

5.1 | User reaction

Let A be a news article in some news outlets and T_{react} the user reaction to A. For this study we use all the news articles in the column Aw.C in Table 1. Figure 6 gives an example of the distribution of T_{react} for Fox News. Table 3 shows the statistics of user response for six news outlets: Daily Mail, Fox News, Washington Post, The Guardian, New York Times, and Wall Street Journal. We omit the rest to have a more legible presentation.

We can draw several somehow unexpected conclusions. First, the user reaction to a news article happens within the first 2 hours in general at some outlets (the column F2Hs in Table 3). This extends the observation of time interval in (Ferraz Costa et al., 2015) from user postings to publication and first comment. The Guardian hosts the user community with the fastest reaction time, less than an hour on average. It is also the most alert news outlet among the 17, given that it has the lowest standard deviation (*SD*). Washington Post hosts the least alert user population, taking on average 4 hours to respond with a *SD* of over 5 hours. Second, and maybe the most surprising, a user may react to a news article even a month later (the long tail in Figure 6). We will provide an explanation in the next section. Third, it appears that the most suitable distribution to model the user response at a news outlet is the power law distribution. This is true in many cases, like the user activity in Digg and Twitter (Lerman & Ghosh, 2010). However, we posit that there is no one-size-fits-all solution to fitting the user response across all outlets. They may require a fit with a mixture of two or more distributions, like the mixture of Poisson and power law distributions. Our explanation is that Twitter, Facebook, and the rest of L-SMWs are the heavy tail of civic engagement, representing the average convergence of engagement behavior. In contrast, the user populations at news outlets represent the civic long tail, they are more diverse and nuanced, and there may not be one model to fit them all.

TABLE 3 User reaction stats. Mean and Standard Deviation (*SD*) are in hours

News outlet	Mean	SD	F2Hs (%)
Daily Mail	1.6	2.88	84.7
Fox News	3.1	4.67	68.3
Washington Post	4.2	5.36	58.7
The Guardian	0.9	1.51	96.16
New York Times	2.5	3.67	67.4
Wall Street Journal	2.7	4.19	67.5

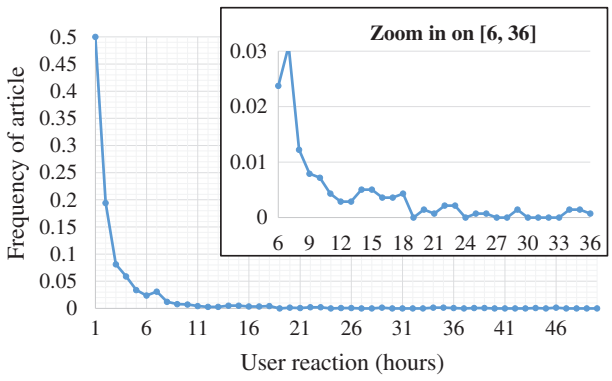


FIGURE 6 Distribution of user reaction at Fox News. User reaction is defined as the time difference between the first comment and the article publication time

The analysis of user reaction is an important tool when modeling user interest in social media, for example, conversation curation (Backstrom et al., 2013). In general, a longer response delay foretells a less-enthusiastic user activity. To our knowledge, this has not been studied in prediction tasks related to news outlets. It is thus worth pursuing the creation of new prediction models that include this dimension along with those extracted from news articles (Tsagkias et al., 2009). We plan to pursue this line of work in the future.

5.2 | Duration of user comments

Figure 7 shows the distribution of the duration of user commenting activity over the news articles with comments in five news outlets (we only present five news outlets here to make the plot clearer).

It is apparent from Figure 7 that the users interest for commenting lasts about 1 day after the first comment is posted in most articles (suggested by the area of peaks in each outlet). The (rather very) long tail is a surprising discovery (the commenting duration lasts into thousands of hours in some articles. Figure 7 shows only the beginning part of the long tail because of the limited space). The phenomenon can be explained in connection with news aggregation service, for example, Google News: a news story may show up in Google News days or even months after it drops out from the front page of a news outlet and many of the old news articles about the story are carried over. At which time, some people discover them and decide to post comments. We know that these old news articles are carried over in because our crawler rediscovers them whenever a story reappears. For example, consider the article "China's artificial islands in before-and-after photos" for the story "South China Sea," most of its comments are posted during October 27–30, 2015. There are, however, two comments posted on February 17, one comment on February 24, one comment on April 29, and one comment on July 20 in 2016. The story "South China Sea" appeared in Google News on October 28, 2015, reappeared on February 17 and 24, on July 12, 2016. If we compare the times when a story reappears and comments are posted for an article in that story, we find that they are consistent with each other. This may also explain the long tail of user reaction in Figure 6.

5.3 | User activity by time

Our goal is to analyze the pattern of user activity by days of week and times of day and understand whether this is different from the user activity reported in L-SMWs.

Figure 8 depicts the distribution of the volume of user comments over different time granularities for 10 news outlets. Figure 8a shows the distribution over the days of the week and Figure 8b shows them over the hours of a day. In Figure 8b, we partition a day into four time ranges: Morning (5:00–11:59 a.m.), Afternoon (12:00–4:59 p.m.), Evening (5:00–8:59 p.m.), and Night (9:00 p.m. to 5:00 a.m.). We can draw a number of interesting observations from these plots. First, we note that most comments are posted during weekdays. The user activity slows down substantially during weekends across all news outlets. It clearly picks up on Mondays, and at some outlets (e.g., Time) it reaches the peak on Fridays. Interestingly, the distribution is quite uniform over the weekdays for most of the news outlets, for example, Washington Post, Daily Mail, and Wall Street Journal. Second, from Figure 8b, we note that the commenting activity is not consistent across the news outlets, the peaks differ between the news outlets. For instance, most of the comments are posted at night at Washington Post and Daily Mail, in the morning at New York Times and BBC, and in the afternoon at Market Watch, CNN, and Time.

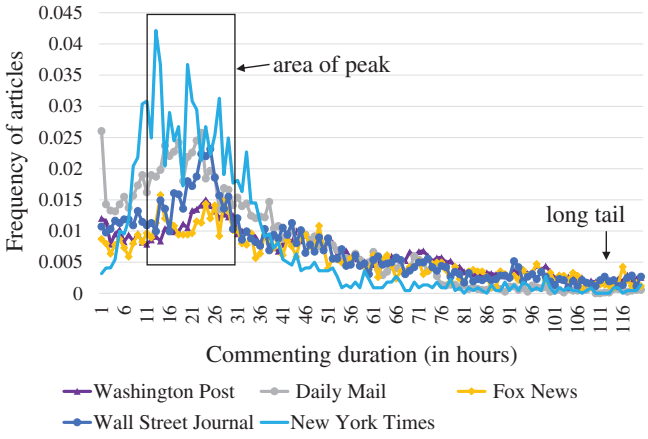


FIGURE 7 Distribution of the duration of user commenting activity at each outlet. The duration is defined as the time difference between the last and first comment in an article

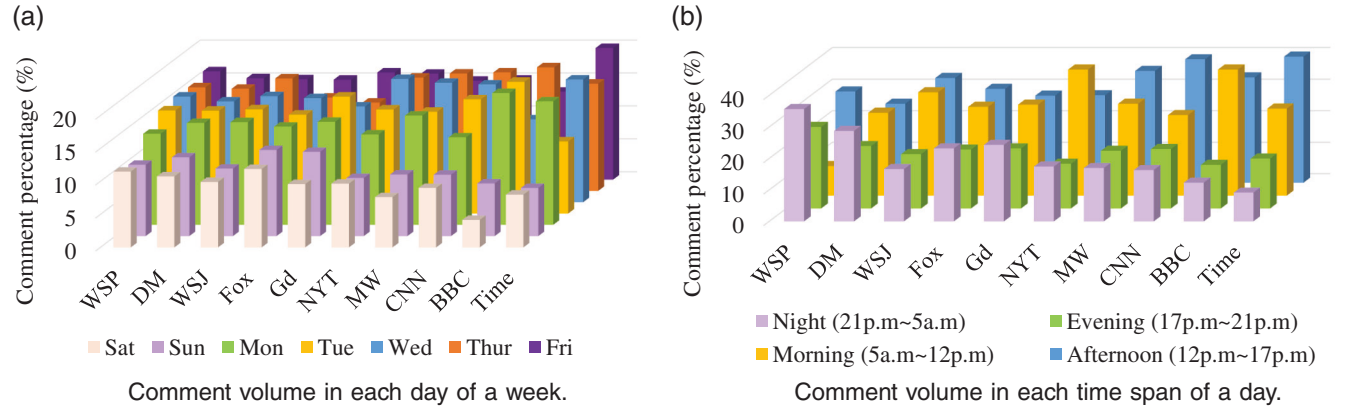


FIGURE 8 The volume of user comments by time at each outlet. WSP, Washington Post; DM, Daily Mail; WSJ, Wall Street Journal; Gd, The Guardian; NYT, New York Times; MW, Market Watch. (a) Comment volume in each day of a week. (b) Comment volume in each time span of a day

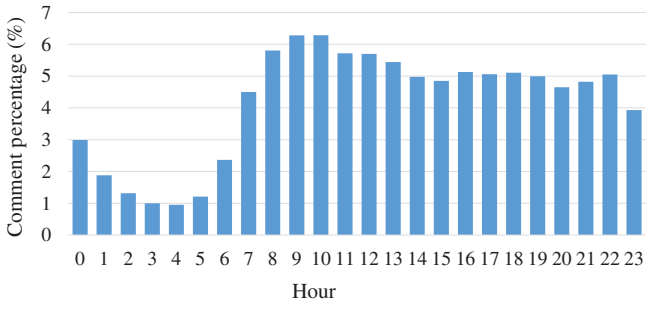


FIGURE 9 Comment volume per hour at Daily Mail, focusing on the users from United Kingdom. The time of comment is displayed by the users time zone

We draw attention to a subtle issue when analyzing this data. Most of these outlets cater to audiences across the globe, hence time zone may slightly shift these distributions in practice. Each user post has two times in general: the time t_u when the user issues the comment and the time t_s when the web server receives the comment. Crawlers have direct access to t_s at any news outlet, but not t_u. Considering the effect of time zones, the comments may shift across the buckets. For example, the Washington Post server receives a comment at t_s = 10:00 p.m. (assume the server is in East United States, timezone UTC−5). If the user is in Los Angeles when posting the comment, then t_u = 7:00 p.m. (timezone UTC-8). Hence, this comment should be counted into the evening bucket instead of the night bucket.

Daily Mail is the only news outlet that gives user location at country and city granularities. This allows us to recover the timezone of most of the comments and compare the distribution according to t_s and that according to t_u. According to our data, the user population at Daily Mail has the following distribution: United Kingdom 40.9%, United States 33.3%, Australia 2.3%, Canada 1.61%, Ireland 0.85%, France 0.68%, and Spain 0.51%. We omit the countries with under 0.5%. The information about city or state is ambiguous or missing in some cases, especially for U.S. users. We discard them from our analysis. We can recover the user timezone given a user's country in most cases because most countries span only a timezone, like United Kingdom. Figure 9 shows the distribution of comment volume by hour after adjusting the server time to the user time.

The Kullback–Leibler divergence between this and the one according to t_s (Figure 8b) is only 0.034. Thus, the analysis according to t_s approximates very well the user-commenting behavior according to t_u for Daily Mail. We believe this to be true for the rest of the news outlets.

If we look over the comment volume by hours of day, we find that the peak of commenting activity is between 10:00 a.m. and 2:00 p.m. at most news outlets, which is earlier than the peaks of user activity at L-SMWs—the peak at Facebook is between 1:00 and 4:00 p.m. and at Twitter 11 a.m. to 3 p.m. (Gillett, 2014). This finding is opposite to the assumed *social-media-to-online-news* news consumption patterns of college students with average age to be 19.8 years old (Tandoc Jr & Johnson, 2016). One possible explanation is that there is little user profile overlap between the user populations at news websites and social media, as the median age of those who visit news websites is 41.4 (Conaghan, 2017). We plan to study this aspect in more detail in the future.

5.4 | User stickiness

In addition to the study on the activity of user commenting behavior over time, we also ***explore the retention of user commenting behavior over time since it is crucial for us to understand the stickiness of user on different outlets. In order to conduct an unbiased analysis, we only investigate the user stickiness on articles related to "Donald Trump" across five major news outlets in 2016, which was the election year. We firstly define the monthly retention rate of user commenting behavior, which is the ratio of the number of users who continue commenting in two consecutive months to the number of users who only comment in the first month. For example, the monthly retention rate between June and July is calculated by dividing the number of users who comment on both June and July by the number of users who only comment in June. Given its definition, we can have 11 monthly retention rates among 12 months of 2016. An example on Wall Street Journal is presented in Figure 10a, where the number of user with commenting activities is represented by the blue bars, and the retention rate between 2 months is represented by the red star between two corresponding blue bars. To have a better sense about user stickiness from the aggregated user activities, we define the user stickiness as the median of monthly user retention rates within a year. Comparison of user stickiness for Wall Street Journal, Fox News, Washington Post, New York Times, and The Guardian is presented in the Figure 10b. We do not compare with Daily Mail here since its data in 2016 is incomplete. As we can see, the median retention rate is at least 0.4 across all five outlets, which indicates that at least 40% of users are likely to keep commenting about election-related articles in more than half of the year in 2016. This significant user stickiness presumably reflects more persistent user behavior on veteran news outlets due to their high-quality posts about import events, which is not observed on social media (https://www.statista.com/chart/2157/twitter-user-retention/). Besides, we note that users are more sticky on more experienced news outlets. For example, users are more sticky on Fox News and Washington Post compared with New York Times and The Guardian. We also surprisingly find that Wall Street Journal has the highest user stickiness compared to others. It is very likely that their election-related articles are more practical to users since they may discuss more about the influence of elections to finance.

5.5 | News stories dual popularity

There has been numerous speculations about the factors employed by Google to promote a story in its GNews platform. In this study we show that the popularity of a story (i.e., time spent) in GNews overlaps with its popularity (user engagement) across news outlets.

For a given news story S, let T_{AS} and T_{DS} be the times when S appears in and disappears from GNews, respectively. Let V_{S} be the volume of cumulative comments across all news outlets for S, as shown in Figure 11. We analyze the relationship between V_S and T_{AS} and T_{DS}. We discretize V_S in increments of 5% and analyze its cumulative behavior relative to both T_{AS} and T_{DS} for all S (874 stories after discarding some stories which have less than 100 comments from the original 1,115 ones) in our dataset. Figure 12a gives the relationship between $V_S^{T_{\text{AS}}}$ and T_{AS}. It shows that 51% of the news stories have up to 5% of their user comments in by the time the story breaks into GNews. Google is late to the party (more than 50% of V_S is in) for 23.5% of the stories. It completely misses on the user popularity (at least 95% of V_S is in) for less than 5% of all stories. Figure 12b shows the relation between $V_S^{T_{\text{DS}}}$ and T_{DS}. 77.3% of the stories disappear from GNews when at least 95% of their

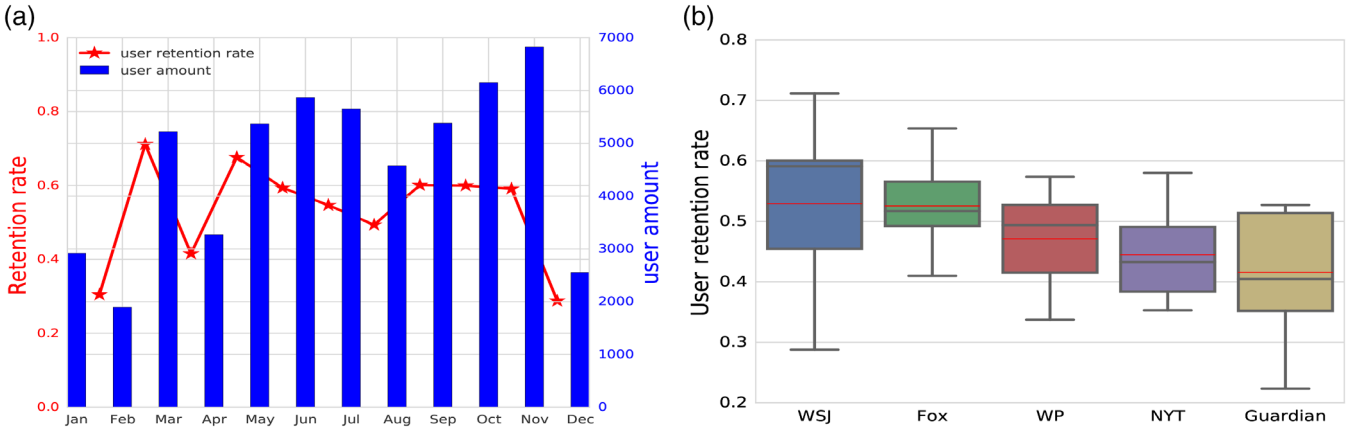


FIGURE 10 User stickiness. (a) Monthly user retention rates about "Donald Trump" related articles in 2016 on Wall Street Journal. (b) User Stickiness: distribution of monthly user retention rate about "Donald Trump" related articles in 2016 across five major news outlets

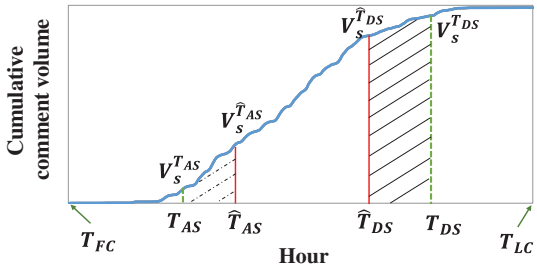


FIGURE 11 Cumulative comment volume over time

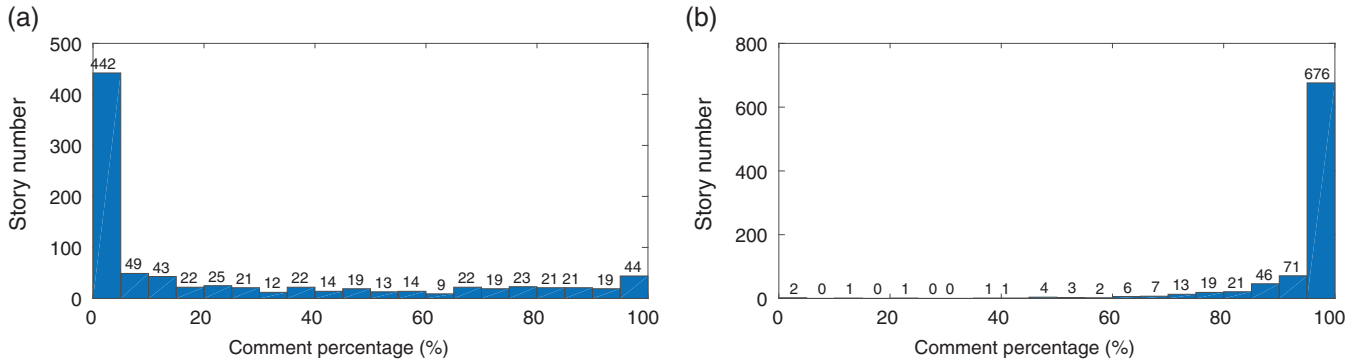


FIGURE 12 User engagement versus story popularity. (a) Appearance in GNews and (b) Disappearance in GNews

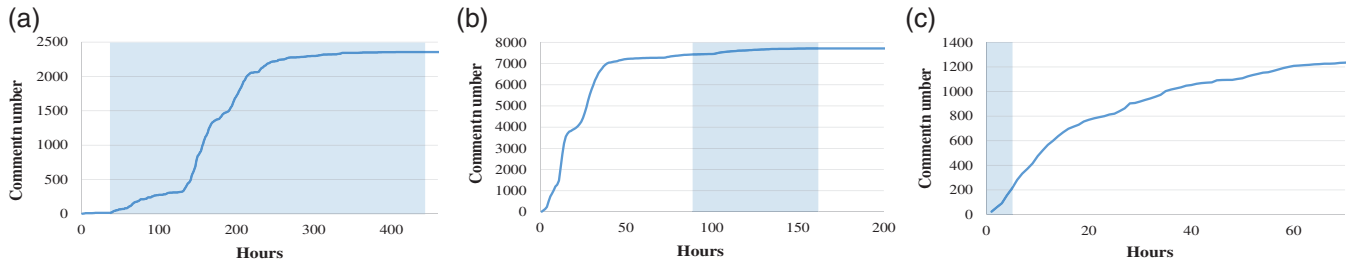


FIGURE 13 Cumulative comment growth over time for story. (a) Apple, (b) Angela Merkel, and (c) Fort McMurray

volumes of user comments is in. Comparing the two graphs, Google is significantly better at sensing when the user popularity drops than when the popularity picks up for a news story. Observe that the two graphs almost perfectly mirror each other.

5.5.1 | Examples

We give here three examples of news stories that illustrate the patterns described above. The first example is "Apple" in Figure 13a. The graph shows the cumulative volume of user comments for "Apple." The shaded area is the duration of the story in Top Stories in GNews. It breaks into GNews when less than 3% of the comments are posted, and it disappears very close to when the user engagement ceased. This is an example of a news story for which GNews recognizes early its popularity, but does not recognize the early signs of dwindling user interest. The second example is the news story "Angela Merkel" in Figure 13b. It is included in Top Stories when the users' interest is almost over. The explanation is that a story with the same name was present in GNews 5 days earlier, which stayed up for 2 days, was then dropped, and reappeared 3 days later. The user engagement was the most fervent during those 3 days when it was not present in GNews. The third example is "Fort McMurray," which briefly appeared in GNews between 12:28 a.m. and 5:42 p.m., May 12. The story reappears 28 hours later in GNews, missing a full day of user discussion, which in general may account for almost a third of the entire volume of user discussion.

6 | DYNAMICS OF ECOSYSTEM VIA USER INTEREST

Here, we aim to elevate the individual user-commenting behavior (along the hierarchy in Figure 1) at the news outlet level and give an aggregated characterization about outlets. We can view each outlet as a meta user and the goal is to create an

aggregated profile for each meta user, on which we compare the news outlets. We build our study upon users' breadth and heterogeneity of interests in news stories.

6.1 | Breadth of user interest

We firstly analyze the breadth of interests of users across our population of news outlets. We propose to quantify a user's breadth of interest (UBI) as the number of news stories on which the user comments (via the news articles in those stories). In Figure 14, we present the averaged UBI in different portions (top-k percentage) of user population across six frequently visited outlets. We note that the growth of UBI is very slow in all outlets. For example, the increment is less than 5 for every 10% of a population. This indicates that only a small portion (e.g., 10%) of users have broad interests, while the vast majority of users (i.e., 90%) are interested in a limited number of news stories. We also observe that though different news outlets have different sizes of user population, their aggregated UBIs of their top 10% users are quite similar, for example, Wall Street Journal and Daily Mail.

6.2 | Diversity of user interest at news outlets

We aim to quantify the overall UBI at each outlet and use it to make statements about the diversity of interests among the users in each of these outlets' populations. The challenge is that of finding a concise representation of the distribution of interest among a user population. We propose to use *heterogeneity* of user interest to measure the diversity of interests among multiple users. If we regard the users of a news outlet as nodes in a graph, then the heterogeneity of the interests of an entire user population is naturally captured using the mean and *SD* of *farness* of user nodes in the graph. The farness of a node is the averaged distance of the node to its neighbors, which is also known as the inverse of closeness centrality (Sabidussi, 1966). The heterogeneity of interest of the user population of outlet k is hence given by:

$$\text{mean} = \frac{1}{N_k}\sum_{i=1}^{N_k}\frac{\text{farness}(i)}{N_k-1} = \frac{\sum_{j\neq i} d\left(U_k^i, U_k^j\right)}{N_k(N_k-1)}$$

$$SD = \left\{\frac{1}{N_k(N_k-1)-1}\sum_{j\neq i}\left\{d\left(U_k^i, U_k^j\right)-\bar{d}\right\}^2\right\}^{\frac{1}{2}}$$

where N_k is the number of users, U_k^i is the vector representation of the interests of user i, and $d(\cdot,\cdot)$ is a distance function. U_k^i is indexed by the news stories. So, U_k^i is a vector of dimension 1,942 in our study. We propose two kinds of vector representations for U_k^i: (a) Binary representation, where the mth entry of U_k^i is 1 if the user is interested in the mth story, and 0 otherwise. The distance between two binary vectors is measured with the normalized Hamming distance. (b) Discrete representation, where the mth entry of U_k^i is the number of articles in mth story on which user i comments. We use the Cosine distance to measure the distance between two U_k^i vectors in this representation.

In this study, the reference user population at each outlet is the top-10% (recall Figure 1). Figure 15a shows the heterogeneity according to the discrete representation. This is a bar plot. The height of every single solid bar stands for the mean of the pairwise distance between user interest. The half length of error bar is the *SD* of the pairwise distance between user interest. We observe that the user population of Daily Mail exhibits a more heterogeneous interest in news stories, while the user

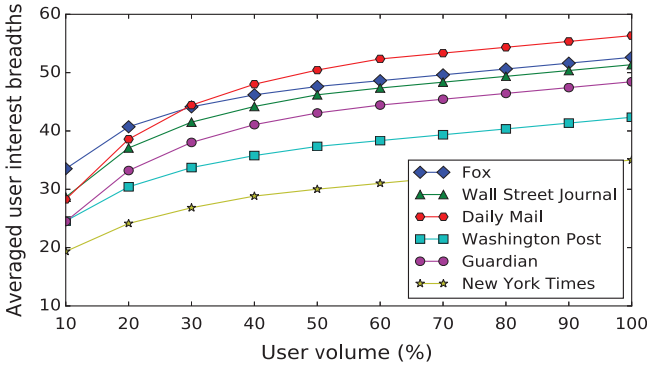


FIGURE 14 Cumulative plot of averaged breadths of user interest (number of distinct stories). Users are ranked by their breadths of interest from large to small

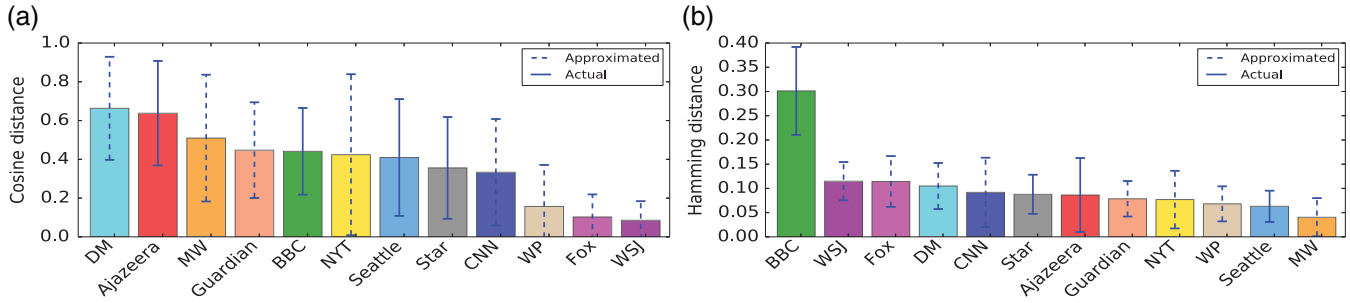


FIGURE 15 Quantifying breadth and heterogeneity of user populations at news outlets in news stories. (a) User interest heterogeneity evaluated on discrete representation and (b) user interest breadth evaluated on binary representation

populations of Wall Street Journal are more homogeneous, although their averaged UBIs are almost identical. The Cosine distance between two discrete representations tells us more about whether two users have similar focus in their interests. For example, user A comments on 90 sport news articles and 10 politics news articles, and user B comments on 10 sport articles and 90 politics articles. Then they are far apart because the Cosine distance between their interests is large even though they both show interest in sports and politics. The binary representation gives a different picture about the heterogeneity of a user population (Figure 15b). It compares the range and overlap of stories between two users. In this representation user A and user B are indistinguishable: they have identical interests. Comparing the two plots, one notices that heterogeneity of Wall Street Journal population is ranked much higher in Figure 15b than in Figure 15a. It indicates that Wall Street Journal users tend to have diverse (heterogeneous) interests but with similar (homogeneous) focus. Likewise, the heterogeneity of Market Watch population is ranked much lower in Figure 15b than in Figure 15a, which means Market Watch users are more likely to have similar interests but with diverse focus.

6.2.1 | Notes on computation

The plots in Figure 15 are computationally expensive to produce. The main reason is that they require the computation of the distance between each pair of users. For instance, 10% of Fox News population is 12,629, which gives roughly 10^9 pairs. It is impractical to exhaustively compute all 10^9 distances. Hence, we provide the approximate statistics for the outlets with large user populations; the error bars are dashed lines for these outlets in the two plots. The approximation procedure is via uniformly sampling multiple contiguous blocks with 1,000 users together with incremental updating the mean and *SD* (Chan, Golub, & LeVeque, 1983).

7 | CONCLUSIONS

In this paper, we study the commenting activity on news articles at 17 news outlets. We find that there is a strong mutual relationship between the article volume and the comment volume. According to the analysis of the user-commenting activity by times of day, we find that the user activity at news outlets has a very different pattern than that of Twitter or Facebook. Besides, we analyze the breadth and heterogeneity of interest in news stories of the user populations across all news outlets.

As a part of future work, we plan to draw from the lessons learned in this study to create novel prediction models for user comment volume of a news article in news outlets. Our study in this paper suggests that properties of user commenting activity at outlets may help improve the existing prediction models.

ACKNOWLEDGMENTS

This work was supported in part by the following U.S. NSF grants: BigData #1838145 and #1838147, EAGER #1842183, and SES #1659998.

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Lihong He: Conceptualization, data curation, formal analysis, resources, software, validation, writing-original draft. **Chao Han**: formal analysis, software, validation, writing-original draft. **Arjun Mukherjee:** Conceptualization, formal analysis, funding acquisition, investigation, methodology, validation, writing-original draft. **Zoran Obradovic:** Conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision. **Eduard Dragut:** Conceptualization, funding acquisition, investigation, methodology, project administration, resources, supervision, validation, writing-original draft.

ORCID

Eduard Dragut https://orcid.org/0000-0002-3103-054X

RELATED WIREs ARTICLES

Social network analysis: An overview

REFERENCES

Aker, A., Kurtic, E., Balamurali, A., Paramita, M., Barker, E., Hepple, M., & Gaizauskas, R. (2016). A graph-based approach to topic clustering for online comments to news. In: *European Conference on Information Retrieval* (pp. 15–29).

Aragón, P., Gómez, V., & Kaltenbrunner, A. (2017). To thread or not to thread: The impact of conversation threading on online discussion. In: *Eleventh International AAAI Conference on Web and Social Media* (pp. 12–21).

Artzi, Y., Pantel, P., & Gamon, M. (2012). Predicting responses to microblog posts. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 602–606).

Backstrom, L., Kleinberg, J., Lee, L., & Danescu-Niculescu-Mizil, C. (2013). Characterizing and curating conversation threads: Expansion, focus, volume, re-entry. In: *Proceedings of the sixth ACM international conference on Web search and data mining* (pp. 13–22).

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, *435*(7039), 207–211.

Bennett, W. L., & Segerberg, A. (2011). Digital media and the personalization of collective action: Social technology and the organization of protests against the global economic crisis. *Information, Communication & Society*, *14*(6), 770–799.

Berry, G., & Taylor, S. J. (2017). Discussion quality diffuses in the digital public square. In: *Proceedings of the 26th International Conference on World Wide Web* (pp. 1371–1380).

Chan, T. F., Golub, G. H., & LeVeque, R. J. (1983). Algorithms for computing the sample variance: Analysis and recommendations. *The American Statistician*, *37*(3), 242–247.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009, November). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703.

Conaghan, J. (2017). *Young, old and in-between: Newspaper platform readers ages are well-distributed*. Retrieved from https://www.newsmediaalliance.org/age-newspaper-readers-platforms/

Cottle, S. (2011). Media and the Arab uprisings of 2011: Research notes. *Journalism*, *12*(5), 647–659.

Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. In: *Proceedings of the National Academy of Sciences*, *105*(41), 15649–15653.

Diakopoulos, N., & Naaman, M. (2011a). Topicality, time, and sentiment in online news comments. Paper presented at the meeting of the 29th Annual CHI Conference on Human Factors in Computing Systems, CHI 2011 (pp. 1405–1410).

Diakopoulos, N., & Naaman, M. (2011b). Towards quality discourse in online news comments. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 133–142).

Diakopoulos, N. A. (2015). The editor's eye: Curation and comment relevance on the New York times. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1153–1157).

Dos Rieis, J. C. S., de Souza, F. B., de Melo, P. O. S. V., Prates, R. O., Kwak, H., & An, J. (2015). *Breaking the news: First impressions matter on online news*. Paper presented at the meeting of the Ninth International AAAI Conference on Web and Social Media.

Fernandes, J., Giurcanu, M., Bowers, K. W., & Neely, J. C. (2010). The writing on the wall: A content analysis of college students' Facebook groups for the 2008 presidential election. *Mathematics and Computer Science*, *13*(5), 653–675.

Ferraz Costa, A., Yamaguchi, Y., Juci Machado Traina, A., Traina, C., Jr., & Faloutsos, C. (2015). Rsc: Mining and modeling temporal activity in social media. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 269–278).

Ghannam, J. (2011). Social media in the Arab world: Leading up to the uprisings of 2011. *Center for International Media Assistance*, *3*, 1–44.

Gillett, R. (2014). *The best (and worst times) to post on social media (infographic)*. Retrieved from https://www.fastcompany.com/3036184/the-best-and-worst-times-to-post-on-social-media-infograph

Hu, M., Liu, S., Wei, F., Wu, Y., Stasko, J., & Ma, K.-L. (2012). Breaking news on twitter. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2751–2754).

Iwata, T., Shah, A., & Ghahramani, Z. (2013). Discovering latent influence in online social activities via shared cascade Poisson processes. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Katti, S., & Rao, A. V. (1968). Handbook of the poisson distribution. *Technometrics*, *10*(2), 412–412.

Kobayashi, R., & Lambiotte, R. (2016). *Tideh: Time-dependent hawkes process for predicting retweet dynamics*. arXiv Preprint arXiv:1603.09449.

Lerman, K., & Ghosh, R. (2010). Information contagion: An empirical study of the spread of news on digg and twitter social networks. In: *Fourth International AAAI Conference on Weblogs and Social Media* (pp. 90–97).

Liu, B. (2015). *Sentiment analysis - Mining opinions, sentiments, and emotions*. Cambridge, England: Cambridge University Press.

Liu, Y., Kliman-Silver, C., & Mislove, A. (2014). The tweets they are a-changin: Evolution of twitter users and behavior. In: *Eighth International AAAI Conference on Weblogs and Social Media* (Vol. 30, pp. 5–314).

Llewellyn, C., Grover, C., & Oberlander, J. (2016). Won't somebody please think of the children? Improving topic model clustering of newspaper comments for summarisation. In: *Proceedings of the ACL 2016 Student Research Workshop* (pp. 43–50).

Ma, Z., Sun, A., Yuan, Q., & Cong, G. (2012). Topic-driven reader comments summarization. In: *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 265–274).

Martins, F., Magalhães, J., & Callan, J. (2016). Barbara made the news: Mining the behavior of crowds for time-aware learning to rank. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*.

Oliveira, J. G., & Barabási, A.-L. (2005). Human dynamics: Darwin and einstein correspondence patterns. *Nature*, *437*(7063), 1251–1251.

Park, D., Sachar, S., Diakopoulos, N., & Elmqvist, N. (2016). Supporting comment moderators in identifying high quality online news comments. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1114–1125).

Pierson, E. (2015). Outnumbered but well-spoken: Female commenters in the New York times. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 1201–1213).

Rizos, G., Papadopoulos, S., & Kompatsiaris, Y. (2016). *Predicting news popularity by mining online discussions*. Paper presented at the meeting of the Proceedings of the 25th International Conference Companion on World Wide Web (pp. 737–742).

Ruiz, C., Domingo, D., Micó, J. L., Díaz-Noci, J., Meso, K., & Masip, P. (2011). Public sphere 2.0? The democratic qualities of citizen debates in online newspapers. *The International Journal of Press/Politics*, *16*(4), 463–487.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika*, *31*(4), 581–603.

Sachar, S. S., & Diakopoulos, N. (2016). Changing names in online news comments at the New York times. In: *Tenth International AAAI Conference on Web and Social Media* (pp. 339–347).

Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests. *The Communication Review*, *14*(3), 197–215.

Siersdorfer, S., Chelaru, S., Pedro, J. S., Altingovde, I. S., & Nejdl, W. (2014). Analyzing and mining comments and comment ratings on the social web. *ACM Transactions on the Web (TWEB)*, *8*(3), 17.

Singer, J. B. (2009). Separate spaces: Discourse about the 2007 scottish elections on a national newspaper web site. *The International Journal of Press/Politics*, *14*(4), 477–496.

Somasundaran, S., & Wiebe, J. (2010). *Recognizing stances in ideological on-line debates*. Paper presented at the meeting of the Workshop at NAACL HLT.

Szabo, G., & Huberman, B. A. (2010). Predicting the popularity of online content. *Communications of ACM*, *53*(8), 80–88.

Tan, C., Friggeri, A., & Adamic, L. (2016). Lost in propagation? Unfolding news cycles from the source. In: *Tenth International AAAI Conference on Web and Social Media* (pp. 378–387).

Tandoc, E. C., Jr., & Johnson, E. (2016). Most students get breaking news first from twitter. *Newspaper Research Journal*, *37*(2), 153–166.

Tatar, A., Antoniadis, P., De Amorim, M. D., & Fdida, S. (2014). From popularity prediction to ranking online news. *Social Network Analysis and Mining*, *4*(1), 174.

Tsagkias, M., Weerkamp, W., & De Rijke, M. (2009). Predicting the volume of comments on online news stories. In: *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 1765–1768).

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., & Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E*, *73*(3), 036127.

Wang, C., & Huberman, B. (2012). Long trend dynamics in social media. *EPJ Data Science*, *1*(1), 2.

Wang, J., Yu, C. T., Yu, P. S., Liu, B., & Meng, W. (2012). Diversionary comments under political blog posts. In: *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1789–1793).

Woolley, J. K., Limperos, A. M., & Oliver, M. B. (2010). The 2008 presidential election, 2.0: A content analysis of user-generated political facebook groups. *Mathematics and Computer Science*, *13*(5), 631–652.

Wu, S., Hofman, J. M., Mason, W. A., & Watts, D. J. (2011). Who says what to whom on twitter. In: *Proceedings of the 20th international conference on World wide web* (pp. 705–714).

Yano, T., & Smith, N. A. (2010). What's worthy of comment? Content and comment volume in political blogs. In: *Fourth International AAAI Conference on Weblogs and Social Media*.

How to cite this article: He L, Han C, Mukherjee A, Obradovic Z, Dragut E. On the dynamics of user engagement in news comment media. *WIREs Data Mining Knowl Discov*. 2020;10:e1342. https://doi.org/10.1002/widm.1342