A macroscopic analysis of news content in Twitter

Momin M. Malik and **Jürgen Pfeffer**

Momin M. Malik and Jürgen Pfeffer. (2016). A macroscopic analysis of news in Twitter. *Digital Journalism* 4 (8): 955-979. doi:10.1080/21670811.2015.1133249.

Previous literature has considered the relevance of Twitter to journalism, for example as a tool for reporters to collect information and for organizations to disseminate news to the public. We consider the reciprocal perspective, carrying out a survey of news media-related content within Twitter. Using a random sample of 1.8 billion tweets over four months in 2014, we look at the distribution of activity across news media and the relative dominance of certain news organizations in terms of relative share of content, the Twitter behavior of news media, the hashtags used in news content versus Twitter as a whole, and the proportion of Twitter activity that is news media-related. We find a small but consistent proportion of Twitter is news media-related (0.8 percent by volume); that news media-related tweets focus on a different set of hashtags than Twitter as a whole, with some hashtags such as those of countries of conflict (Arab Spring countries, Ukraine) reaching over 15 percent of tweets being news media-related; and we find that news organizations' accounts, across all major organizations, largely use Twitter as a professionalized, one-way communication medium to promote their own reporting. Using Latent Dirichlet Allocation topic modeling, we also examine how the proportion of news content varies across topics within 100,000 #Egypt tweets, finding that the relative proportion of news media-related tweets varies vastly across different subtopics. Over-time analysis reveals that news media were among the earliest adopters of certain #Egypt subtopics, providing a necessary (although not sufficient) condition for influence.

KEYWORDS: computational; news media; social media; topic modeling; Twitter

Introduction

The rise of the internet and social media has been involved with a crisis in journalism (Chouliaraki and Blaagaard 2013; Franklin 2012; Hirst 2010; McChesney 2012; Picard 2014). As part of attempts to adapt professional journalism to shifting patterns of consumption, there has been an increasing amount of work about social media, and in particular Twitter, looking at how social media users consume news (Bastos 2015; Hermida et al. 2012; Nielsen and Schrøder 2014), as well as newsroom studies about how news organizations use social media (Armstrong and Gao 2010; Artwick 2013; Bosch 2014; Broersma and Graham 2013; Cozma and Chen 2013; El Gody 2014; Engesser and Humprecht 2015; Hermida 2010, 2012, 2013; Ju, Jeong, and Chyi 2014; Mare 2014; Revers 2014; Thurman and Walters 2013; Verweij and van Noort 2014; Vis 2013).

At the same time, there has been a discussion of the possibilities of bringing computational analysis to journalism, both in the actual profession (Lewis 2015; Lewis and Westlund 2015; Parasie 2015; Young and Hermida 2015) including from mining Twitter data (Vis 2013), as well as in studying journalism and news media (Flaounas et al. 2013; Jang and Pasek 2015). Aggregate data is an attractive alternative to labor-intensive hand-coding, which has practical limitations on scale (Flaounas et al. 2013). Yet there has been less discussion about using computational methods to study news organizations' use of social media (and only a few examples, e.g., Lotan et al. 2011). This is surprising, given how the social media platforms are engineered specifically to allow for exactly such large-scale analysis (Gehl 2014)—specifically, for reducing users to computationally legible units that can be data mined to optimize advertising—and how news organizations themselves are seeking out insights that come from the same style of data mining and analytics as is used in advertising (Castillo et al. 2014).[1]

Using the Twitter "decahose," a random 10 percent sample of all Tweets, we investigate three theoretical directions. First, given the discussions in journalism literature about how news organizations may best use Twitter, how are different news organizations using Twitter? There have been newsroom case studies for some individual organizations, but no comprehensive view across all news media and journalists. Second, drawing on attention in literature to the importance of the politics, economics, and culture of social media platforms, we look at how journalism fits into Twitter as a whole. Lastly, motivated by literature on agenda-setting, we look at what topics are addressed by news media comparative to Twitter as a whole, and what topics have a greater presence of news tweets versus all other tweets.

Previous Work

Twitter

There is now a body of literature providing detailed overviews of Twitter's mechanisms and conventions (Honeycutt and Herring 2009; Marwick and boyd 2011), structure (Bruns and Moe 2013; Gaffney and Puschmann 2013), and history (van Dijck 2013; Rogers 2013). Computer science was the first to conduct research around Twitter (Java et al. 2007; Krishnamurthy, Gill, and Arlitt 2008), as the large volumes of temporal, network, linguistic,

[1] See also "Success Stories: Media, News & Publishing" (https://biz.twitter.com/success-stories/industry/media-news-publishing, accessed December 10, 2015), where Twitter pitches "success stories" of "curated content" to potential clients. However, note that success is shown by many different metrics. While this may be attributed to companies having different goals in their Twitter use, and Twitter itself gives a typology of strategies at https://business.twitter.com/, when we look at comparable products with presumably comparable goals (such as the "products" of Mitt Romney and Barack Obama's respective 2012 presidential campaigns), we see that even then different metrics are presented as evidence of success.

and even geographic data generated in aggregate by millions of users created rich opportunities to develop computational tools to process such data. Computer scientists also used the data as a test bed to describe and model phenomena such as tie formation, communities in networks, topic emergence, sentiment and opinion, and viral spread and other information flows (Cheong and Lee 2010). By 2013, there were already over a thousand papers from multiple disciplines, looking across tweet text, user behavior, or improving or building on the technology of Twitter (Williams, Terras, and Warwick 2013).

Despite this use of data, Twitter does not necessarily reflect the wider world. For example, a Pew study found that Twitter often does not match public opinion (Mitchell and Hitlin 2013). Also even when correspondences are found between Twitter data and the world at large, they can break down under slight changes in context (Cohen and Ruths 2013; Gayo-Avello 2012a, 2012b). Considering work on the users and culture of Twitter, this divergence is unsurprising (Ruths and Pfeffer 2014; Tufekci 2014) despite the platform creators' aspirations for Twitter to be a "neutral utility" (van Dijck 2013, 69).

There is work showing how Twitter users are not demographically representative using both representative surveys (Duggan et al. 2015) and comparisons of Twitter data to Census data (Hecht and Stephens 2014; Malik et al. 2015; Mislove et al. 2011). Then, there is work about Twitter's idiosyncratic conventions and cultural norms (boyd, Golder, and Lotan 2010; Java et al. 2007; Kwak et al. 2010), including an ugly side of enormous hostility towards and harassment of women and those of marginalized identities (Matias et al. 2015). Furthermore, even this work on the extent and specific types of biases in Twitter data are largely based on cross-sectional studies from the United States that may not generalize, as Twitter is itself neither globally uniform (Poblete et al. 2011) nor a static, stable environment across years (Liu, Kliman-Silver, and Mislove 2014).

Van Dijck (2013) pushes further, theorizing that work about Twitter's technology, users and usage, and content only considers the platform as a techno-cultural construct. Also critical is consideration of how social media platforms are socioeconomic structures, whose ownership, governance, and business models have important implications both for understanding the platforms at large but even for understanding the data produced on social media. For Twitter, we see how the possibility of making money from link farming (Ghosh et al. 2012) or from selling bots to inflate metrics (Donath 2007) has attracted spammers. Indeed, as anybody who analyzes Twitter data quickly finds, spam is widespread (Thomas et al. 2013), despite Twitter's attempts to filter it out (Thomas et al. 2011), and this can distort research findings (Ghosh et al. 2012). Also crucially, the most accessible channel of data (allowing for specific queries) that Twitter makes available for free, the Streaming Application Programming Interface (API) (Gaffney and Puschmann 2013), has strict rate limits within which sampling is not necessarily random (Morstatter, Pfeffer, and Liu 2014; Morstatter et al. 2013). This nonrandom sampling distorts not only absolute frequency (how often something appears on Twitter) but even relative frequencies (whether one thing or another is more frequent). An alternative, the Sample API, is a random sample

so frequencies are proportional to incidence Twitter overall; but at 1 percent of Twitter, and no ability to request data about specific users, hashtags, languages, etc., there is not enough statistical power to detect small phenomena.

In our analysis, we use access to the Twitter decahose, also known as the gardenhose (compared to the commercial firehose consisting of all public Twitter data). This is a scaled-up version of the Sample API (using the same method of sampling which successfully achieves randomness) that gives a 10 percent random sample of tweets (Kergl, Roedler, and Seeber 2014).

There are two other theoretical considerations that we do not consider but recognize. First is how social media platforms are not independent of one another but, through competition, shared users, and corporate and political links, form an "ecosystem of connected media" (van Dijck 2013, 18–23). Certainly, a given organization will coordinate its actions across multiple social media platforms (Bastos 2015; Skogerbø and Krumsvik 2015). Second is how social media platforms are designed to have user labor act as "affective processers" to produce data about users, data which are then stored—but not made accessible to users—in massive "archives of affect, sites of decontextualized data that can be rearranged by site owners to construct particular forms of knowledge about social media users" (Gehl 2014, 43). This relates to discussions of the blurring boundaries between citizen journalism and professional journalism (Ha ̈nska-Ahy and Shapour 2013), but also to larger questions of how the news media, with its agenda-setting power, is interacting with social media companies, with their power of managing and shaping what happens on their platforms.

Twitter and Journalism

Journalism literature about Twitter emerged in 2010, contemporaneously with other social science interest (Marwick and boyd 2011), with Hermida (2010) theorizing Twitter use as "ambient journalism." The possibility of using digital media as a way to "save" (Picard 2014) or perhaps transform journalism (Hermida 2013), makes it relevant to look at the consumption side of how people use Twitter for news, the production side of the Twitter strategies of news organizations, and the hybrid "prosumption" of interaction and feedback such as with dialogue and citizen journalism.

Research on news consumption includes surveys as well as case studies. For surveys, the Pew Research Center's Journalism Project carries out representative sampling within the United States, the most recent of which (Mitchell, Gottfried, and Matsa 2015) found 14 percent of surveyed internet-using 18–33 year olds getting political news from Twitter in a given week, compared to 9 percent of 34–49 year olds, and 5 percent of 50–68 year olds. Sixty-one percent of 18–33 year olds reported getting political news on Facebook in a given week, with respective percentages of 50 and 39 for 34–59 year olds and 50–68 year olds,

although it seems this is mostly incidental exposure (Mitchell, Gottfried, and Matsa 2015, 9). The survey also measured interest in political news, and knowledge and trust of various sources, and found less interest and knowledge in the younger group but no difference in trust. Nielsen and Schrøder (2014) used the results of the 2013 Reuters Digital News Survey (an online survey given to a sample of 1000 people each in Denmark, France, Germany, Italy, Japan, and Spain, and 2000 each in the United States and United Kingdom) and found that, overall, television is still the dominant source of news but social media is of growing importance. In Canada, Hermida et al. (2012) ran an online survey with a sample of 1600 Canadians and similarly found a minority (two-fifths) of social media users using social media as a source of news; but those that did were on average younger, and for more than two-thirds of users, access to news and views was a major motivation for their use of social media. In the United States, Holton et al. (2015) carried out an online survey with a national sample of 1813 Americans, finding that being engaged in reciprocal information exchanges on social media explains greater news consumption as well as content creation. Less literature looks at the interplay between consumers and news media on a large scale; one example is Hermida (2014, 136), who found (using a social media analytical tool Topsy Pro) that of five million tweets about the disappearance of Malaysian flight MH370, a full four million (80 percent) were retweets, with the remaining million mostly being "media organizations sharing the latest news."

Two early studies (Armstrong and Gao 2010; Holcomb, Gross, and Mitchell 2011) gathered and analyzed tweets to study how some specific US news organizations use Twitter. Subsequently, there has been literature about the production-side perspective of newsrooms both from major western outlets (Thurman and Walters 2013) as well as news media across the world (El Gody 2014; Mabweazara 2014; Mare 2014; Verweij and van Noort 2014).

The findings of Armstrong and Gao (2010) and Holcomb, Gross, and Mitchell (2011), conducted respectively from hand-coding of a four-month sample of tweets and of a one-week sample of tweets, were that news organizations by and large just tweet out headlines of articles with a corresponding link. Tweets were intended to drive traffic to news sites, and only a minority were intended as public service announcements (e.g., about road closings or hazardous weather). Almost no tweets solicited information (either to inform a story or feedback), including those from the most active news organizations. Holcomb, Gross, and Mitchell found almost no retweeting, and those retweets that did exist were of tweets belonging to the same news organization.

Subsequent work supports and extends these findings. Thurman and Walters (2013) studied Guardian.co.uk's use of Live Blogs; they found reporters there making use of information from Twitter and posting tweets. But the information flow was oneway, as Live Blogs were housed on The Guardian website and did not feed back into the Twitter ecosystem. Broersma and Graham (2013) also looked at the use of tweets as a source of evidence, arguing that reporters quoting tweets as evidence or examples of opinion alters the bal-

ance of power between journalists and sources, although without discussing the presence or absence of engagement between journalists and the people producing newsworthy tweets. Lawrence et al. (2014), in a study of all tweets from 400 political journalists during the 2012 US presidential campaign, similarly found that, despite the inclusion of opinion and expression, largely there was no greater transparency; the traditional ideas of gatekeeping had not been disrupted, and the journalism there existed in the same "bubble" as before.

There is, however, a difference between types of media outlets; Lasorsa, Lewis, and Holton (2012) used content analysis of 22,248 tweets in 2009 from the 500 mostfollowed journalists and found that for journalists on Twitter, those employed by less elite organizations were more comfortable "sharing their stage with other news gatherers and commentators." Transparency about everyday lives was similarly associated with journalists from less elite organizations, and Lasorsa (2012) found this associated more with female journalists than with male journalists.

Other work focuses on journalists, and has shown that reporters can have behavior distinct from the accounts of news organizations. Revers (2014) conducted a field study at the New York State Capitol building in Albany between 2009 and 2011, with interviews with 35 people, 300 hours of observations, and analysis of some 4492 tweets, finding variation in Twitter adoption as well as in the professional pressures associated with using Twitter (e.g., keeping up with competition in information gathering, or generating advertising revenue). Artwick (2013) addressed similar themes, using a sample of 2733 tweets from 51 journalists in 2011, finding that reporters engaged in a mix of "service," with reports tweeting public service announcements and retweeting citizen voices, and "product," with reporters self-promoting by linking to newsroom content they had produced. Vis (2013) used a Guardian/Twitter database of 2.6 million tweets on the 2011 UK riots, collected from the Twitter firehose, to identify the top 1000 most-mentioned accounts; in addition to analyzing the makeup of this sample, Vis used the REST API (see Gaffney and Puschmann 2013) to collect all tweets for two journalists on which to conduct content analysis. She found that accounts of mainstream media received the greatest proportion of mentions, which she relates to findings about the dominance in conversations of a few sources of information. The content analysis showed a variety of activities undertaken by the journalists, but the plurality (about a third) of tweet content was the two journalists' own eyewitness reporting.

While much analysis (ours included) looks at the most visible entities, a special issue of Digital Journalism looked at social media practices in newsrooms in Africa. Bosch (2014) covered three community radio stations in South Africa, finding that their use of social media has improved news-gathering as well as increased access and participation for audiences. Also in South Africa, Verweij and van Noort (2014) looked at networks among 500 journalists, finding high density among an elite group but low amounts of connection otherwise, similar to patterns found in the United Kingdom and the Netherlands. El Gody (2014) described how ICTs are used in daily routines in three established Egyptian newsrooms; de-

spite pressure from Egyptian audiences organizing into networks with their own communication systems, he found cases of management prohibiting stories based mainly on internet information, and of journalists treating ICTs as tools for personal use or to pass time. From a field study in Mozambique, conducted at a newly founded community newspaper, Mare (2014) observed innovative uses, including interacting with audiences for feedback and getting information (although noting that the newspaper is new and thus the success of its strategies is yet to be proven).

Guided by existing work, we pursue two research directions. First, what sort of Twitter usage does our large-scale data support? Second, what news organizations are associated with the greatest share of content and activity? We would hypothesize that our approach will give evidence supporting previous literature, and specifically,

H1: News media currently use Twitter in largely a "push" model, rather than for transparency or dialogue.

H2: Large news organizations continue to dominate.

We also investigate something not found in literature thus far: the difference in focus between news media content on Twitter, and Twitter content at large. Social media represents, on the one hand, a kind of public opinion that, as previous work has shown, is determined by news media (McCombs and Shaw 1972), and, on the other hand, is a kind of media in itself that, as from research on inter-media agenda-setting (Golan 2006), may be something from which news media take cues (Dearing and Rogers 1996, 33). But, conditional on H1 being true, we would assume that news media do not interact with social media enough to be influenced, so either there is little relationship between news media content on Twitter and content at large, or else news media focus precedes larger attention. This leads to our third hypothesis,

H3: On Twitter, news media focus on a set of topics distinct from Twitter at large.

Method

Data Collection: News Organizations on Twitter

We combined an online list of news organizations and journalists from Alexa[2] white pages with an extensive search through online listings of global news organizations, using a series of heuristics (looking for "twitter.com/..." in the HTML of the website homepage, collecting whatever comes after the slash, and manually reviewing and revising the results), resulting in 6103 Twitter handles that we identified as belonging to news outlets or journalists.

[2] Accessed August 16, 2015.

We recognize first that our list has not been independently verified, for which reason we make this list available for download,[3] and second that it is not necessarily appropriate to have a category of "news organizations." International and western news media may be very different from news media in Latin America, Africa (Mabweazara 2014), and the Asia-Pacific region; local news media may be different from regional and national news media; and there may be differences in the Twitter strategies and behavior of newspapers compared to television, or either one compared to multimedia news outlets. The category is also complicated by the emergence of citizen-journalists, "blogs," and digital media organizations such as Gawker, Mashable, The Huffington Post and Buzzfeed that (in differing amounts) mix entertainment and commentary with news and investigative reporting. We allow a broad and inclusive definition of professional journalism that includes such cases.

Furthermore, given that we cannot guarantee that we have a comprehensive list of news media and journalists (even given some boundary definition of news media and journalism), we likely underestimate results. However, for volume of news content (see below for definition), the top 100 news media-related accounts account for 65.7 percent of all tweets, such that the many smaller organizations and individual journalists we undoubtedly miss (the "tail" of the rank list) have little impact on the aggregate figures.

Data Collection: Twitter

From Twitter's decahose (described above), we collected two different datasets, both within the time period from March 1 to June 30, 2014; first, all 1,783,704,266 English-language tweets (filtering by "lang = 'en'" in the tweet meta-data), and second, a set of 100,000 tweets containing the hashtag "#Egypt" (for a description of raw Twitter data and how it may be manipulated, see Kumar, Morstatter, and Liu 2014). As a way of looking not just at mentions, or tweets, or links to news websites, we propose combining these three aspects to classify a tweet as news media-related or not using the following inclusion criteria:

a) A tweet is sent by news media. We collected 6103 Twitter users from websites of news outlets and from online lists including news media and journalists and use this list to determine whether a tweet was sent by news media or not.
b) A tweet mentioning news media. Tweets can mention other Twitter users. We looked for tweets that mention at least one of the 6103 news media Twitter users, e.g., a retweet of a tweet originally sent by a news media account.
c) A tweet linking to news media. From alexa.com we extracted 6535 URLs from the news category.[4] If a tweet contains a link to (the domain of) one of these websites we catego-

[3] The list of news media organization Twitter handles used in this article can be accessed at http://www.pfeffer.at/data/news-on-twitter/.
[4] From http://www.alexa.com/topsites/category/Top/News (accessed August 16, 2015).

rize the tweet as news media-related. We use the "expanded_url" from the tweet metadata but we do not de-code third-party shortened URLs (e.g., bit.ly).[5]

Importantly, this excludes newsworthy tweets, for example tweets from nonjournalists or citizen journalists that are important in actually breaking a news story. As our concern is with the behavior of professional journalism, we look only at tweets connected with news media and not tweets relating to news more broadly. Also, we undoubtedly miss accounts of individual journalists, news shows, and editors; this is not to say that these are not important, only that our macroscopic perspective is not the appropriate methodology for describing their importance.

Also importantly, we did not choose the time period to include any specific news stories or events, which means that we did not consciously introduce bias towards some "high-news" period. However, after looking at events taking place during that period, we decided to focus on Egypt-related tweets for the second part of the analysis because of turmoil around general elections stretching across these months.

Topic Modeling

We extracted 104,698 tweets containing the hashtag "#Egypt" (as there were several events involving Egypt over the time period considered) and applied what is called a "topic model" to see if different clusters (topics) show different levels of involvement by news media. Topic modeling is almost synonymous with the most popular technique used to do it (Schmidt 2013), Latent Dirichlet Allocation (LDA), invented in 2003 by Blei, Ng, and Jordan (2003).[6] LDA takes in a collection of documents and outputs several clusters of words in which each cluster is supposed to capture a "topic" present across the documents. The analyst will look at the various clusters and pick a descriptive name for the topic they think the words represent.[7] The main selling point of LDA is that it can take in huge amounts of text

[5] This means we miss tweets linking to news media websites through these short- ened URLs, so we likely undercount the number of tweets with news URLs. For reference, there are 37,462,621 bit.ly links, 8,713,644 ow.ly links, 7,988,958 goo.gl links, and 7,714,077 tinyurl links among the 1.8 billion collected tweets. Note that since February 2013, Twitter forcibly displays all URLs in 23 characters using its own "t.co" shortening service (see "Twitter Now Reducing Some Tweets to 117 Characters," http://mashable.com/2013/02/20/twitter-tco-length/, and "Posting Links in a Tweet," https://support.twitter.com/articles/78124, accessed December 10, 2015), removing the incentive that users would have previously had to use URL shortening services to save space (Antoniades et al. 2011; Grier et al. 2010; Maggi et al. 2013; Wang et al. 2013). The one paper analyzing URL shortening services after the changeover (Gupta, Aggarwal, and Kumaraguru 2014) does not consider the uses of such services other than for spam, and do not have an esti- mate of what proportion of tweets with shortened URLs are spam.

[6] The eponym Dirichlet is of the nineteenth-century German mathematician; a probability distribution based on his work was named after him, and this proba- bility distribution is the basis for LDA.

[7] This part of the process is very similar to a technique common in social science, Principle Component Analysis, in how the analyst interprets what loadings represent.

data, and output topic clusters that are reasonable.

Important to note is that LDA is a "bag-of-words" approach and uses only frequencies and co-occurrences and not semantics, context, or any structural properties of language; as such, it does not and cannot extract meaning, but it is reasonable to use the presence of certain words as proxies for what a person would (subjectively) identify as topics when reading through text. The subjectivity means that caution is required; studies have shown that, when people are inclined to see structure, they will interpret random collections of words as representing a coherent topic (Zhu, Gibson, and Rogers 2009). One study asks whether interpreting LDA outputs is akin to "reading tea leaves" (Chang et al. 2009). However, validations of LDA have shown that, while the clusters found by LDA are not the same as what human readers code, there is reasonable correspondence (Morstatter et al. 2015). We avoid some of the interpretational problems by not trying to interpret the clusters found by LDA; we are primarily interested in differences across clusters related to news media presence.

Results

Applying the above-mentioned three criteria in order to identify news mediarelated tweets, we identified 14,276,925 of 1,783,704,266 tweets (0.800 percent) as news media-related.[8] Table 1 shows more details about this analysis step and Figure 1 illustrates the percentage of news media-related tweets per week (daily minimum 0.516 percent, daily maximum 1.105 percent, SD = 0.134 percent) showing relatively stable values between 0.75 and 0.85 percent over time.

TABLE 1
Tweets related to news media

Criteria	Number of Tweets	Percentage of Twitter total
Tweets from news media handles	576,316	0.032%
Tweets mentioning news media	10,909,971	0.612%
Tweets with news media URLs	3,390,664	0.190%
All	14,276,925	0.800%

[8] Note that for estimating just this proportion, it was not necessary to have a sample as large as the decahose; we could have also estimated this with the Sample API, or even 1/10,000th of the Sample API. The advantage of having the full data, or a far larger sample, is in being able to get accurate estimates of observations in the tail of the rankings, as here observations are very sparse and hence it is much harder to get accurate estimates from smaller samples. Furthermore, large samples allow for drawing subsamples (such as the one we take for #Egypt) large enough to perform meaningful inference.

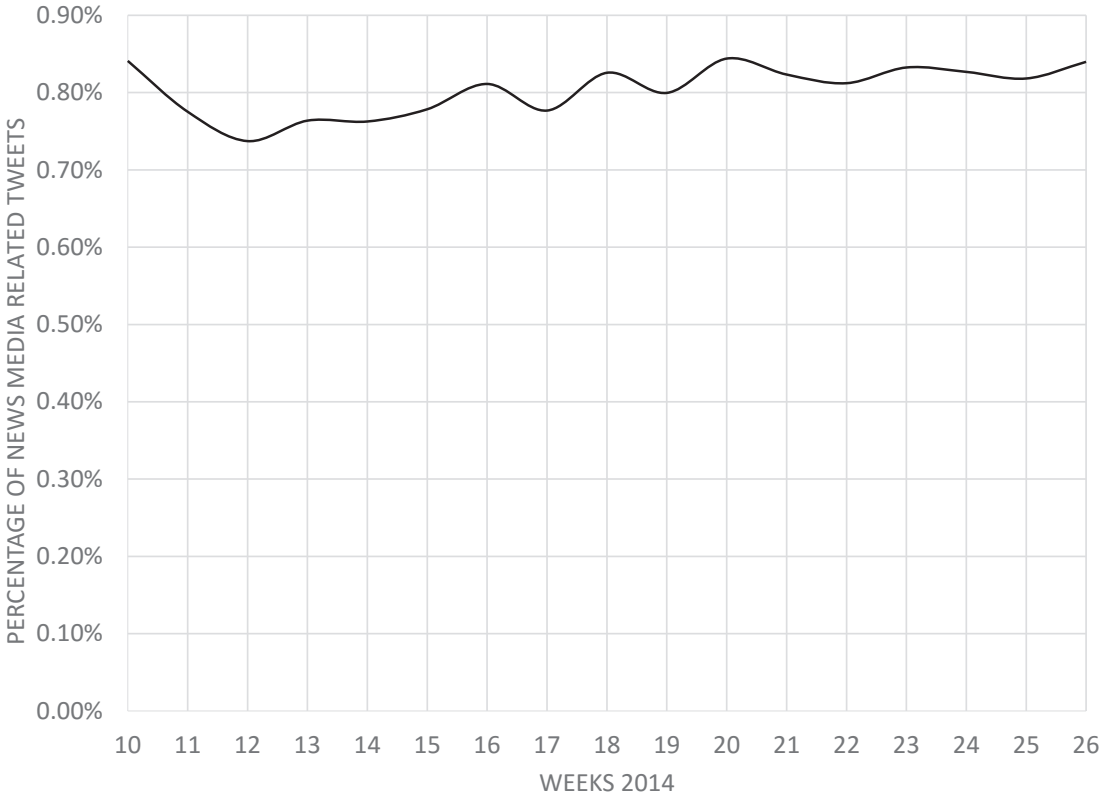

FIGURE 1
Weekly aggregation of percentage of news media-related tweets.

The overall number of tweets that are news media-related is low, but respectable considering that we have about 6000 accounts out of more than 316 million.

Next, we want to analyze these news media-related tweets in more detail to find possible differences based on topics. A straightforward approach of identifying topics on Twitter is to use hashtags, as they are created to be a machine-readable (and therefore easily identified) way for Twitter users to define a "topic" and communicate around it. For all hashtags in the approximately 1.8 billion tweets we count occurrence and calculate the percentage of tweets per hashtag that are news media-related. Figure 2 shows the top 656 hashtags that occur at least one time per 5000 tweets (=0.005 percent) in our data. The x-axis shows the number of tweets and the y-axis represents the percentage of news media-related tweets.

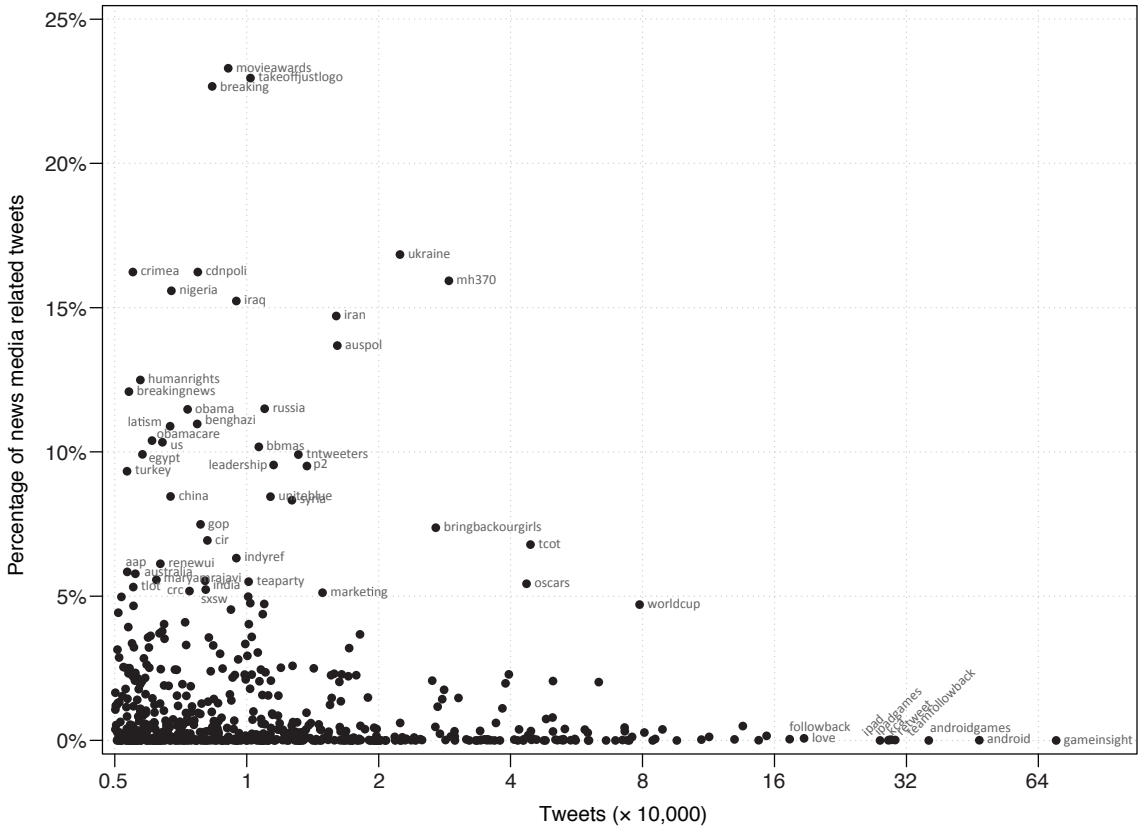

FIGURE 2
Top topics and percentage of tweets related to news media.

Five different aspects of this graph are striking. First, many of the topics with more than 5 percent news media-related tweets are related to countries in conflict including Arab Spring countries in northern Africa and the Middle East but also Russia, Ukraine and Scotland. For Scotland, we identified #indyref as referring to the Scotland independence referendum taking place in September 2014. Secondly, we can find several political campaigns that receive high media attention (#auspol, #uniteblue). Thirdly, the missing aircraft flight Malaysia Airline MH370 can be seen as example of topics where all news available comes mediated through news media and no real information comes from people "on the ground;" this matches what Hermida (2014) found for all original tweets around MH370 coming from news media, and is potentially an example of what Vasterman (2005) calls "media-hypes," self-perpetuating coverage that continues despite not having anything new to report. Fourth, the most used hashtags in our data are not news media-related at all and discuss primarily mobile games. Lastly, there are the three outlier hashtags with the highest proportion of use by news media. #breaking is understandable; it is a journalistic convention, adapted to Twitter, but likely used far less by non-news entities. #takeoffjustlogo is a protest around a designer using a logo similar to a sacred Sufi symbol

on a perfume.[9] #movieawards, referring to the MTV Movie Awards, is attributable to news media-related activity around the single entity of MTV. These results relate to the findings of Bastos (2015) who finds, for The Guardian and The New York Times, differing emphasis by topic from readers and editors. It also relates to the findings of Kwak et al. (2010), from a study of complete Twitter data crawled in 2009, about overlap between but differences in top hashtags used in Twitter and top topics of coverage in news media.

"Push" Model of Twitter Use

As much of the activity is accounted for by mentions and URLs, we were curious to investigate the extent to which this activity is self-generated linkage, addressing the first hypothesis about news media using Twitter largely in a "push" model. Indeed, as discussed above, news organizations use Twitter to refer to their own stories (Armstrong and Gao 2010; Holcomb, Gross, and Mitchell 2011; Vis 2013). From our list of news media Twitter accounts we took the top 51 accounts, accounting for ~50 percent of all news media-related tweets (as defined above) in our data, and extracted a random set of 1000 tweets from these 51 users to manually inspect. Confirming that previous results have held since studies carried out in 2009, we found that the vast majority (89.7 percent of all tweets from these top news media accounts) have a URL to the website of the respective sender of the tweet or mention a Twitter account associated with the sending organization. That is, the accounts with the highest associated volume of news media-related tweets are indeed largely self-references. This can be seen as an indicator that most of these tweets were written to drive traffic to the individual news sites, that is, news media employ Twitter as mere news dissemination tool. Hence, retweets of these tweets are a significant factor in the large amount of news media-related tweets for these organizations described above.

From our selection of 51 users, 2.8 percent of tweets point to other news sources and another 1.0 percent point to the sending organization in the form of hashtags for a specific show (without Web link or mention). Just 6.5 percent of tweets are not related to news media's websites, Twitter accounts, or hashtags, and these tweets come from a very small number of accounts, e.g. MTV linking to artists. In addition, we found not a single tweet among the 1000 that did not have at least one hashtag, mention, or URL. This supports previous findings about how news organizations have interpreted "professionalism" on Twitter to mean using Twitter affordances to maximum effect (Armstrong and Gao 2010; Lawrence et al. 2014). It has not meant adopting the cultural norms of Twitter even (or especially) when they conflict with the traditional practice of journalism, nor the abandonment of professionalism as an ideal in favor of transparency or dialogue (Hornmoen and Steensen 2014; Lawrence et al. 2014).

[9] See http://www.theguardian.com/fashion/2014/may/29/roberto-cavalli-perfume-offends-sufi-students (accessed December 10, 2015).

Relative Dominance

Addressing the second hypothesis, we ask, what is the relative dominance of established organizations across this news media-related behavior? We find a characteristically skewed "long-tailed" plot. That is, a small number of news entities are responsible for the majority of activity. Looking only at the subset of news media-related tweets pointing to URLs of news organizations, the distribution is similar; a link to the bbc.co.uk domain appears in 166,449 tweets and a link to theguardian.com domain appears in 158,950 tweets.

Figure 3 shows the "rank-frequency" plots (Brookes and Griffiths 1978) in logarithmic scale. The figure shows how the vast majority of news media only have a very small number of news media-related tweets and links to their respective organizations' URLs, whereas the accounts associated with the most Twitter activity dwarf all others. Cumulatively, the top 10 handles alone account for 19 percent of news media-related activity, the top 25 for 34 percent, the top 50 for 49 percent, the top 100 for 65 percent, and the top 500 for 95 percent. That is, the remaining 5603 of the 6103 handles account for only 5 percent of the volume of news media-related tweets.

From Figure 3a we see that, unsurprisingly, sports and entertainment news (ESPN, MTV, Billboard) and new, digital news media (Mashable, The Huffington Post) are associated with the largest volume of news media-related content. However, The Guardian, The New York Times, the Associated Press, CNN, and Forbes also have large amounts tweets related to their content. This is in contrast to other traditional news organizations, such as the Washington Post and Wall Street Journal, which have almost half the number of news media-related tweets associated with them (respectively, 87,034 and 85,494 tweets, the 17th and 18th highest volumes). In a similar comparison between news wires, we can see that Reuters has less news media-related content associated with it than does the Associated Press, as @reuters has 59,388 associated tweets, less than half of @ap. CNN has higher associated volumes than other television news: @foxnews has 89,101 associated tweets, @abc has 78,402, and @nbcnews has 50,906. When considering that CNN also has @cnnbrk with 94,767 tweets, we see that it accounts for a large proportion of Twitter activity. The BBC is a notable absence among the accounts with the most associated news media-related tweets, but this is because it has employed a strategy of dividing up its Twitter activity between multiple accounts. @bbcsport has 99,588 associated tweets, ranking 14th; @bbcworld has 72,900 tweets, ranking 22nd; @bbcbreaking has 59,769 tweets, ranking 25th; and @bbcnews has 55,473 tweets, ranking 29th. With these accounts together, the BBC has more associated tweets than does ESPN. It is important to recall that the data is a random 10 percent sample of all of Twitter. Consequently, these numbers can be used to estimate the overall volume on Twitter by multiplying by 10.

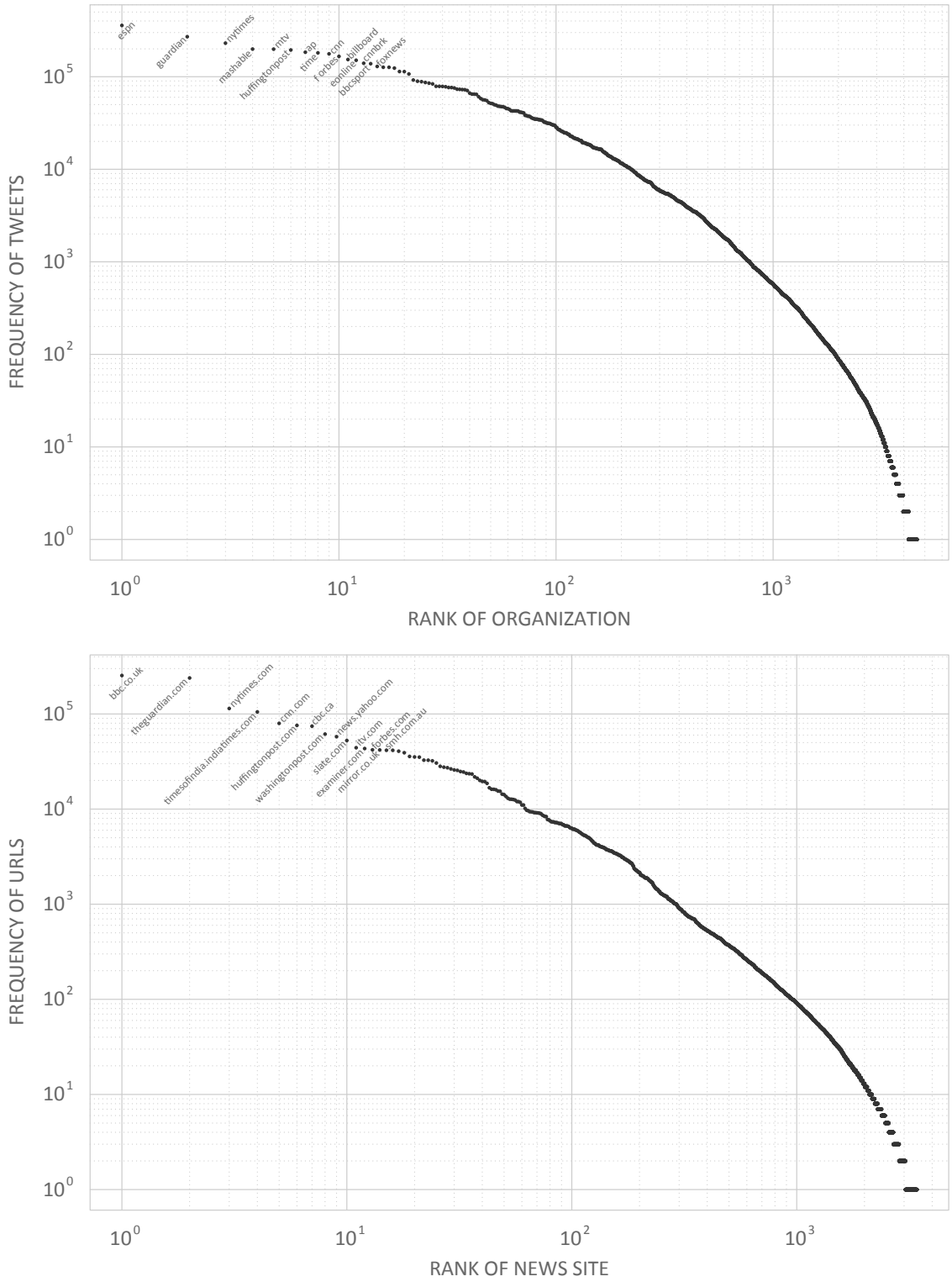

FIGURE 3
Rank-frequency plots of (a) highest volume tweeters and (b) highest linked to news agency websites.

When looking only at URLs, there are slight differences; while English-language media are dominated by American and British outlets, there is one non-western media outlet that appears high on the list, with the Times of India having almost as many URLs as The New York Times (versus @timesofindia ranking 37th in overall volume of associated news media-related tweets with 49,202 tweets). Overall, the news media with the most frequent URLs are dominated by traditional organizations (other than The Huffington Post and Yahoo News), perhaps mostly retweets of news media tweets.

Local Analysis: Egypt

To analyze the importance of news media in more detail, and investigate our third hypothesis about whether Twitter content takes cues from news media or vice versa beyond differing hashtag distributions, we selected a hashtag related to one country in conflict and performed additional analysis. Identifying topics of substantive interest that occurred through our period of data collection, we chose Egypt as a case study, as this time included the contentious overthrow of the previously elected government of the Muslim Brotherhood and transition to a government headed by the general Abdel Fattah el-Sisi (Borge-Holthoefer et al. 2015; El Issawi 2014). We identified 104,698 tweets from March 1 to June 30, 2014 including the hashtag "#Egypt."

With the three above-described approaches to detect news media-related tweets we were able to identify 8474 tweets (8.094 percent); 391 (0.373 percent) sent by news media, 5195 (4.962 percent) mentioning news media, and 3276 (3.129 percent) linking to news media websites. Figure 4 shows that news media-related tweets are stable between 6 and 10 percent per week over the four-month data collection period.

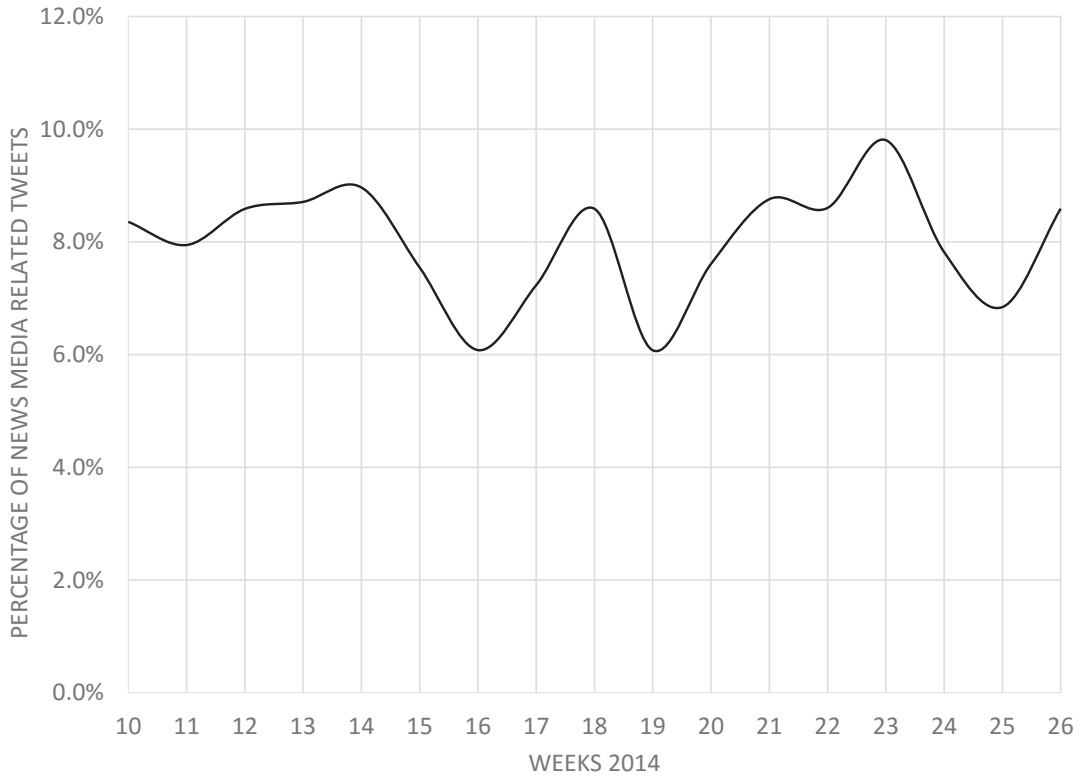

FIGURE 4
News media-related tweets per week for tweets including the hashtag “Egypt”

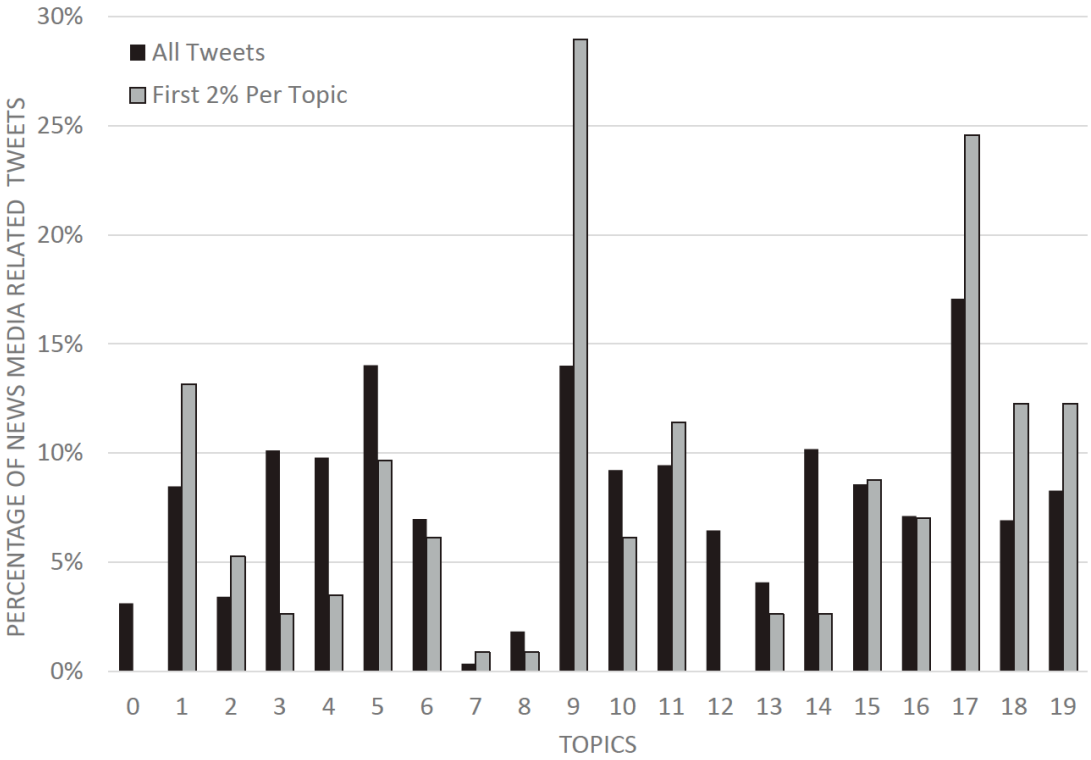

FIGURE 5
Percentage of news media-related tweets in #Egypt subtopics

We applied LDA, as described above, with 20 topics to the #Egypt data and counted the news media-related tweets in these #Egypt subtopics. The black bars in Figure 5 visualize these results. As one can see, news media-related tweets are very different in these subtopics and go from almost 0 percent (topic 7) to 17 percent (topic 17). (In LDA output, the numbers assigned to clusters are just an indexing tool, they do not say anything about the clusters.)

To better understand the role of news media in these subtopics, we use the timestamps of the tweets in these 20 topics and count news media-related tweets in the earliest 2 percent of all tweets per topic. By looking just at the earliest tweets, we hope to approximate the first adopters of the topics. Note that early tweets cannot be said to be the source or cause; meaningfully modeling influence is an enormously difficult problem, as there are many confounding factors. One thing we can say is that being an "early adopter" is a necessary but not sufficient condition for influence.

The gray bars in Figure 5 show these results. Intriguingly, two topics reach almost 25 percent (topic 17) and 30 percent (topic 9). We do not analyze all topics in detail but want to have a closer look at these two topics. A central feature of LDA is that it gives a set of top words that are most important for the topic clustering and, consequently, most descriptive for the topics. Table 2 shows the top 10 words for the clusters. In particular, the words in clusters 9 and 17 point to a very specific incident: three Al-Jazeera journalists, charged with terrorism and imprisoned on December 29, 2013, were found guilty on June 23 in a trial that received international condemnation (BBC News 2014).

Although LDA is exploratory, we feel comfortable saying that within Egypt, we identify that news organizations' Twitter outputs focus disproportionately on a specific event, apart from the general level of interest. It is telling but unsurprising that the event relates to journalism and press freedom.

TABLE 2
Top words of LDA clusters of the #Egypt data.

Topic	Top Words
1	#news read power #sinai #egyptian top fresh torture latest
2	#anticoup today coup military university forces cairo students security
3	world military democracy regime aid freedom press state #africa
4	day presidential elections #sisi sisi #egypt's election vote #egyelections
5	president sisi #egypt's #sisi sexual egypt's chief harassment brother
6	#egypt's justice man action back minister historic country madness
7	#iran iran #news #world #cnn #un #usa #london #fox
8	#uae #saudi #kuwait #ksa #ff free #bahrain #qatar watch
9	#freeajstaff court trial crime today journalism #ajtrial mohamed photo
10	political innocent iraq maliki group report calls #egypt's killer
11	egyptian news women media tv show state pm #gaza

12	#syria #iraq #libya #ukraine #israel #palestine foreign arab #lebanon
13	#cairo time good great morning #giza young happy hope
14	brotherhood muslim #mb_europe supporters #mb mb leader terrorist terrorism
15	police killed army &amp cairo protesters breaking shot killing
16	support #maryamrajavi government campaign call million sign #humanrights committed
17	journalists years al prison days journalist jail jazeera jailed
18	protest rights law human #rnn el live life yesterday
19	death sentenced people mass stop sentences executions speak sentence
20	egypt #travel #photography #tourism #discover_egypt_come #art #design #journey #welcometoegypt

Conclusions

It is difficult to say what we should have expected for the volume of tweets that are news media-related; on the one hand, only about 6000 accounts out of 300 million "active monthly users"[10] would, if we take a uniform distribution as our null model, predict only 0.0019 percent tweets being sent, whereas we see 0.032 percent of the volume of tweets coming from our news accounts. Given that Twitter in general has extremely skewed distributions of activity, we could more accurately say that news accounts are over-represented among frequent tweeters and heavy users. Alternatively, if we take other large organizations as the more appropriate comparison set for large news organizations, then news media has a good showing but are not among the most dominant entities. Taking our more expansive notion of news media-related tweets that include mentions and URLs, the finding that 0.8 percent of the total volume relates to news media is an interesting alternative way to consider the importance of news media to Twitter (e.g., versus considering only overlap in topical focus, as in Kwak et al. (2010); it is hard to contextualize in itself, but may be a baseline for future study, to see if the proportion changes over larger periods of time as more users join Twitter; as smaller media organizations increase adoption and usage of Twitter; or as existing patterns of usage by general users or by news media-related entities changes. Alternatively, we also present the number of tweets, mentions, and URLs, for studying any one of these independently over time.

Our study is a compliment to the smaller studies previously done; using largescale computational analysis is not as accurate as content analysis and certainly not as informative about actual practice as field studies, but by using a very different methodology to arrive at the same conclusions, we strengthen and reinforce the findings of previous work in terms of the two hypotheses we consider. First, across a range of organizations, wider than just those in the case studies previously considered, we find continued evidence that Twitter practices of news media consist largely of dissemination. There is little evidence for engagement that breaks previously established journalistic norms such as professionalism, such as by prioritizing transparency or dialogue. Second, we found (in something not yet

[10] See https://about.twitter.com/company (accessed December 10, 2015).

explicitly demonstrated or quantified), the inequality of the distribution of news content over different organizations, with large, established organizations accounting for most news media-related content.

For looking at the direction of agenda-setting in the social media environment, we first found differentiation: over a randomly chosen period in 2014, news media-related content was focused on a different set of hashtags than Twitter in general, often with far greater representation of such content. From topic modeling, we were able to look within a specific topic area, that of Egypt across a coup and government change, to see how certain subtopics—in particular, ones relating to reporting and journalism itself—have a far higher proportion of content from, directed at, or linking to news entities, and that news media were the earliest to focus on these topics.

The difference between news media and other content also connects back to one of our original theoretical motivations: to look at news consumption and dissemination activities on Twitter not just in terms of affordances and emergent practices, but also in terms of the larger context of the Twitter platform. Computational methodologies that aggregate and analyze at a large scale provide an important perspective on global context, exploiting for research purposes the ways in which Twitter as a platform is designed to enable commodification and data-mining (Gehl 2014). This suggests further avenues of investigation around how user consumption splits between news and non-news content. Further investigating the role played by news media in Twitter may also demonstrate how the presence of news media is valued by users, thereby creating value for the Twitter platform. Again recalling how Twitter is a commercial enterprise with a private governance structure and not a neutral public utility (van Dijck 2013), knowledge of the relative value of news media will help anticipate how such media may be treated in the future by the platform, and perhaps advocate for specific privileges built into the platform, and generally support the pursuit by news media of reach, impact, and influence on digital media platforms.

ACKNOWLEDGEMENT

Thanks to Cornelia Brantner for discussion and guidance around journalism theory.

DISCLOSURE STATEMENT

The authors have no affiliation with Twitter, any commercial aggregator of Twitter content, or any news organization, and do not have any financial interest or benefit arising from the direct applications of this research.

FUNDING

This work was supported in part by the Office of Naval Research under MINERVA [grant number N000141310835]. Momin is supported in part by a grant from the ARCS Foundation.

REFERENCES

Antoniades, Demetris, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, Sotiris Ioannidis, Evangelos P. Markatos, and Thomas Karagiannis. 2011. "We.B: The Web of Short URLs." In *Proceedings of the 20th International World Wide Web Conference (WWW 2011)*, 715–724. New York: ACM Press.

Armstrong, Corey L., and Fangfang Gao. 2010. "Now Tweet This: How News Organizations Use Twitter." *Electronic News* 4 (4): 218–235. doi:10.1177/1931243110389457.

Artwick, Claudette G. 2013. "Reporters on Twitter: Product or Service?" *Digital Journalism* 1 (2): 212–228. doi:10.1080/21670811.2012.744555.

Bastos, Marco Toledo. 2015. "Shares, Pins, and Tweets: News Readership from Daily Papers to Social Media." *Journalism Studies* 16 (3): 305–325. doi:10.1080/1461670X.2014.891857.

BBC News. 2014. "Egypt Trial: Outcry over Al-Jazeera Trio's Sentencing." BBC News. June 23. http://www.bbc.com/news/world-middle-east-27982732.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (March): 993–1022.

Borge-Holthoefer, Javier, Walid Magdy, Kareem Darwish, and Ingmar Weber. 2015. "Content and Network Dynamics behind Egyptian Political Polarization on Twitter." In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*, 700–711. New York: ACM Press. doi:10.1145/2675133.2675163.

Bosch, Tanja. 2014. "Social Media and Community Radio Journalism in South Africa." *Digital Journalism* 2 (1): 29–43. doi:10.1080/21670811.2013.850199.

boyd, danah, Scott Golder, and Gilad Lotan. 2010. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences (HICSS-43)*, 1–10. Los Alamitos, CA: IEEE Computer Society. doi:10.1109/HICSS.2010.412.

Broersma, Marcel, and Todd Graham. 2013. "Twitter as a News Source: How Dutch and British Newspapers Used Tweets in Their News Coverage, 2007–2011." *Journalism Practice* 7 (4): 446–464. doi:10.1080/17512786.2013.802481.

Brookes, Bertram C., and Jose M. Griffiths. 1978. "Frequency-Rank Distributions." *Journal of the American Society for Information Science* 29 (1): 5–13. doi:10.1002/asi.4630290104.

Bruns, Axel, and Hallvard Moe. 2013. "Structural Layers of Communication on Twitter." In *Twitter and Society*, edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt,

and Cornelius Puschmann, 15–28. Digital Formations. New York: Peter Lang.

Castillo, Carlos, Mohammed El-Haddad, Jürgen Pfeffer, and Matt Stempeck. 2014. "Characterizing the Life Cycle of Online News Stories Using Social Media Reactions." In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*, 211–223. New York: ACM Press. doi:10.1145/2531602.2531623.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, 288–296. Curran Associates, Inc. http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

Cheong, Marc, and Vincent Lee. 2010. "Dissecting Twitter: A Review on Current Microblogging Research and Lessons from Related Fields." In *From Sociology to Computing in Social Networks*, edited by Nasrullah Memon and Reda Alhajj, 343–362. Vienna: Springer Vienna. http://www.springerlink.com/index/10.1007/978-3-7091-0294-7_18.

Chouliaraki, Lilie, and Bolette Blaagaard. 2013. "Introduction: Cosmopolitanism and the New News Media." *Journalism Studies* 14 (2): 150–155. doi:10.1080/1461670X.2012.718542.

Cohen, Raviv, and Derek Ruths. 2013. "Classifying Political Orientation on Twitter: It's Not Easy!" In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13)*, 91–99. Palo Alto, California: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6128.

Cozma, Raluca, and Kuan-Ju Chen. 2013. "What's in a Tweet? Foreign Correspondents' Use of Social Media." *Journalism Practice* 7 (1): 33–46. doi:10.1080/17512786.2012.683340.

Dearing, James W., and Everett Rogers. 1996. Agenda-Setting. *Communication Concepts* 6. Thousand Oaks: SAGE Publications, Inc.

van Dijck, José. 2013. *The Culture of Connectivity: A Critical History of Social Media.* New York: Oxford University Press.

Donath, Judith. 2007. "Signals in Social Supernets." *Journal of Computer-Mediated Communication* 13 (1): 231–251. doi:10.1111/j.1083-6101.2007.00394.x.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. "Demographics of Key Social Networking Platforms." Pew Research Center: Internet, Science & Tech. http://www.pewinternet.org/2015/01/09/demographics-of-key-socialnetworking-platforms-2/.

El Gody, Ahmed. 2014. "The Use of Information and Communication Technologies in Three Egyptian Newsrooms." *Digital Journalism* 2 (1): 77–97. doi:10.1080/21670811.2013.850202.

El Issawi, Fatima. 2014. "Egyptian Media under Transition: In the Name of the Regime... in the Name of the People?" LSE Research Online Documents on Economics. London School of Economics and Political Science, LSE Library. http://eprints.lse.ac.uk/59868/.

Engesser, Sven, and Edda Humprecht. 2015. "Frequency or Skillfulness: How Professional News Media Use Twitter in Five Western Countries." *Journalism Studies* 16 (4): 513–529. doi:10.1080/1461670X.2014.939849.

Flaounas, Ilias, Omar Ali, Thomas Lansdall-Welfare, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2013. "Research Methods in the Age of Digital Journalism: MassiveScale Automated Analysis of News-Content—Topics, Style and Gender." *Digital Journalism* 1 (1): 102–116. doi:10.1080/21670811.2012.714928.

Franklin, Bob. 2012. "The Future of Journalism: Developments and Debates." *Journalism Studies* 13 (5-6): 663–681. doi:10.1080/1461670X.2012.712301.

Gaffney, Devin, and Cornelius Puschmann. 2013. "Data Collection on Twitter." In *Twitter and Society*, edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, 55–68. Digital Formations. New York: Peter Lang.

Gayo-Avello, Daniel. 2012a. "'I Wanted to Predict Elections with Twitter and All I Got Was This Lousy Paper' — A Balanced Survey on Election Prediction Using Twitter Data." ArXiv:1204.6441, April. http://arxiv.org/abs/1204.6441.

Gayo-Avello, Daniel. 2012b. "No, You Cannot Predict Elections with Twitter." *IEEE Internet Computing* 16 (6): 91–94. doi:10.1109/MIC.2012.137.

Gehl, Robert W. 2014. *Reverse Engineering Social Media: Software, Culture, and Political Economy in New Media Capitalism*. Philadelphia, PA: Temple University Press.

Ghosh, Saptarshi, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. 2012. "Understanding and Combating Link Farming in the Twitter Social Network." In *Proceedings of the 21st International Conference on World Wide Web (WWW 2012)*, 61–70. New York: ACM Press. doi:10.1145/2187836.2187846.

Golan, Guy. 2006. "Inter-Media Agenda Setting and Global News Coverage: Assessing the Influence of the New York times on Three Network Television Evening News Programs." *Journalism Studies* 7 (2): 323–333. doi:10.1080/14616700500533643.

Grier, Chris, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. "@Spam: The Underground on 140 Characters or Less." In *Proceedings of the 17th ACM Conference on Computer and Communications Security (CCS '10)*, 27–37. New York: ACM Press. dpi:10.1145/1866307.1866311.

Gupta, Neha, Anupama Aggarwal, and Ponnurangam Kumaraguru. 2014. "bit.ly/malicious: Deep Dive into Short URL Based E-Crime Detection." In *Proceedings of the 2014 APWG Symposium on Electronic Crime Research (ECrime)*, 14–24. Los Alamitos, CA: IEEE Computer Society. doi:10.1109/ECRIME.2014.6963161.

Hänska-Ahy, Maximillian T., and Roxanna Shapour. 2013. "Who's Reporting the Protests? Converging Practices of Citizen Journalists and Two BBC World Service Newsrooms, from Iran's Election Protests to the Arab Uprisings." *Journalism Studies* 14 (1): 29–45. doi:10.1080/1461670X.2012.657908.

Hecht, Brent, and Monica Stephens. 2014. "A Tale of Cities: Urban Biases in Volunteered Geographic Information." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 197–205. Palo Alto, CA: AAAI Press. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8114.

Hermida, Alfred. 2010. "Twittering the News: The Emergence of Ambient Journalism." *Journalism Practice* 4 (3): 297–308. doi:10.1080/17512781003640703.

Hermida, Alfred. 2012. "Tweets and Truth: Journalism as a Discipline of Collaborative Verification." *Journalism Practice* 6 (5-6): 659–668. doi:10.1080/17512786.2012.667269.

Hermida, Alfred. 2013. "#Journalism: Reconfiguring Journalism Research about Twitter, One Tweet at a Time." *Digital Journalism* 1 (3): 295–313. doi:10.1080/21670811.2013.808456. Hermida, Alfred. 2014. *#TellEveryone: Why We Share and Why It Matters.* Toronto: Doubleday Canada.

Hermida, Alfred, Fred Fletcher, Darryl Korell, and Donna Logan. 2012. "Share, Like, Recommend: Decoding the Social Media News Consumer." *Journalism Studies* 13 (5–6): 815–824. doi:10.1080/1461670X.2012.664430.

Hirst, Martin. 2010. *News 2.0: Can Journalism Survive the Internet?* Crows Nest, N.S.W.: Allen & Unwin.

Holcomb, Jesse, Kim Gross, and Amy Mitchell. 2011. "How Mainstream Media Outlets Use Twitter: Content Analysis Shows an Evolving Relationship." The Project for Excellence in Journalism, Pew Research Center, and the George Washington University School of Media and Public Affairs. http://www.journalism.org/2011/11/14/how-mainstream-media-outlets-use-twitter/.

Holton, Avery E., Mark Coddington, Seth C. Lewis, and Homero Gil De Zúñiga. 2015. "Reciprocity and the News: The Role of Personal and Social Media Reciprocity in News Creation and Consumption." *International Journal of Communication* 9: 2526–2547.

Honeycutt, Courtenay, and Susan C. Herring. 2009. "Beyond Microblogging: Conversation and Collaboration via Twitter." In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS-47)*, 1–10. Los Alamitos, California: IEEE Computer Society. doi:10.1109/HICSS.2009.602.

Hornmoen, Harald, and Steen Steensen. 2014. "Dialogue as a Journalistic Ideal." Journalism Studies 15 (5): 543–554. doi:10.1080/1461670X.2014.894358.

Jang, S. Mo, and Josh Pasek. 2015. "Assessing the Carrying Capacity of Twitter and Online News." *Mass Communication and Society* 18 (5): 577–598. doi:10.1080/15205436.2015.1035397.

Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. 2007. "Why We Twitter: Understanding Microblogging Usage and Communities." In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis (WebKDD/SNA-KDD '07)*, 56–65. New York: ACM Press. doi:10.1145/1348549.1348556.

Ju, Alice, Sun Ho Jeong, and Hsiang Iris Chyi. 2014. "Will Social Media save Newspapers? Examining the Effectiveness of Facebook and Twitter as News Platforms." *Journalism Practice* 8 (1): 1–17. doi:10.1080/17512786.2013.794022.

Kergl, Dennis, Robert Roedler, and Sebastian Seeber. 2014. "On the Endogenesis of Twitter's Spritzer and Gardenhose Sample Streams." In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 357–364. Los Alamitos, CA: IEEE Computer Society.

doi:10.1109/ASONAM.2014.6921610.

Krishnamurthy, Balachander, Phillipa Gill, and Martin Arlitt. 2008. "A Few Chirps about Twitter." In *Proceedings of the First Workshop on Online Social Networks (WOSN '08)*, 19–24. New York: ACM Press. doi:10.1145/1397735.1397741.

Kumar, Shamanth, Fred Morstatter, and Huan Liu. 2014. *Twitter Data Analytics.* SpringerBriefs in Computer Science. New York: Springer New York. http://link.springer.com/10.1007/978-1-4614-9372-3.

Kwak, Haewoon, Changhyun Lee, Hosung Park, and Sue Moon. 2010. "What is Twitter, a Social Network or a News Media?" In *Proceedings of the 19th International Conference on World Wide Web (WWW 2010)*, 591–600. New York: ACM Press. doi:10.1145/1772690.1772751.

Lasorsa, Dominic. 2012. "Transparency and Other Journalistic Norms on Twitter: The Role of Gender." *Journalism Studies* 13 (3): 402–417. doi:10.1080/1461670X.2012.657909.

Lasorsa, Dominic, Seth C. Lewis, and Avery E. Holton. 2012. "Normalizing Twitter: Journalism Practice in an Emerging Communication Space." *Journalism Studies* 13 (1): 19–36. doi:10.1080/1461670X.2011.571825.

Lawrence, Regina G., Logan Molyneux, Mark Coddington, and Avery Holton. 2014. "Tweeting Conventions: Political Journalists' Use of Twitter to Cover the 2012 Presidential Campaign." *Journalism Studies* 15 (6): 789–806. doi:10.1080/1461670X.2013.836378.

Lewis, Seth C. 2015. "Journalism in an Era of Big Data." *Digital Journalism* 3 (3): 321–330. doi:10.1080/21670811.2014.976399.

Lewis, Seth C., and Oscar Westlund. 2015. "Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics." *Digital Journalism* 3 (3): 447–466. doi:10.1080/21670811.2014.976418.

Liu, Yabing, Chloe Kliman-Silver, and Alan Mislove. 2014. "The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 305–314. Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8043.

Lotan, Gilad, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, and danah boyd. 2011. "The Revolutions Were Tweeted: Information Flows during the 2011 Tunisian and Egyptian Revolutions." *International Journal of Communication* 5: 1375–1405.

Mabweazara, Hayes Mawindi. 2014. "Introduction: 'Digital Technologies and the Evolving African Newsroom': Towards an African Digital Journalism Epistemology." *Digital Journalism* 2 (1): 2–11. doi:10.1080/21670811.2013.850195.

Maggi, Federico, Alessandro Frossi, Stefano Zanero, Gianluca Stringhini, Brett Stone-Gross, Christopher Kruegel, and Giovanni Vigna. 2013. "Two Years of Short URLs Internet Measurement: Security Threats and Countermeasures." In *Proceedings of the 22nd International Conference on World Wide Web (WWW 2013)*, 861–872. New York: ACM Press.

Malik, Momin M., Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. "Population Bias in Geotagged Tweets." In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*, 18–27.

Palo Alto, California: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662.

Mare, Admire. 2014. "New Media Technologies and Internal Newsroom Creativity in Mozambique: The Case of @Verdade." *Digital Journalism* 2 (1): 12–28. doi:10.1080/21670811.2013.850196.

Marwick, Alice E., and danah boyd. 2011. "I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience." *New Media & Society* 13 (1): 114–133. doi:10.1177/1461444810365313.

Matias, J. Nathan, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jacklyn Friedman, and Charlie DeTar. 2015. "Reporting, Reviewing, and Responding to Harassment on Twitter: Women, Action, and the Media." Women Action Media. http://www.womenaction media.org/twitter-report/.

McChesney, Robert W. 2012. "Farewell to Journalism? Time for a Rethinking." *Journalism Studies* 13 (5-6): 682–694. doi:10.1080/1461670X.2012.679868.

McCombs, Maxwell E., and Donald L. Shaw. 1972. "The Agenda-Setting Function of Mass Media." *Public Opinion Quarterly* 36 (2): 176–185. doi:10.1086/267990.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. "Understanding the Demographics of Twitter Users." In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-11)*, 554–557. Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2816.

Mitchell, Amy, and Paul Hitlin. 2013. "Twitter Reaction to Events Often at Odds with Overall Public Opinion." Pew Research Center. http://www.pewresearch.org/2013/03/04/twitter-reaction-to-events-often-at-odds-with-overall-public-opinion/.

Mitchell, Amy, Jeffrey Gottfried, and Katerina Eva Matsa. 2015. "Millennials and Political News: Social Media—The Local TV for the Next Generation?" Pew Research Center's Journalism Project. http://www.journalism.org/2015/06/01/millennials-political-news/.

Morstatter, Fred, Jürgen Pfeffer, Huan Liu, and Kathleen M. Carley. 2013. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM-13)*, 400–408. Palo Alto, CA: AAAI Press. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6071.

Morstatter, Fred, Jürgen Pfeffer, and Huan Liu. 2014. "When is It Biased?: Assessing the Representativeness of Twitter's Streaming API." In *Companion to the Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)*, 555–556. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee. doi:10.1145/2567948.2576952.

Morstatter, Fred, Jürgen Pfeffer, Katja Mayer, and Huan Liu. 2015. "Texts, Topics, and Turkers: A Consensus Measure for Statistical Topics." In *Proceedings of 26th ACM Conference on Hypertext and Social Media (HT '15)*, 123–131. New York: ACM Press. doi:10.1145/2700171.2791028.

Nielsen, Rasmus Kleis, and Kim Christian Schrøder. 2014. "The Relative Importance of Social Media for Accessing, Finding, and Engaging with News: An Eight-Country Cross-Media Comparison." *Digital Journalism* 2 (4): 472–489. doi:10.1080/21670811.2013.872420.

Parasie, Sylvain. 2015. "Data-Driven Revelation? Epistemological Tensions in Investigative Journalism in the Age of 'Big Data'." *Digital Journalism* 3 (3): 364–380. doi:10.1080/21670811.2014.976408.

Picard, Robert G. 2014. "Twilight or New Dawn of Journalism? Evidence from the Changing News Ecosystem." *Journalism Practice* 8 (5): 488–498. doi:10.1080/17512786.2014.905338.

Poblete, Barbara, Ruth Garcia, Marcelo Mendoza, and Alejandro Jaimes. 2011. "Do All Birds Tweet the Same? Characterizing Twitter around the World." In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011)*, 1025–1030. New York: ACM Press. doi:10.1145/2063576.2063724.

Revers, Matthias. 2014. "The Twitterization of News Making: Transparency and Journalistic Professionalism." *Journal of Communication* 64 (5): 806–826. doi:10.1111/jcom.12111.

Rogers, Richard. 2013. "Foreword: Debanalising Twitter: The Transformation of an Object of Study." In *Twitter and Society*, edited by Katrin Weller, Axel Bruns, Jean Burgess, Merja Mahrt, and Cornelius Puschmann, ix–xxvi. Digital Formations. New York: Peter Lang.

Ruths, Derek, and Ju ̈rgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346 (6213): 1063–1064. doi:10.1126/science.346.6213.1063.

Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities* 2 (1). http://journalofdigitalhumanities.org/2-1/wordsalone-by-benjamin-m-schmidt/.

Skogerbø, Eli, and Arne H. Krumsvik. 2015. "Newspapers, Facebook and Twitter: Intermedial Agenda Setting in Local Election Campaigns." *Journalism Practice* 9 (3): 350–366. doi:10.1080/17512786.2014.950471.

Thomas, Kurt, Chris Grier, Dawn Song, and Vern Paxson. 2011. "Suspended Accounts in Retrospect: An Analysis of Twitter Spam." In *Proceedings of the 2011 ACM SIGCOMM Internet Measurement Conference (ICM '11)*, 243–258. New York: ACM Press. doi:10.1145/2068816.2068840.

Thomas, Kurt, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. 2013. "Trafficking Fraudulent Accounts: The Role of the Underground Market in Twitter Spam and Abuse." In *Proceedings of the 22nd USENIX Conference on Security (SEC '13)*, 195–210. Berkeley, CA, USA: USENIX Association. http://dl.acm.org/citation.cfm?id=2534766.2534784.

Thurman, Neil, and Anna Walters. 2013. "Live Blogging—Digital Journalism's Pivotal Platform? A Case Study of the Production, Consumption, and Form of Live Blogs at Guardian.co.uk." *Digital Journalism* 1 (1): 82–101. doi:10.1080/21670811.2012.714935.

Tufekci, Zeynep. 2014. "Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls." In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM-14)*, 505–514. Palo Alto, CA: AAAI Press. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062.

Vasterman, P. L. M. 2005. "Media-Hype: Self-Reinforcing News Waves, Journalistic Standards and the Construction of Social Problems." *European Journal of Communication* 20 (4): 508–530. doi:10.1177/0267323105058254.

Verweij, Peter, and Elvira van Noort. 2014. "Journalists' Twitter Networks, Public Debates and Relationships in South Africa." *Digital Journalism* 2 (1): 98–114. doi:10.1080/21670811.2013.850573.

Vis, Farida. 2013. "Twitter as a Reporting Tool for Breaking News: Journalists Tweeting the 2011 UK Riots." *Digital Journalism* 1 (1): 27–47. doi:10.1080/21670811.2012.741316.

Wang, De, Shamkant B. Navathe, Ling Liu, Danesh Irani, Acar Tamersoy, and Calton Pu. 2013. Click Traffic Analysis of Short URL Spam on Twitter (Invited Paper). In *Proceedings of the 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2013)*, 250–259. Los Alamitos, CA: IEEE Computer Society. doi:10.4108/icst.collaboratecom.2013.254084.

Williams, Shirley A., Melissa M. Terras, and Claire Warwick. 2013. "What Do People Study When They Study Twitter? Classifying Twitter Related Academic Papers." *Journal of Documentation* 69 (3): 384–410. doi:10.1108/JD-03-2012-0027.

Young, Mary Lynn, and Alfred Hermida. 2015. "From Mr. and Mrs. Outlier to Central Tendencies: Computational Journalism and Crime Reporting at the Los Angeles times." *Digital Journalism* 3 (3): 381–397. doi:10.1080/21670811.2014.976409.

Zhu, Xiaojin, R. Bryan Gibson, and T. Timothy Rogers. 2009. "Human Rademacher Complexity." In *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, 2322–2330. Curran Associates, Inc. http://papers.nips.cc/paper/3771-human-rademacher-complexity.pdf.

Momin M. Malik, (author to whom correspondence should be addressed), Institute for Software Research, School of Computer Science, Carnegie Mellon University, USA. E-mail: jpfeffer@cs.cmu.edu; Corresponding author. E-mail: momin.malik@cs.cmu.edu. ORCID http://orcid.org/0000-0002-4871-0429

Jürgen Pfeffer, Institute for Software Research, School of Computer Science, Carnegie Mellon University, USA. E-mail: jpfeffer@cs.cmu.edu. ORCID http://orcid. org/0000-0002-1677-150X.