Searching and Clustering Methodologies: Connecting Political Communication Content across Platforms

By
KEVIN DRISCOLL
and
KJERSTIN THORSON

People create, consume, and share content online in increasingly complex ways, often including multiple news, entertainment, and social media platforms. This article explores methods for tracing political media content across overlapping communication infrastructures. Using the 2011 Occupy Movement protests and 2013 consumer boycotts as cases, we illustrate methods for creating integrated datasets of political event-related social media content by (1) using fixed URLs to link posts across platforms (*URL-based* integration) and (2) using semiautomated text clustering to identify similar posts across social networking services (*thematic* integration). These approaches help to reveal biases in the way that we characterize political communication practices that may occur when we focus on a single platform in isolation.

Keywords: political communication; social media; content analysis; cluster analysis; text mining

We are witnessing the emergence of a remarkable array of methodological approaches for exploring the spread of political-event related expressions on social media. However, one limitation of many of these studies is the exclusive focus on analysis of a single social platform, such as Twitter, Facebook, or YouTube (e.g., Bruns and Burgess 2011; Lotan et al. 2011; Kim, Kim, and Yoo 2014). There is much to be learned from these studies, but we observe that the analysis of social data collected from a single platform does not accurately reflect the lived experiences of users, whose complex repertoires of content creation,

Kevin Driscoll is a postdoctoral researcher at Microsoft Research. His research concerns the histories, politics, and popular cultures of personal computer networks.

Kjerstin Thorson is an assistant professor in USC's Annenberg School for Communication and Journalism. Her research explores the effects of digital and social media on political engagement, activism, and persuasion, particularly among young adults.

DOI: 10.1177/0002716215570570

consumption, and sharing increasingly arc across social media, websites, blogs, and so on. As it becomes routine for users to load multiple tabs in their web browsers and applications on their smartphones, it is crucial to observe social media use across many different sites and services in simultaneity. Scholarly focus on analysis of a single communication platform can act to "abstract new social media out of more complex contexts" (Segerberg and Bennett 2011, 199), leading to the "fetishization" of specific platforms such as Twitter and Facebook.

Our focus here is on exploring methodologies for integrating data collected on multiple social platforms. Assembling such datasets is methodologically challenging, adding another layer of complexity to already difficult "big data" collection processes. We describe two approaches for unifying and reducing diverse datasets. The first involves integrating multiplatform datasets by tracking a specific type of media content across multiple social media sites, an approach we term *URL-based* integration. Here, a fixed URL serves as the key to unite posts collected across distinct social media platforms. We illustrate this approach by describing our study of the use of video in the discourses about the 2011 Occupy Movement protests. The second approach is *thematic* integration. We use an unsupervised text clustering procedure to identify related discourses across social networking services and group posts from distinct platforms by theme. We illustrate the thematic approach using a large corpus of posts related to consumer boycotts collected from Twitter, Facebook, Google+, and Disqus. For researchers asking questions of large-scale social data, both of these techniques offer an entry point into analyses of political communication as it occurs across social media platforms.

The Challenges of a Multiplatform Approach

Many everyday conversations that were once hidden or ephemeral are now tacitly recorded by the technical infrastructures of social media systems. This new visibility represents a significant opportunity for communication research. Scholars of political communication in particular are bringing new data collection and analysis procedures to bear on questions as diverse as how citizens talk about political events, how voters are mobilized, how contentious actions are organized, and the extent of partisan polarization in online political communication.

An array of existing methodological tools has been applied to big data collections of interest to political communication researchers. For example, Lewis, Zamith, and Hermida (2013) outlined procedures to apply content analysis to a collection of Twitter data. Others have combined content analysis with network analyses to understand the ebb and flow of Twitter conversations (Papacharissi and de Fatima Oliveira 2012; Himelboim, McCreery, and Smith 2013). One commonality across many of these studies is their focus on a single social media platform, a state of affairs that Mattoni and Treré (2014) call "the one-medium bias." The result of a literature dominated by single-platform studies is a growing knowledge base concerning political communication processes within, for

example, Twitter, but much less about the role played by any given social media service within the broader ecosystem around social movements, elections, and other political processes.

One approach to produce more coherent knowledge about digital communication concerning a political event is to collect data from a single platform (e.g., Twitter) and build a dataset outward to examine how links to other forms of media were circulated within that platform. Segerberg and Bennett (2011) examined the media links shared by activists within Twitter hashtags related to two climate change protests. In an analysis of tweets related to the Arab Spring, Aday et al. (2013) extracted links to the URL shortener bit.ly as a way to explore what kind of media were consumed by readers of the Arab Spring hashtags. Although studies like these provide extraordinary insight by beginning to build bridges across forms of digital media, they prioritize communication on the topic as it originates in a single platform. The alternative approaches we pose here each begin with a keyword-based data collection across multiple social networking services. That is, rather than conducting a keyword search of a single platform and expanding the dataset via analysis of linked media within that platform, a multiplatform study begins with a keyword search of more than one site (e.g., Facebook and Twitter, or Twitter and YouTube).

The difficulties of data collection are multiplied when we attempt to collect and integrate social data across multiple social media platforms. Each social media infrastructure and each tool used for collection carries with it a distinct set of difficulties. These include challenges concerned with (1) the availability of data (e.g., cost, privacy settings, unknowns about sample representativeness), (2) whether the site is indexed by the tools the researcher uses for collection and the proprietary nature of the search procedures undertaken by commercial collection software (e.g., Radian6; DiscoverText), (3) incompatibility of formats in which the data are returned, and (4) inconsistencies in the metadata available to the researcher.

The second set of challenges arises when we attempt to integrate posts that were created under the auspices of distinct social media infrastructures into a single dataset for comparison. Many social media services are superficially similar to one another in that users interact with a chronological stream of text, images, and videos, but these technical similarities belie the extent to which the content produced within such systems resist comparison. The researcher must identify stable points of contact among two or more social media systems. For some research questions, these points of contact are plainly accessible—say, a link to a piece of media. In this case, a *URL-based* integration may be used. This approach should be used when the research focus is on understanding how a particular piece of media content (e.g., a news story, an image) or genre of media content (videos, news stories in general) is used in posts that appeared across social platforms. URL-based analysis is made possible by the presence of a piece of content with a persistent link, such as a story on the *New York Times* website or a link to a video on YouTube. Relatively small-scale studies of this kind are not uncommon. Baym and Shah's (2011) study of the circulation of environmental advocacy clips from news satire programs or Wallsten's (2010) analysis of the "Yes We Can"

video from the 2008 Obama campaign are examples. Below, we detail a large-scale version of such studies in which we developed a database of videos related to the Occupy Movement in 2011 via searches of both YouTube and Twitter.

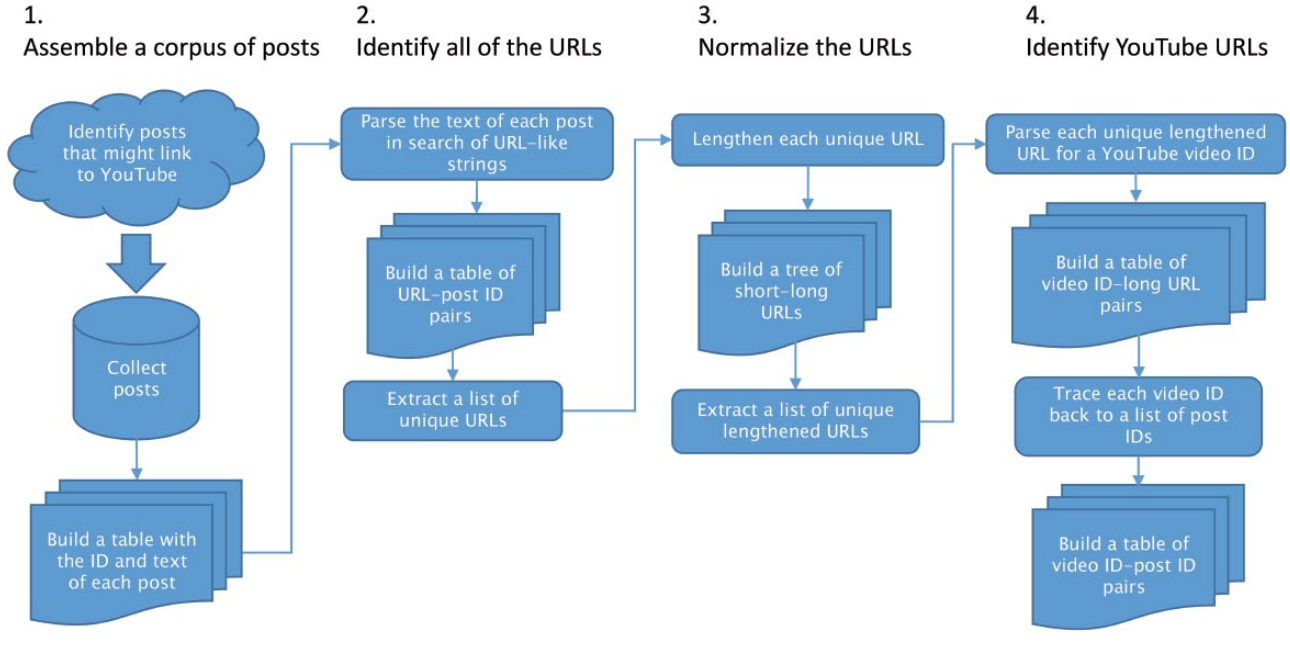
The *URL-based* approach is appropriate when the research question is focused on the circulation of a particular item or type of media content across multiple social platforms. This approach will not work for questions that require identification of related discourses on distinct media platforms. In cases like these, the researcher can attempt to integrate the data based on *thematic* similarities of posts. This is a more difficult task than URL-based integration as there is no shared "key" to unite, say, a Facebook post with a Tweet as there would be if both posts linked to the same news story. For this second approach, we explore the use of clustering algorithms to identify related texts collected across different social media sites. The two cases below illustrate use of the *URL* and *thematic* forms of data reduction.

Case 1: URL-Based Analysis of Videos Related to the Occupy Movement

In November 2011 we conducted a large-scale study of the circulation of video clips among participants in the Occupy Movement (Thorson et al. 2013). During the period of observation, the movement was decentralized structurally, ideologically, and geographically. In addition to activists camped out in cities and towns across the United States, tens of thousands of people were discussing the movement online. The instrumental goal of our data collection procedure was to create a set of videos relevant to participants in the Occupy phenomenon, broadly defined. The research design focused on two primary arenas of activity: YouTube and Twitter. Participant observations of the movement suggested the two systems were frequently used in tandem, each providing a complementary set of technical features for creating, uploading, sharing, searching, sorting, discovering, and commenting on digital videos. In some cases, activists would upload videos to YouTube and share them via Occupy-related hashtags on Twitter. There were also instances when videos were uploaded to YouTube and were not shared—YouTube was used as a storage platform as well a jumping off point for video circulation. We realized that simply looking at videos shared to Occupy via Twitter would bias our sample toward the first instance, and entirely miss the second.

We designed a data collection methodology that gathered videos from independent searches of YouTube (for videos tagged with Occupy keywords, that is, videos *about* Occupy) and Twitter (by extracting links to videos from Twitter posts that carried the same set of keywords or related hashtags, that is, videos used to *talk about or to* Occupy). We did so by creating a comprehensive list of terms, phrases, and hashtags relevant to the Occupy discourse and using identical search terms to collect tweets (by searching Twitter) and videos (by searching YouTube). We then used the unique YouTube URL structure to join the

FIGURE 1
Process Diagram for Extracting Unique YouTube URLs from a Corpus of Text Posts



collections, enabling us to trace the circulation of video clips holistically across the two platforms. URLs can be deceptively messy resources, however, and significant cleaning was required to perform a reliable URL-based integration (see Figure 1).

We searched Twitter in real time using the Gnip PowerTrack (Gnip 2014) streaming service, which provided us with access to the full "fire hose" of tweets (Driscoll and Walker 2014). Commercial services such as Gnip PowerTrack yield high-volume streams of Twitter activity, but they do not assist in the management of these data. Members of our team were responsible for building and maintaining a local system for storing, sorting, parsing, and cleaning the thousands of tweets streaming into our lab daily. During the month of November, 4,899,554 tweets matching 371 Occupy-related keywords were streamed from Gnip to our servers.

Whereas Gnip PowerTrack enabled us to collect traces of Twitter activity in real time using hundreds of keywords, it did not allow us to search for videos posted to YouTube. For this task, we used Radian6 (Salesforce 2014), a social media analytics tool aimed at marketing professionals, to conduct searches of YouTube for the time period matching our Twitter collection. For a given keyword, Radian6's historical search feature returned a list of videos in which the term appeared somewhere in the title, description, or tags. The results of these searches identified 43,378 videos on YouTube that matched our keywords.

To integrate these two datasets, we needed to identify the subset of tweets containing a link to a YouTube video. This task is more difficult than it seems because many of the links that appear on Twitter are masked by one of hundreds of URL shortening services such as bit.ly or is.gd. We used a custom Python script to "lengthen" each shortened URL and to identify URLs that originated from YouTube. The platform assigns an eleven-character ID to each video that can be easily extracted from the typical YouTube URL using a regular expression, for example, "P2wvZ1E9a34" in "http://www.youtube.com/watch?v=P2wvZ1E9a34."[1] The output of this multistage filtering process was a list of 22,768 unique YouTube video IDs found among all of the Occupy-related tweets sent in November 2011.

The next step was to merge the list of tweeted videos with the videos identified from YouTube, and to classify each item in this corpus as either Twitter-only, YouTube-only, or both Twitter and YouTube. We expected to see a large overlap between these two datasets. To our surprise, only 5,770 video IDs appeared in both the search of YouTube and the collection of video IDs extracted from Twitter. This rather small overlap of the two datasets was our first clue that the majority of YouTube videos tweeted to Occupy were not explicitly about Occupy (that is, they were not tagged with Occupy-specific keywords on YouTube), and the majority of videos about Occupy posted to YouTube had not been tweeted.

This discovery—one made apparent due to our multiplatform data collection—led to a number of substantive findings as our analysis of the dataset proceeded. For example, we found that many Twitter users treated YouTube as an archive, tweeting videos that had been posted months and even years before. These included music videos (e.g., Billy Bragg singing the International Socialist Anthem, *The Internationale*), historical footage of civil rights protests, stand-up

comedy, and news coverage of earlier political events. None of these videos were about Occupy, but they were brought into conversation with Occupy via individual tweeting practices. The relevance of these videos to Occupy only surfaced through the active construction of discourse by users of Twitter.

Conversely, the large store of videos about Occupy that had not been tweeted gave us a glimpse into the long tail of individual media creation practices. These videos were, on average, less "produced" than the tweeted videos. They were rarely shaped by their creator with the goal of circulation—they were uncut and unedited. These videos often took the form of long and uneventful cellphone footage of a protest or occupation. Analysis of non-tweeted videos helped us to understand how video could be used simply as an artifact of having witnessed an event (Gregory et al. 2005; Kahn and Kellner 2004). Thinking about YouTube as an archival medium in addition to a social network or video-sharing service gave us a different perspective on videos that had attracted few viewers. Small view count may not indicate a failure to attract attention, but may instead open a window into video creation in service of personal identity expression and maintenance (Pingree 2007). In general, the act of integrating data collected from two platforms revealed a much richer picture of the use of video in the Occupy movement than would have emerged from examining videos found on YouTube or Twitter alone.

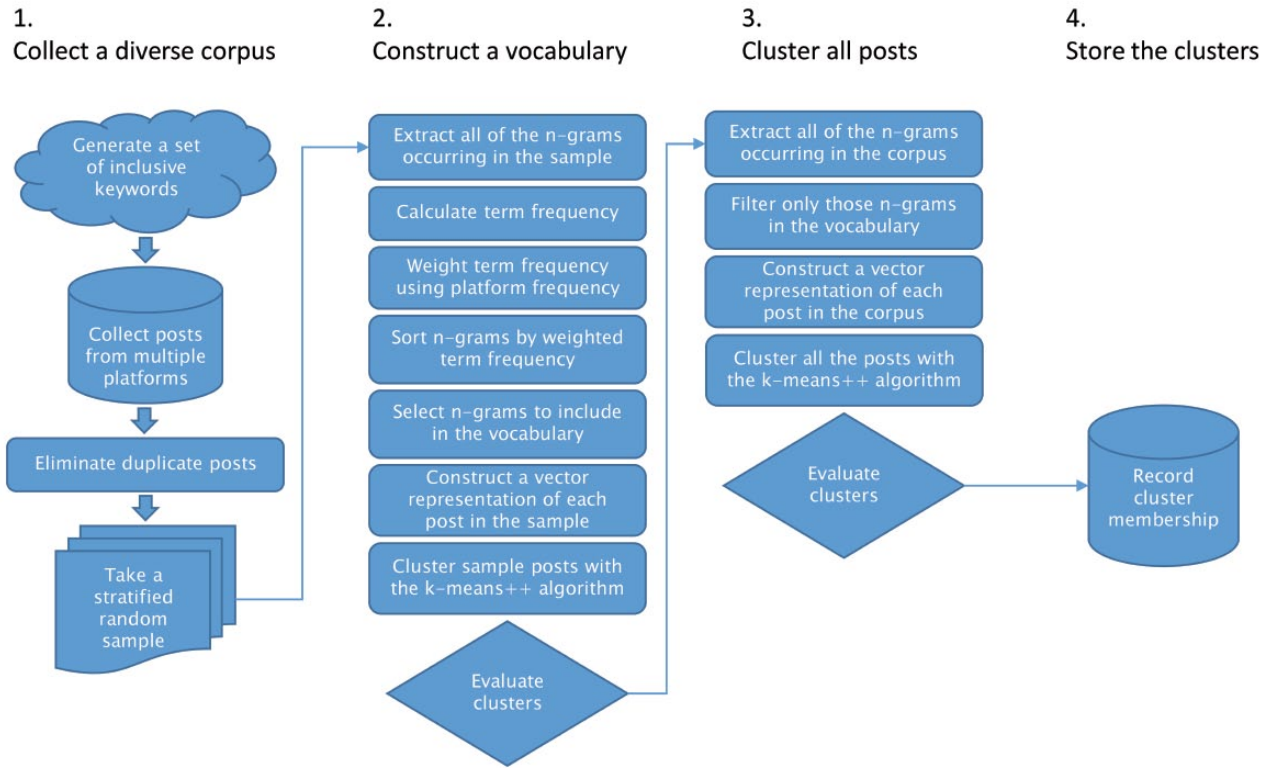
Case 2: Thematic Integration of Social Media Posts Promoting Holiday-Season Boycotts

Other research questions do not lend themselves to URL-based data integration. For example, a researcher might want to identify "hot topics" of discussion across a range of online communities (for health applications, see Chen 2012; Lu et al. 2013) or to identify subtopics of discussion within broader discourses about an election. These questions require a method to identify distinct themes across a large, diverse collection of text. We developed such a method as part of a larger project on political consumerism as a form of political participation. Specifically, we developed a multiplatform approach to cluster analysis that allowed us to identify discussion of boycotts undertaken across social media during the 2013 holiday shopping season in the United States. Unlike the Occupy case, it was not possible to develop a comprehensive set of keywords and terms in advance of collecting data—the list of boycotts that would be discussed online was still unknown. We proceeded in three steps: (1) data collection and cleaning, (2) generation of text "features" to help us discriminate among themes, and (3) clustering and theme identification (see Figure 2).

Data collection and cleaning

We began with a search for the term "boycott" across several different platforms. We used the web-based platform DiscoverText (2014) to collect posts

FIGURE 2
Process Diagram for Thematic Integration



from Disqus (a service that enables threaded discussions on websites), Facebook (public profiles only), Twitter, and Google+. DiscoverText includes features to cluster posts within but not across different platforms. We combined posts from all these sites into a single collection by transforming each post according to a standard format including the text of the post and the platform on which it was created. The resulting collection included 176,694 posts containing the word "boycott"—97,985 from Facebook, 51,275 from Twitter, 23,835 from Disqus, and 3,599 from Google+.

Next, we eliminated duplicate messages. Our definition of "duplicate" included platform-specific copies, such as Twitter "retweets," as well as exact, character-for-character replicas, as in a block of text manually pasted into the comment sections of multiple blogs. The duplicates were excluded using a custom filtering process. We were then left with 118,426 unique posts matching the keyword "boycott." Notably, messages from Twitter were the most affected by the filtering process. Just over half the tweets were identified as retweets and removed.

Generation of text features

Cluster analysis involves iteratively comparing posts to one another and assessing their similarity. There are dozens of linguistic features that might be used to compare two texts—their overall length, use of punctuation, frequency of spelling errors, and so on. To cluster the collection by theme, we first extracted a "vocabulary" of all the words and phrases occurring within any of the posts. This comprehensive vocabulary was massive and included hundreds of thousands of unique terms, most of which appeared just once and a small number of which appeared in thousands of posts. Terms at either extreme—that is, those that appear too frequently or infrequently—are not useful for clustering because they cannot help us to evaluate the similarity of two posts. Instead of using the full vocabulary, we developed a strategic vocabulary inclusive of only those terms that defined the thematic boundaries within our corpus.

To do so, we took a random sample of posts stratified proportionally to the population across the four social media platforms and used this sample to identify a set of two-word phrases that either (1) frequently occurred within a single platform or (2) frequently occurred across multiple platforms. We first identified all the two-word phrases ("2-grams") present in the sample (N = 20,668 2-grams, of which 18,636 were unique). The most frequently occurring 2-gram, "duck dynasty"—a clear reference to boycotts of the A&E television show—appeared 60 times across the sample, while 17,525 (94 percent) of the terms appeared in just one post each. We then computed a *term frequency* for each term by dividing the number of occurrences of each term by the total number of occurrences of any term and assessed the distribution of terms across each social media platform. Exploring the frequency of terms within each platform, it became immediately clear that Facebook was overrepresented in the overall frequency (see Table 1). To prevent a single platform from dominating the vocabulary, the term

TABLE 1
Top Five Frequently Occurring 2-Grams in the Sample, by Platform

OVERALL	FACEBOOK	TWITTER	DISQUS	GOOGLE+
duck dynasty	south africa	boycott top	mr simons	pick n
boycott ae	duck dynasty	duck dynasty	boycott ae	n pay
south africa	boycott ae	declares boycott	jewish people	american studies
phil robertson	phil robertson	winner declares	duck dynasty	land swaps
cracker barrel	shared boycott	science journals	nation state	phil Robertson

frequencies were then weighted by a *platform frequency* that gave a slight preference to 2-grams that appeared in two or more of the platforms.

The top (5 percent) of the terms in the weighted distribution was selected for our final vocabulary. This vocabulary included 467 unique 2-grams. Of the 467 terms, 316 (68 percent) appeared in posts found on two or more platforms (see Figure 3). We next parsed the text of our sample posts and created a vector of binary values (1 or 0), representing the presence or absence of each of the terms in the vocabulary. Using these purely numeric representations, we proceeded to assemble clusters from within the sample.

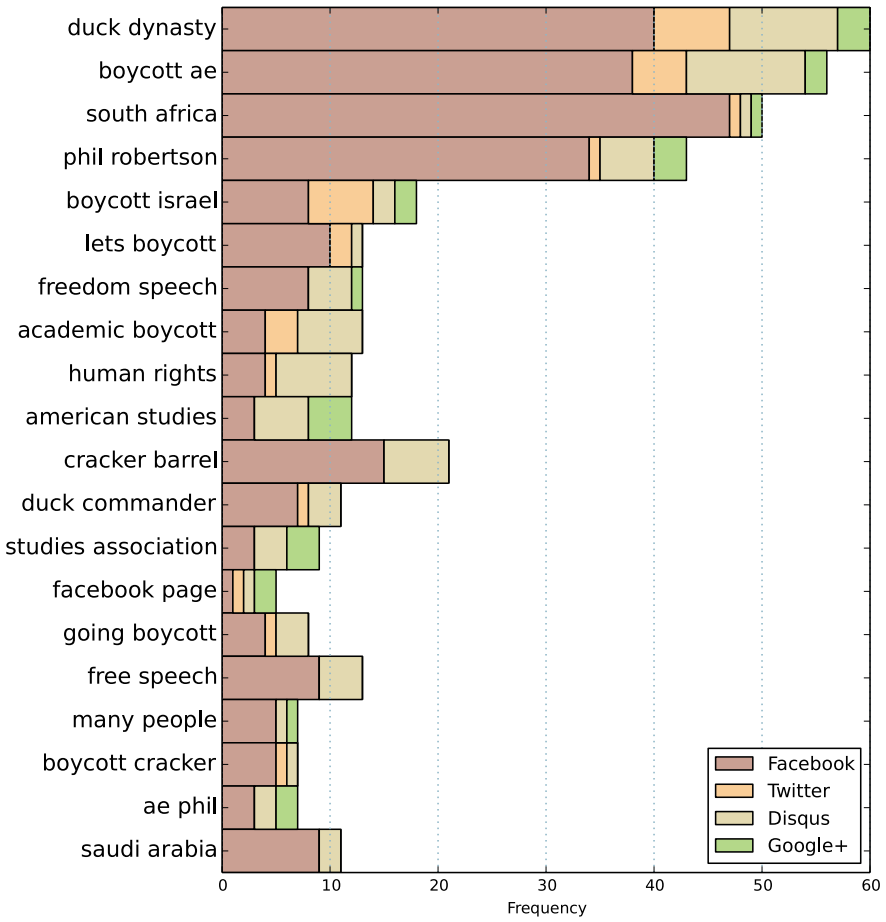
Clustering and theme identification

To cluster the sample posts, we used the K-Means algorithm as implemented in the scikit-learn machine learning library (Arthur and Vassilvitskii 2007; Jain 2010; Lloyd 1982; Pedregosa et al. 2011). First, we selected twenty posts at random to act as the "seeds" of the sample clusters to be assembled. We ran a series of experiments to cluster the data and found that twenty clusters produced an ideal solution, generating thematic consistency within clusters and thematic diversity across clusters. The second and third steps in the K-Means algorithm run in a loop, evaluating each post and assigning it to one of the existing clusters until the set of clusters reaches relative stability.

Interpreting the results of cluster analysis is a hands-on, subjective process that requires domain-specific knowledge of the content and context of the data. For a collection of loosely related social media posts, there is no single "correct" set of clusters. Instead, the purpose of clustering these posts was to identify broad themes within the data and as a route to identify cases for additional analysis. To assess the effectiveness of our custom vocabulary, we examined the twenty clusters to ensure that they represented qualitatively different themes within the broader boycott-related discourse. Satisfied with the results of the sample clustering, we then used the clustering algorithm to generate twenty clusters from all of the 118,426 posts.

We were successful in identifying conversations about distinct boycotts within the dataset. Three boycotts stood out due to their appearance in large, clearly defined clusters: the American Studies Association formal boycott of Israeli

FIGURE 3
Top Twenty Frequently Occurring 2-Grams in the Vocabulary, Weighted by Platform



academic institutions, a grassroots boycott of the 2014 Winter Olympics in Sochi, and a consumer boycott of the A&E network following the suspension of Phil Robertson from the television series *Duck Dynasty*. Robertson was suspended for making antigay statements during a magazine interview (see Carter 2013). The controversy regarding Robertson and *Duck Dynasty* encompassed several subthemes that appeared in their own clusters. For example, after Cracker Barrel withdrew its sponsorship of *Duck Dynasty*, a large number of Robertson supporters began to call for a boycott of the restaurant chain. The posts related to the Cracker Barrel boycott appeared in a distinct cluster from those clusters concerned with other aspects of the controversy.

We also observed clusters that grouped boycotts together by broader social concerns. In December 2013 gay rights organizations called for boycotts of both

the Winter Olympics and *Duck Dynasty*. These boycotts tended to co-occur in clusters because of the prevalence of 2-grams like "anti gay" and "gay rights." These sorts of overlaps were due, in part, to the dimensionality of the analysis. Increasing the size of the vocabulary creates additional dimensions for comparing posts to one another at the cost of additional computational burden. In this case, we found that increasing the dimensionality did not significantly improve the outcome of the clustering process.

Identifying clusters in this way enables researchers to develop a valuable familiarity with their data while reducing the number of individual posts to be examined. It facilitates conversations about the validity of the data collection procedures and can identify problems in the data well in advance of more labor-intensive analyses, such as a human-coded content analysis. Most importantly, the clusters enable researchers to see related political discourses as they appear within and across different platforms. Finally, the process of clustering is usefully implemented as a data reduction technique that should inform the selection of data for a content analysis or close reading of individual texts.

The boycott dataset demonstrated a few characteristics that researchers using a similar approach are likely to encounter. First, there will always be one large, "catchall" cluster that includes dissimilar posts. In some cases, it may be desirable to repeat the clustering process within this subgroup to establish another level of precision. Second, these procedures easily identify autogenerated posts or posts that were copied and pasted with small modifications. Two clusters that consistently emerged out of our data consisted of posts generated by the Facebook platform when a user shared a photo or a link in a group with the word "boycott" in its name. The corpus included more than one thousand tweets with nearly identical text save for a different set of mentions and hashtags. These posts were likely created either by a computer program or by a human following a mechanical process.

Finally, as might be expected, boycotts contained within a single nation tended to group together. For example, a small number of posts originating within India and addressing an Indian audience were clustered together. One reason for this outcome is that our clustering procedure did not account for linguistic differences. As a result, languages other than English were consistently grouped together, regardless of theme or geography. Depending on the goals of the research, the clustering procedure may be repeated within each linguistic cluster to generate more meaningful groupings.

Conclusion

At first glance, social media systems seem flush with prepackaged data just waiting to be collected. But the objects produced by these infrastructures are rarely easy to compare. Superficially similar objects—say, a tweet and a Facebook post—are produced within different social contexts and technological regimes.

Designing research around multiple platforms therefore involves defining the relationships between the various systems under observation and developing a specialized apparatus for collecting, cleaning, and combining data that reflects this ontology.

Given the multiple difficulties of collecting and analyzing data across multiple social media platforms, researchers must ask whether there are substantial benefits to engaging in the process. It is rapidly becoming clear that political events or aspects of social movements mediated by computer networks rarely remain within the arbitrary boundaries of one or another system. A user may shoot video of a protest on a mobile device, upload it to YouTube for storage, and then use Facebook and Twitter for distribution and promotion. A conversation about a boycott on Facebook may develop into an attempt to mobilize participants via Twitter. The ability to write lengthy blocks of text on Facebook or Disqus may lead political conversations in these spaces down a very different path than topically similar interactions on Twitter. And these conversations may or may not be connected. This transplatform fluidity is not only enabled but also encouraged by the architectures of websites and mobile applications that invite readers to "Tweet this!" or "Share with your friends on Instagram!" The challenge to researchers is to find ways to identify and analyze these processes systematically.

Studies designed around common URLs, such as our case on the Occupy Movement, are useful for research questions focused on the practices of sharing media artifacts among users of different systems. Researchers wishing to understand the use of video in conversations about the Occupy Movement would be misled if they simply searched YouTube for videos tagged "occupy." Key practices within the movement could only be seen when videos found on YouTube were combined with videos extracted from tweets about Occupy.

Thematic integration, as in our case on consumer boycotts, is appropriate for open-ended discourses that unfold simultaneously on multiple sites and services, regardless of whether there are explicit points of contact. The purpose of the boycott study was to identify a range of boycotts occurring simultaneously among a variety of different people with diverse political commitments, with an eye to further analysis within each case. Social media is not evenly adopted across all populations so boycotts discussed on one platform may not be discussed on another. Indeed, some boycotts were mentioned on just one system, while others were discussed in multiple contexts.

For users and scholars alike, social media systems are increasingly vital sites for political advocacy, argument, humor, and learning. Both URL and thematic approaches to multiplatform research can help to reveal biases in the way we characterize political communication practices that may occur when we focus on a single platform in isolation. Many social media systems offer similar technical features, but users and groups adopt them to address different social and political needs. Extending research across multiple systems offers a first step toward understanding the ways that users integrate different media systems into their repertoire of communication habits.

Note

1. The script used to lengthen short URLs is freely available. See Driscoll (2014).

References

Aday, Sean, Henry Farrell, Deen Freelon, Marc Lynch, John Sides, and Michael Dewar. 2013. Watching from afar: Media consumption patterns around the Arab Spring. *American Behavioral Scientist* 57 (7): 899–919.

Arthur, David, and Sergei Vassilvitskii. 2007. K-Means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–35. New York, NY: ACM.

Baym, Geoffrey, and Chirag Shah. 2011. Circulating struggle: The on-line flow of environmental advocacy clips from The Daily Show and The Colbert Report. *Information, Communication & Society* 14 (7): 1017–38.

Bruns, Axel, and Jean E. Burgess. 2011. New methodologies for researching news discussion on Twitter. Paper presented at The Future of Journalism 2011 Conference, 8–9 September, Cardiff, UK.

Carter, Bill. 19 December 2013. Fans criticize move to suspend "Duck Dynasty" star. *New York Times*.

Chen, Annie T. 2012. Exploring online support spaces: Using cluster analysis to examine breast cancer, diabetes and fibromyalgia support groups. *Patient Education and Counseling* 87 (2): 250–57.

DiscoverText. 2014. [Computer software]. Available from https://discovertext.com/.

Driscoll, Kevin. 2014. Batch URL Expander [Computer software]. Available from https://github.com/driscoll/batch-url-expander.

Driscoll, Kevin, and Shawn Walker. 2014. Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication* 8:1745–64.

Gnip. 2014. Gnip Powertrack [Computer software]. Available from http://gnip.com/products/realtime/powertrack/.

Gregory, Sam, Gillian Caldwell, Ronit Avni, and Thomas Harding. 2005. *Video for change: A guide for advocacy and activism*. Ann Arbor, MI: Pluto Press.

Himelboim, Itai, Stephen McCreery, and Marc Smith. 2013. Birds of a feather tweet together: Integrating network and content analyses to examine cross-ideology exposure on Twitter. *Journal of Computer-Mediated Communication* 18 (2): 40–60.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters, Award Winning Papers from the 19th International Conference on Pattern Recognition (ICPR) 19th International Conference in Pattern Recognition (ICPR)* 31 (8): 651–66.

Kahn, Richard, and Douglas Kellner. 2004. New media and Internet activism: From the "Battle of Seattle" to blogging. *New Media & Society* 6 (1): 87–95.

Kim, Ji won, Yonghwan Kim, and Joseph Jai-sung Yoo. 2014. The public as active agents in social movement: Facebook and Gangjeong movement. *Computers in Human Behavior* 37:144–51.

Lewis, Seth C., Rodrigo Zamith, and Alfred Hermida. 2013. Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media* 57 (1): 34–52.

Lloyd, Stuart P. 1982. Least squares quantization in PCM. *IEEE Transactions on Information Theory* 28 (2): 129–37.

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* 5:31.

Lu, Yingjie, Pengzhu Zhang, Jingfang Liu, Jia Li, and Shasha Deng. 2013. Health-related hot topic detection in online communities using text clustering. *PLoS ONE* 8 (2):e56221.

Mattoni, Alice, and Emiliano Treré. 2014. Media practices, mediation processes, and mediatization in the study of social movements. *Communication Theory* 24 (3): 252–71.

Papacharissi, Zizi, and Maria de Fatima Oliveira. 2012. Affective news and networked publics: The rhythms of news storytelling on #Egypt. *Journal of Communication* 62 (2): 266–82.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–30.

Pingree, Raymond J. 2007. How messages affect their senders: A more general model of message effects and implications for deliberation. *Communication Theory* 17:439–61.

Salesforce. 2014. Radian6. Available from: http://www.exacttarget.com/products/social-media-marketing/radian6.

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

Thorson, Kjerstin, Kevin Driscoll, Brian Ekdale, Stephanie Edgerly, Liana Gamber Thompson, Andrew Schrock, Lana Swartz, Emily K. Vraga, and Chris Wells. 2013. YouTube, Twitter, and the Occupy Movement: Connecting content to circulation practices. *Information, Communication & Society* 16 (3): 421–51.

Wallsten, Kevin. 2010. "Yes we can": How online viewership, blog discussion, campaign statements, and mainstream media coverage produced a viral video phenomenon. *Journal of Information Technology & Politics* 7 (2–3): 163–81.