COMPUTER-AIDED CONTENT ANALYSIS OF DIGITALLY ENABLED MOVEMENTS[*]

Alexander Hanna[†]

With the emergence of the Arab Spring and the Occupy movements, interest in the study of movements that use the Internet and social networking sites has grown exponentially. However, our inability to easily and cheaply analyze the large amount of content these movements produce limits our study of them. This article attempts to address this methodological lacuna by detailing procedures for collecting data from Facebook and presenting a class of computer-aided content analysis methods. I apply one of these methods in the analysis of mobilization patterns of Egypt's April 6 youth movement. I corroborate the method with in-depth interviews from movement participants. I conclude by discussing the difficulties and pitfalls of using this type of data in content analysis and in using automated methods for coding textual data in multiple languages.

Contemporary movements almost invariably incorporate online tools into their tactical repertoires. Although high-profile cases such as the 2011 Egyptian revolution and Occupy Wall Street have garnered the most attention, some of the fundamental work of movement activists—coordination, persuasion, and discussion—has been facilitated by the integration of social media and other information and communication technologies (ICTs). Accordingly, movement scholars are turning their focus to ICTs to better understand what is happening in these spaces and to incorporate them into the existing social movement literature. Social networking sites such as Facebook and Twitter provide us with the ability to observe what an activist or a would-be activist is saying and doing as an instance of contention unfolds. Their profile characteristics display relevant information about them, their utterances suggest particular courses of action, and the connections they make with others highlight the networked nature of these sites.

At the same time, the data available to researchers studying these movements have far surpassed the amount that had been previously available for analysis. The sheer scale of these data could give us unprecedented leverage in answering questions about social movement activity. While many studies of mobilization with ICTs have been studied in terms of network structure and message volume, most have neglected to focus on the content of messages at scale. These data can be better leveraged with computer-aided content analysis methods that can classify message content.

---

[*] I would like to thank Pamela Oliver, Chaeyoon Lim, and James Montgomery for guiding this project from start to finish. I would also like to thank Richard Aviles, Geoff Bakken, Trey Causey, Peter Hart-Brinson, Jonathan Latner, Matt Moehr, Dhavan Shah, Adam Slez, the online community that has formed around the collaborative Bad Hessian blog, and the coders Nana Elhariry and Mohammed Magdi. Lastly, I give a special thanks to the editor of this special issue, Neal Caren, and the two anonymous reviewers for their helpful comments. This article benefited from a session organized by the Middle East sociology working group at the 2012 American Sociological Association. My research was supported by a fieldwork grant from the International Institute at the University of Wisconsin-Madison and the NSF Graduate Research Fellowship.

[†] Alexander Hanna is a Ph.D. candidate in the Department of Sociology at the University of Wisconsin-Madison. Please direct all correspondence to ahanna@ssc.wisc.edu.

© 2013 Mobilization: An International Quarterly 18(4):367-388

THE GROWTH OF ONLINE ACTIVISM AND METHODS FOR RESEARCH

In light of recent movements, such as the Arab Spring and Occupy Wall Street, and the press that has surrounded the role of the Internet in fomenting protest, social scientists have highlighted the potential uses of the Internet for politics. Farrell (2012) discusses the different mechanisms through which the Internet may be impactful for political outcomes, while Earl et al. (2010) construct a four-category typology to suggest the ideational and mobilizing uses of ICTs. Earl and Kimport (2012) suggest a continuum by discussing the levels of "affordances" that are granted by the Internet for movement activity. "E-mobilizations" are those movements that attempt to mobilize for offline action through online means, "e-tactics" are tactics that combine online action into a more complete repertoire of action, and "e-movements" are those movements that conduct all of their action online. Dividing up movements that utilize ICTs into these different typologies disentangles the many different ways in which activists may use online tools and highlights how each type of activity contributes to particular parts of mobilization processes.

Given the immense range of the types of digitally enabled movements that have emerged since the first years of the Internet, the methodology to study them has dramatically evolved as well. Studies of early Internet protests focused on particular websites and online social communities as the unit of analysis, and adopted a type of "online ethnography" approach to understand the processes at work in those groups. Gurak (1997) looks at two protests against proposed software changes and privacy intrusions by investigating the Usenet and FTP (File Transfer Protocol) message boards. Eaton (2010) focuses on how MoveOn.org attempted to create an imagined community to impact elections and to challenge conservative legislation, while Kidd (2003) studies the tactics and strategies of Indymedia.org in constructing an alternative media. A significant amount of Internet research has focused on surveying the U.S. population on their Internet usage and their associated political participation; the Pew Research Center's Internet and American Life Project, which began in March 2000, has been one of the longest and most comprehensive research initiatives to undertake this project (Pew 2012). Given the fluidity of movement participants and the anonymous nature of online mobilizations, however, comparable survey research has not been conducted on activists involved with digitally enabled movements.

Ubiquitous use of the Internet has exponentially increased the amount of data at the disposal of social scientists. It is no longer tenable to conduct in-depth ethnographies of particular online movement communities, in which the researcher reads every message that the community produces. This has led scholars to adopt different strategies in analysis. Some scholars have sampled from several different sites within the same social movement field and attempted to perform manual content analysis on the sample (Earl and Kimport 2012; Earl and Schussman 2003). Others have focused on performing social network analysis on hyperlinked connections to understand the centrality of particular websites and actors in movement activity (Bennett, Foot, Xenos 2011; Garrido and Halavais 2003), or they analyze websites at the level of categories that are built into software functionality, such as message board topics and user categories (Caren, Jowers, and Gaby 2012; Hagemann 2002). But as new technologies emerge and online content grows, these strategies are becoming unwieldy and cost prohibitive. Hand coding what could be considered a representative sample could require hundreds of hours of work. In terms of network analyses, these studies have been mostly limited to the study of webpage interconnections, which assumes that links between webpage dyads are similar in quality and nature. Moreover, aggregating statistics about the number of users and the volume of postings in particular categories does not tell us much about the particular affordances that are granted by these communities.

Similar issues emerge when we turn to the study of digitally enabled movements that utilize social networking sites (SNSs). As relatively new technologies, there have not been many studies that focus on SNSs in movements. Work in this area thus far has focused on the

networked aspect of SNSs and how they may enable movement diffusion. Focusing on Facebook, Gaby and Caren (2012) suggest that the spread of the Occupy movement flourished through the diffusion of status messages and photos. The heightened emphasis on Twitter (the microblogging site in which individuals can type up to 140 characters) in the Arab Spring, Occupy Wall Street, and the #15M Spanish movements has spurred a flurry of research—much of it coming from outside of sociology. Gaffney (2010), looking at Iran's green movement, attempts to classify Twitter activity based on a number of metrics, including the top words used, self-reported location, and the time of account creation. Howard et al. (2011) collected Twitter data on a number of Arab Spring hashtags (short keywords that begin with the "#" or "hash" character that are used to unite conversations under a basic theme) and attempt to classify Arab Spring activity based on the volume of tweets and the occurrence of particular Arab Spring-related keywords within the tweet text. González-Bailón et al. (2011) highlight the network dynamics of information cascades within the #15M Spanish movement. In a case of other political communication, Conover et al. (2011) demonstrate the polarizing nature of retweets (when one user tweets the exact same content of another user, usually providing full attribution to the original user) and hashtag usage.

However, there are a number of difficulties with these methods to studying movement activity on SNSs. First, volume and number of members on particular sites may tell us that a certain movement has gained some amount of prominence, but it does not tell us much about the character of the activity that is occurring within these groups. Second, these studies tend to collect data based on particular hashtags or keywords. Collecting data this way omits important activity that may occur in messages that do not use those particular hashtags. This is particularly salient for measuring the interactions between users who tend to omit hashtags in conversation. Third, those studies that conduct network analysis are restricted to Twitter and, even then, to the analysis of users mentioning one another or retweeting each other.[1] Treating those networked interactions as the same kind may gloss over important qualitative differences—some retweets may denote agreement while others may be intended solely for information dissemination. Lastly, those studies that do attempt to understand the particular meanings of tweets do so only by counting the presence of a particular keyword or set of keywords, therefore ignoring context and parts of speech.

Theoretical Focus

The theoretical issue of how activists use new social media in movement mobilization is intertwined with the methodological issue of how to process and analyze data from them. New media are evolving rapidly, and activists are changing their use of them at the same time as scholars are trying to develop ways of studying how they are being used. Theoretical expectations about how activists are using media are following rather than preceding data collection. It is, therefore, necessary to use both general deductive principles of movement mobilization to guide research on how the media are being used, in addition to using inductive methods that are open to new information and surprises in studying a constantly changing phenomenon.

In this article, I am primarily focused on methods of extracting information from SNSs, including some lessons learned from mistakes or unexpected difficulties encountered in the data. As will be explained in more detail below, these data are focused on a specific event, which necessarily affects the character of the data and the questions that can be asked about it. Data were initially collected to address three types of questions. (1) *Mobilization*: What was the purpose of communication? Was it used to coordinate action on the ground, to coordinate Internet-specific action, or as an alternative source of news? (2) *Discourse or issues*: What topics were discussed? (3) *Persuasion*: Was there evidence of advocacy for or against the action?

The research was exploratory and open-ended in its initial design. It turned out that there was little useful information about patterns of discourse or persuasion, so these are not discussed

in detail in this paper. However, patterns of mobilization emerge from the data and make practical sense in light of the specifics of this particular case. I focus on five types of mobilization drawn both deductively from prior literature on the use of the Internet in digitally enabled movements, and inductively from interviews with movement participants and the social media data themselves.

1. Offline Coordination. Several authors have noted the potential for using the Internet as a tool for the coordination of offline activities (Earl et al. 2010; Earl and Kimport 2012; Farrell 2012). Therefore, messages can instruct people on what to do or where to go, referring to offline activities. This can include discussion of using various tactics and strategies, and warnings about police presence in certain areas.

2. Internet Action. The discussion of e-tactics (Earl and Kimport 2012) focuses on those activities that are solely online but contribute to a larger movement. Messages instruct people to use a particular technology, warn them about how not to use it, and coordinate efforts that take place specifically online. The author can, for instance, tell readers to not click strange links, to address each other with the @ symbol, to join one group and not another, or to change their profile picture.

3. Media and Press. Often, the goal of a protest action is to get the attention of the media or media personalities, something that has been well-documented in studies of purely offline mobilization (Gamson and Modigliani 1989; Klandermans 1992). Successful action depends upon making the protest known and using this leverage to embarrass or confront actors with authority. The interviews with movement activists suggested the mobilization of media and press, although in two different ways. One interviewee, Mona,[2] recounts how many of her messages were a concerted effort to gain the attention of popular television hosts, hoping they might announce the general strike planned for April 6. Another interviewee, Khaled, however, talks about how state media unintentionally worked towards the ends of protesters. The message could also link to a website that talks about the protest action, or the author might talk about seeing something about it on TV or in the newspapers.

4. Reporting on Events. Social media often serve as alternative media, with activists playing the role of citizen journalists. This becomes more important under autocratic regimes that expend significant resources developing television and newspaper outlets that dominate the media environment. As Diamond (2010: 20) notes, technology "enables citizens to report news [and] expose wrongdoing." Citizen journalists document encounters with security forces, document instances of police brutality and torture, and record protest events.

5. Request for Information. Many Facebook group members asked for information from other group members. These users presumably saw the group as a resource for tactical and protest information and could make explicit requests for it.

COMPUTER-AIDED CONTENT ANALYSIS: APPLICATIONS AND LIMITATIONS

Computer-aided (also called automated) content analysis seeks to unite two bodies of work (Monroe and Schrodt 2008): the more familiar content analysis of texts performed by humans (Krippendorff 2004) and the burgeoning and fast-paced literature from computer science on computational linguistics, natural language processing, and machine learning (Bird, Klein, and Loper 2009; Jurafsky and Martin 2008). While these approaches have been growing within political science (e.g., the special issue for *Political Analysis*, winter 2008) it has been applied less in sociology in general and to social movement research in particular.

Approaches to computer-aided content analysis vary in their mechanics based on the task at hand. Computer-aided content analysis often proceeds from simple counts of words that relate to particular outcomes of interest. Laver, Benoit, and Garry (2003) use this approach to measure the position of a political party on particular policies based on party manifestos. Jasperson et al. (1998) identify a set of framings on the federal budget and attempt to show their presence in

popular news media. In a case closer to the data at hand, Golder and Macy (2011) attempt to measure diurnal and seasonal mood patterns by detecting positive and negative affect through Twitter. However, these approaches often overly rely on humans to curate those words that are of interest, making them susceptible to the introduction of error by omission of important words. The researcher must know *a priori* which particular word and word combinations will express the preferred coding category with most validity. For some texts and topics, dictionaries have been developed that estimate particular categories (e.g., affect dictionaries that are included in the LIWC package used by Golder and Macy), but these dictionaries do not allow us to code for things of interest to movement scholars. Furthermore, they are usually restricted to English and other Western European languages.

There are two lines of modeling that do not rely so heavily on word counts and the manual construction of word dictionaries—language modeling and statistical modeling (Monroe and Schrodt 2008). Language modeling attempts to leverage as much information it can out of a single document; it attempts to identify parts of speech in a given document and allows us to see the *who, what, when, where,* and *how* of a message. In a wire report that reads, "Egyptian activists take Tahrir Square," language modeling would identify the actor ("Egyptian activists"), their action ("taking"), and the object of their action and its location ("Tahrir Square"). This type of modeling attempts to resolve ambiguity ("taking" in this case means "occupying" or "controlling") and provide context (Tahrir Square in Cairo, not in another city). The downside of this class of methods is that it requires a more thorough knowledge of the language with which the researcher is dealing, and it must handle different languages in a piecewise fashion. These methods also need more information to resolve ambiguity and understand context. Language modeling has been used in political science to analyze political events and conflict but has been restricted to analysis of English wire reports (Schrodt, Davis, and Weddle 1994; King and Lowe 2003; Leetaru and Schrodt 2013).

Statistical modeling is based on a "bag-of-words" approach; it does not pay attention to syntax but to likelihoods of co-occurrence of words or phrases. These approaches are usually referred to as *machine learning* and come in two variants: supervised and unsupervised. Supervised machine learning is a process by which human coders "train" the machine to infer meaning from a certain pattern of words (word profiles) by manually coding a subset of documents that is then applied to the rest of the corpus (the full body of documents). This is the approach used by Hopkins and King (2010) and which will be applied to the case study below. The drawback of this approach is that categories need to be defined *a priori*, and human coders must label a subset of messages. Unsupervised machine learning is similar but without the human component. The machine learns from statistical co-occurrence of words, grouping documents that belong together. Latent Dirichlet allocation and topic modeling (Blei, Ng, and Jordan 2003; Blei and Lafferty 2009) exemplify this approach. These methods can be applied across a set of large corpora and across multiple languages, hypothetically without having a coder precode any messages. The downside to unsupervised methods is that the researcher must search through "topics" generated to sort out the signal from the noise. Both of these machine learning approaches suffer from the fact that we get limited information from an individual document—we do not get the specifics of *who*, *what*, *when*, *where*, and *how*. Helpful analogies can be made between these methods and the methods with which sociologists have more experience. Supervised methods are closer to standard regression analysis, in which the messages that human coders have labeled are the "independent variables" and the rest of the corpus is the "dependent variable." Unsupervised methods are closer to a factor analysis.[3]

CASE STUDY: THE APRIL 6 YOUTH MOVEMENT

Three years before the 2011 Egyptian revolution, another movement highlighted the role of social media in popular Egyptian mobilization. In March 2008, two Egyptian youths, Ahmed

Maher and Esraa Abdel Fattah, started a group on Facebook to support the strike of textile workers in the industrial city of Mahalla al-Kobra.[4] Activists involved with liberal-leaning, pro-labor reform organizations, such as the *Kefaya* ("Enough") movement and Youth for Change, wanted to stand in solidarity with these striking workers and bring some kind of political action to the urban centers of Cairo and Alexandria. A Facebook group, intended to attract only a few hundred regular political activists, mushroomed to over 70,000 members and became the largest Facebook group in Egypt that dealt with any political issues at that time. The group's dramatic growth sparked a vibrant discussion of its goals, principles, and tactics. The political substance of discussion messages was broad, including support for free speech, economic improvement, and condemnation of the corrupt Mubarak government. The group decided that a general strike would be the most effective action and, thus, encouraged supporters to wear black or stay home. The online component of the movement was not its only one; the movement had a significant in-person, corporeal component with 20-30 people stationed in Cairo, Alexandria, Mahalla al-Kobra, and other smaller Egyptian cities. These activists included not only members of *Kefaya* but also other Egyptians in their twenties who were drawn in by the movement's online presence. Some of these participants promoted the protests via the Facebook page and kept updates on a blog.[5]

When the day of action came, workers in Mahalla struck en masse. The size of the Mahalla strike was undeniable, but activists and the government disputed the extent of the strike elsewhere. For instance, in Cairo and Alexandria, activists considered the strike successful, with independent newspapers reporting high truancy among public employees and few people in street markets. Although the action called for people to stay home, those involved with older movement organizations participated in protests at significant meeting points, such as Tahrir Square and the lawyers' syndicate. The government-run press, however, claimed that things were more or less operating at the *status quo*. However, Esraa Abdel Fattah had been taken into police custody but was eventually released on April 27. The group attempted to mobilize a second general strike on May 4, Mubarak's birthday, but by all accounts this attempt ended in failure. After these actions, the Egyptian government started to restrict access to Internet cafes by forcing its customers to register at the door, monitoring online activities more closely, and increasing police presence at the movement's other protests.

Since this event, the group has undergone a number of transformations, including several internal splits and contestations, but it has more or less remained a coherent political force. Its members participated widely in the 2011 Egyptian revolution, and the group is positioned to remain an important force in Egyptian politics.

Expectations

Because these data are centered on one specific event, I expect that posts that have the purpose of coordinating or mobilizing for the event should happen before and on the day of the event, and should drop off precipitously afterward. If the Facebook group is being used for e-mobilization there should be more talk of coordination before the April 6 event and then a precipitous drop off afterwards. Because movement activists often attempt to gain press for their event, I expect more mentions of media and press on the actual day of the action.

In addition, if the target event somehow draws attention to or fosters a subsequent movement or discussion, there may be more communication after the event. However, we should not expect post-event communication to be focused on mobilization, but rather on further discussion about the event that influences opinions or future actions in a more diffuse way. That is, the mix of types of communication should change after the event. The literature on mobilization predicts that communication about when and where to show up and exhortations to participate will occur before an event, but there is not much literature to draw on to suggest what will happen after an event. Prior research on news coverage of protest events (Earl et al. 2004; Ortiz et al. 2005) has

Figure 1. Screenshot of April 6 Youth Movement Page

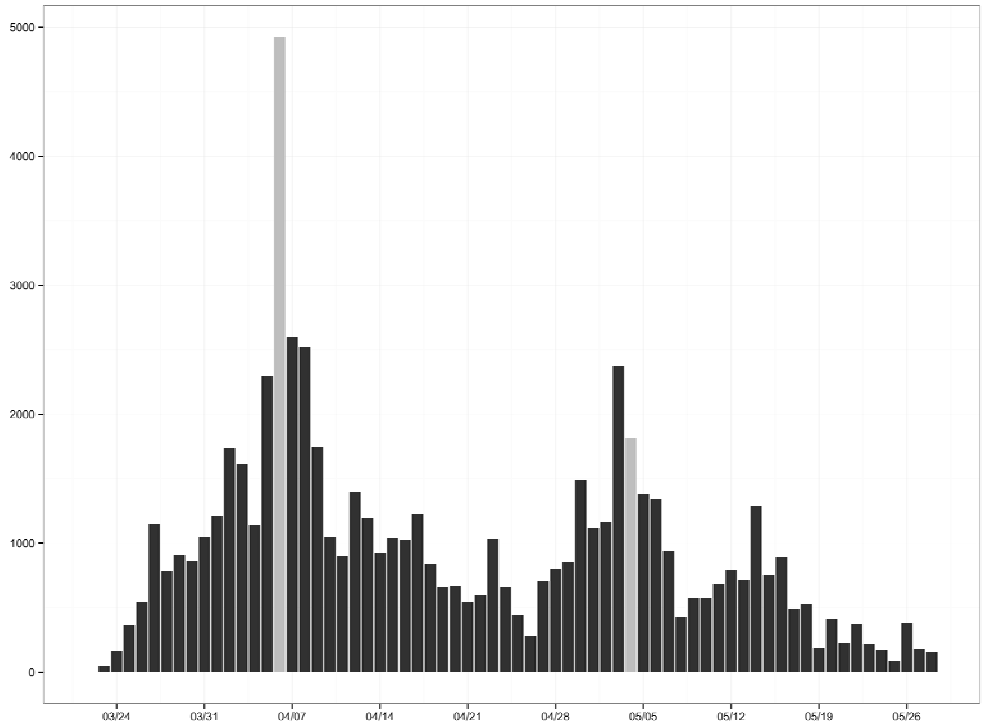
Note: The image on the left reads "The Revolution Continues" near the top and "Egypt First" near the bottom. The posting on the right show the Facebook group wall on November 4, 2011.

focused on whether events get mentioned or not, but not on how the content of coverage changes across time after the event. My analysis addresses this limitation in the literature by exploring how communication changes after the event.

Data Collection

The messages from the April 6 Facebook group were gathered in the fall and winter of 2009. These messages were drawn solely from the "wall" of the group, a landing page which users will first see when they arrive to a page. The collection period starts on March 23, 2008 (the date of the earliest message) and ends on May 29, 2008 (the date on which we observe a significant drop in group activity). To simulate public access, I wrote a program to authenticate as a Facebook user who had no friends and joined as a member of the "April 6 Strike" group and made automated requests to Facebook's database to retrieve the messages.[6] I collected all the messages from this time period, totaling 64,197 messages, with 60,078 of these messages having unique content. This discrepancy may be because of users posting repeated messages or a bug in the way that Facebook loads messages when it receives a request to load more messages. Only 3,841 distinct users posted messages, which is a fraction of the 70,000 users who had reportedly become members of the group. Figure 2 shows the fluctuation in volume of messages per day, with significant peaks on the days of action, April 6 and May 4.

Figure 2. Message Frequencies for All Users, March 23, 2008 to May 29, 2008



Note: April 6 and May 4 are highlighted in grey.

Messages on the Facebook wall were written in a mix of Arabic, English, and what is known as "Franco Arabic" (also known as "Franco," "chat Arabic," or "Arabizi"), which uses the Latin alphabet and numbers to spell out Arabic words. Numbers or punctuation marks are used in place of letters and phonemes that are not available in the Latin alphabet. For instance, the Arabic letter `ayn has no phonetic equivalent in English and can be represented by the number 3, an apostrophe, or simply the letter "a." Since words are not normalized, there is no "correct" way to spell Franco-ized words, although in this text corpus certain spellings appear much more frequently than others. Users write messages in different languages based on their linguistic ability, their technical competence for using and switching languages on computing devices, and the audience(s) that they are trying to reach. Facebook did not officially debut in Arabic until 2009, so users needed to have at least a rudimentary reading comprehension of English to access the site. *The Guardian* claims that, at the time of the Arabic-language debut, Egypt already had 900,000 Facebook users (Black and Kiss 2009).

In order to perform some of the subsequent analyses, we must be able to distinguish between messages that are written in any of the three languages. However, since Franco Arabic is an informal language, language detection tools (such as those available through the Google Translate API)[7] will not work on these messages. With no other tools available, I used a very basic language-classifying algorithm to guess the language of messages. For each message, I counted the number of "stopwords"—common words that occur at high frequency, such as conjunctions and prepositions—in each language. The message was categorized as English, Franco, or Arabic if it had the most stopwords from that language in it. This method does not take into account the use of more than one language in a posting. Nonetheless, the results seem to align with my own visual observations of these data. Using this method, 39,904 (62.2 percent) of

messages were coded as Arabic, 15,113 (23.5 percent) were coded as Franco, 3,813 (5.9 percent) were coded as English, and 5,367 (8.4 percent) could not be coded because of a lack of stopwords. Messages that were not coded could be those that contained only links to other pages and Facebook groups but no words.

METHODS

In order to standardize the text for the content analysis phase, I used a standard method to preprocess the text prior to analysis. The short social media fragments were stripped of punctuation and stopwords and converted to lowercase (for English and Franco, since Arabic is an uncased language). The list for common English and Arabic stopwords was taken from the Natural Language Toolkit (NLTK).[8] However, I also created custom lists of Franco and Arabic stopwords to account for common words that are specific to Egyptian colloquial Arabic and this dataset. Words were also normalized through stemming, the process of reducing words with common roots to their stems. For instance, "consisted," "consisting," and "consistency" would be reduced to "consist." For English, I used the well-established Porter algorithm (Porter 1980). The task of stemming in Arabic is somewhat more difficult, given that the task consists of much more than removing suffixes. One Arabic word can constitute a full sentence. I used the stemmer developed by Taghva (2005), which removes common prefixes and suffixes from longest to shortest, then checks the word against a set number of common Arabic word forms to account for infixes (affixes inserted in the middle of word stems). Both stemmers were included as part of the NLTK. A further difficulty in stemming concerns how to handle Franco words. As noted earlier, there is no particular agreement between posters on how to transliterate words from Arabic into Franco. For example, "to work" was transliterated as *3ml*, but it could also be written as *3amel*, *3eml*, or *3amal*. While this is somewhat unavoidable, there are some similar strategies that I used to minimize two words with the same meaning as being categorized differently. Based on the same algorithm as the Taghva stemmer, I stripped prefixes and suffixes in descending order of length that I assessed by looking at several thousand of the most commonly used Franco words in the dataset. I did not, however, attempt to match the words to any common pattern of word forms. Future work that deals with automated analysis of Franco words may need to rely on a dictionary-based method that can perform proximate matching on words based on common Franco variations or that can convert Franco text back into Arabic script (e.g. Darwish 2013). Similar strategies can be taken with languages that are informally transliterated into the Latin alphabet.

As noted above, the theoretical apparatus that informs the coding schema was developed both inductively from recent work on digitally enabled movements and deductively from the social media data and fieldwork undertaken in summer 2011. I conducted two interviews with April 6 activists. Khaled is a leader and founder of the organization with a deep history of political activism prior to the group's formation. He was, however, not a very active member in the development of the Facebook group itself. I identified him through other Egyptian activists. The other, Mona, is an Egyptian student studying abroad and had been participating with the group online only. I identified her by finding the most active members of the Facebook group during the period under study and sending each of them a message in English and Arabic. While I do not claim that these two interviews are generalizable to and representative of the larger experience of all members of the group, they represent two important poles of experience—one, a highly politicized member involved with the central organizing aspect of the group but removed from the operation of the online activities, and the other a generally nonpolitical member who got deeply involved from afar. This could be considered an instance of purposive sampling for maximum variation (Miles and Huberman 1994) in order to document diverse approaches to the event by different movement actors. Khaled and Mona used the group in contrasting ways, which allows me to develop a wider breadth of coding categories.

RESULTS FROM OTHER FORMS OF ANALYSIS

Before getting to the supervised machine learning method, I look at two other common ways to analyze social media data. The first method attempts to characterize the dataset by describing the distribution of user messages across the dataset. For this analysis, I generate two metrics for measuring user activity—the number of messages they posted and the calendar period that they spent posting on the Facebook wall. I then turn to the more basic content analysis method of doing word counts across time.

User Distribution of Content

As is common with Internet data, the pattern of message posting follows a power law distribution. The top panel of figure 3 plots the cumulative proportion of messages across the cumulative proportion of authors, sorted from the least active authors to the most active ones. Had there been perfect equality of message posting, the curve would match the diagonal line. This follows other studies of Internet content generation that have looked at how individuals contribute to particular message boards (Hagemann 2002), the distribution of linking to blogs (Hindman 2008), and how network connections are distributed through a web infrastructure (Barábasi and Albert 1999). Similar to these studies, only a small number of users in the group were very active during the collection period. Most users (2,392 users or 62.2 percent) posted only once. The bottom panel of figure 3 shows the number of authors that were active for more than one day, calculated by the days between their first and last posting on the group's wall during the collection period.[9] Again, the distribution of the calendar time spent posting on the Facebook wall follows a power law distribution, with only a minority posting frequently.

While these are useful metrics, they cannot tell us much about what particular items are being discussed in digitally enabled movements. The uneven distribution of message posting is not a new result in Internet studies. Moreover, it may be more substantively important to look into what those particular high frequency posters are attempting to accomplish through their activity. For that, we must look at the actual content of the messages.

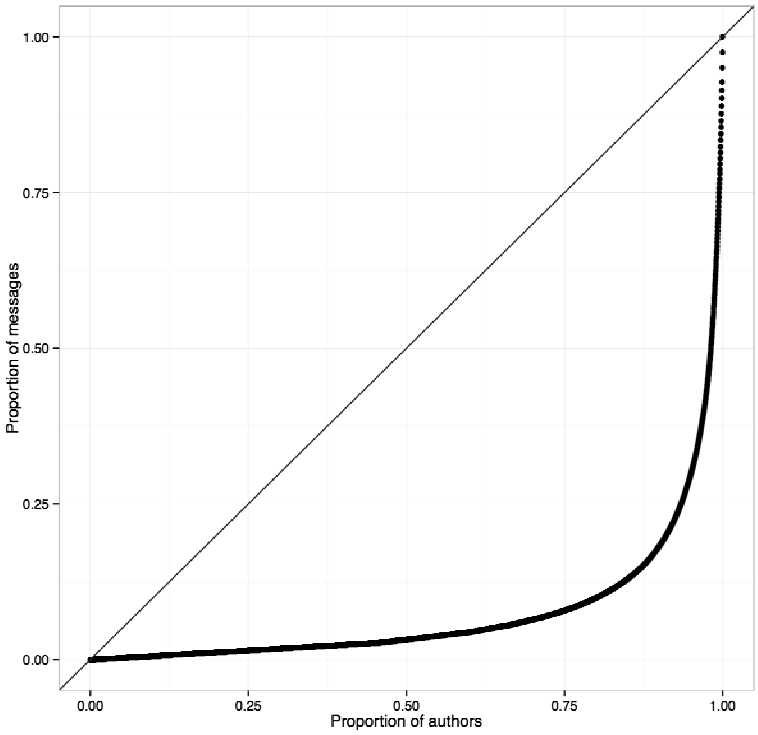
Word Counts

As mentioned above, a common strategy for analyzing content is to use word counts as a metric for estimating the outcome of interest. I used this method to generate measures for two categories of words—*offline coordination* and *media*—as well as for the frequency of words associated with *strike* to use as a comparison. I associated particular words in Arabic, Franco, and English with each code and searched for them in the stemmed text. For instance, in the *media* category, I used the word "press" and its Arabic and Franco equivalent, as well as the names of several press organizations, such as the BBC.

Figure 4 plots the word count estimates of each of these codes across the collection period. The *strike* graph is somewhat sporadic, reporting more mentions early on but dropping off after the April 6. There is some resurgence before May 4 but there is not a strong showing during the event. This may mean that talking about the *strike* has something to do with coordination. However, given the ambiguity of the word, it is not possible to disentangle the code's use without setting more specific parameters. *Offline coordination* does not follow the expected pattern, although there is a small showing on April 6. For *media*, as expected, there are peaks on April 6 and May 4, with another peak on April 27, which was the day that Esraa Abdel Fattah was released from prison.

While a few clear patterns emerge, there are those that do not show up at all. One criticism of this approach would be that these results are highly sensitive to the terms one uses to define a particular type of activity. Admittedly, measuring the frequency of references to media is some-

Figure 3. User Concentration of Messages and Duration of Activity on the Message Board



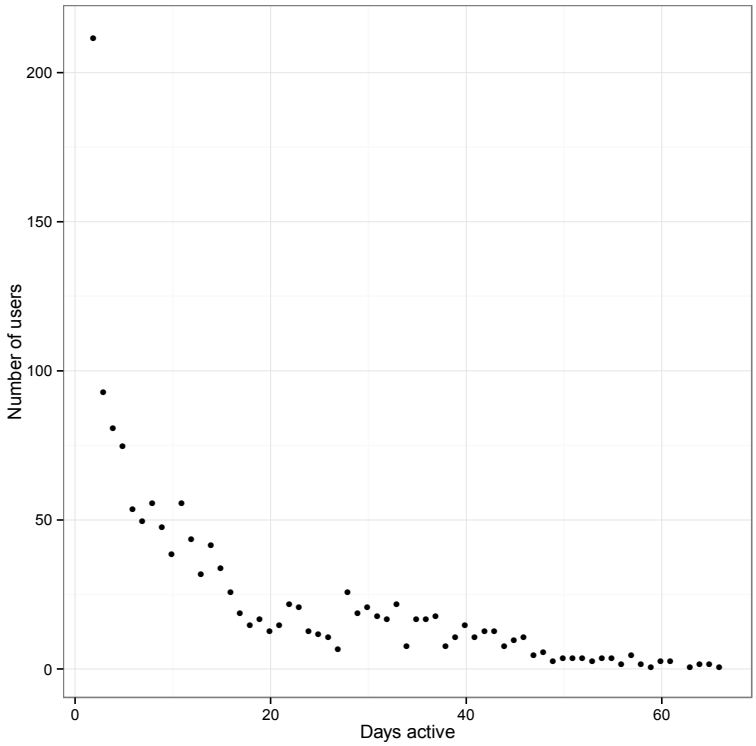
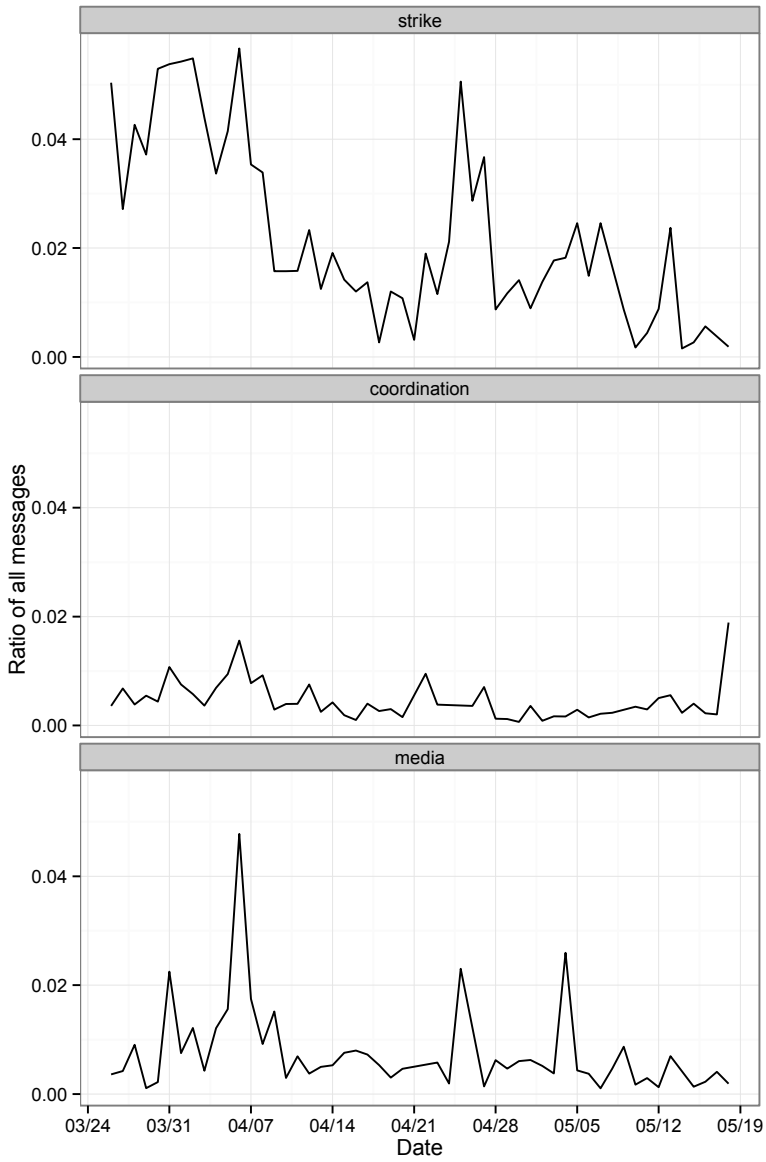


Figure 4. Ratio of Selected Words as a Proportion of All Messages across Time



what simpler, given that words associated with the media can be proper pronouns, like names of media organizations, which are generally not commonly used words and are standardized across all languages. For more complex ideas such as *offline coordination*, the words may not be so well known. Furthermore, the language used in these particular situations may have emerged as a product of the group's formation itself. For instance, a common practice on Twitter is for communities of individuals following similar events to agree upon a common hashtag at an early stage of the event or prior to it (e.g. "#asa13", the hashtag for the 2013 American Sociological Association annual meeting). The researcher must attempt to get a thorough sense of the content before attempting a word count. Another option not explored here is to use Boolean operators (e.g., AND, OR, AND NOT) to find word co-occurrence within each docu-

ment. While a useful and important first step of any content analysis, word counts do not seem adequate to detect more sophisticated movement-centric processes embedded in social media content. For this, I now move to the statistical learning method.

SUPERVISED MACHINE LEARNING METHOD

As defined above, supervised machine learning is a process by which human coders "train" the machine to infer meaning from a certain pattern of words (word profiles) by manually coding a subset of documents that is then applied to the rest of the corpus (the full body of documents). I used a nonparametric supervised machine learning method outlined by Hopkins and King (2010). This method is known as a *classifier* since the end goal is to classify messages into a set of discrete categories. As a *supervised* machine learning method, it requires creating a "training set" by hand coding a subset of the messages into mutually exclusive categories. It then uses the training set as input to assess the categorization of the rest of the messages, the "test set." In contrast with other supervised machine learning methods, this method does not attempt to categorize every individual message and aggregate them to construct a proportion of the prevalence of a code in the population. Instead, it attempts to classify the proportion of messages that are associated with a particular code in the corpus. Hopkins and King take this approach in order to eliminate possible bias introduced as a result of misclassification, mismatches between the hand-coded and the population classifications, and the particular sampling strategy for defining the training set. The method does not require the training set to be a random subsample of the population. As long as the pattern of words within a particular message is associated with a particular meaning, the condition is met.[10]

The method is implemented as an R package called ReadMe (Hopkins and King 2011). The R package was modified slightly such that it would be able to use a mix of Arabic and English words as inputs. The package generates standard errors using a standard bootstrapping approach.

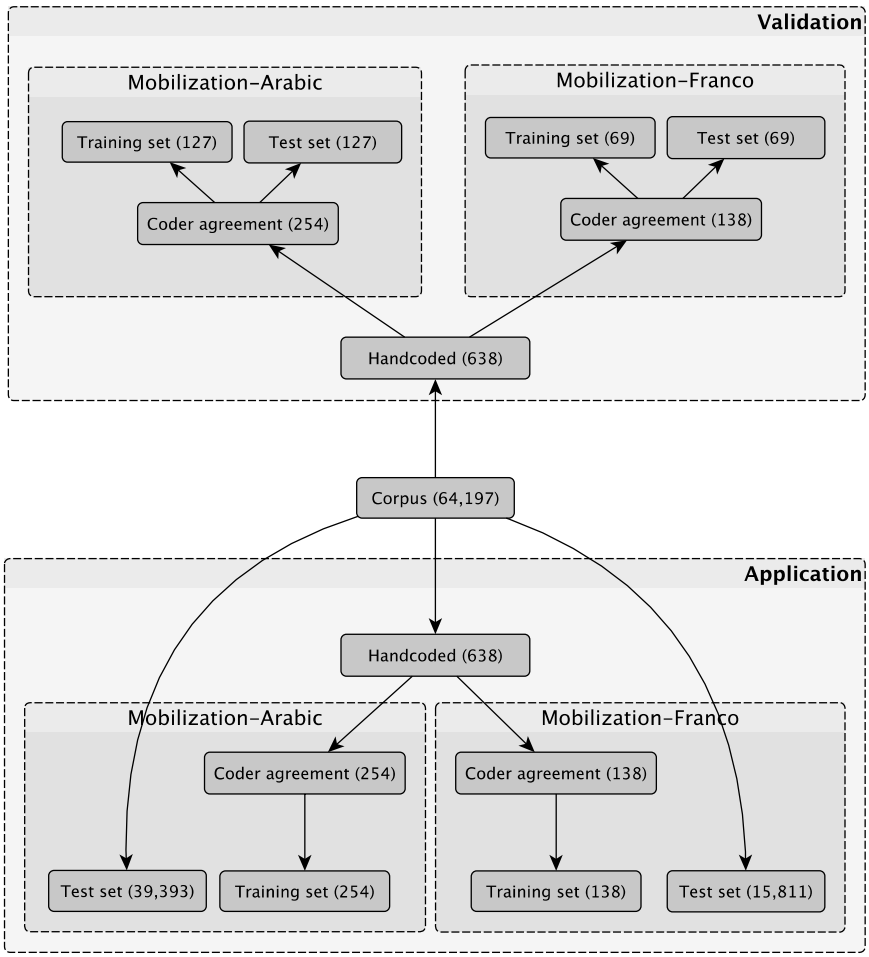
Coding

Two native speakers of Egyptian colloquial Arabic who were also fluent in English performed the coding over several months in fall 2011 and winter 2012. They coded messages for the five types of mobilization outlined earlier*: offline coordination, Internet action*, *media and press*, *reporting on events*, and *requests for information*. Coders could pick one or none of these five codes. While not theoretically mutually exclusive, the coder was restricted to picking only one of these items. In cases that were not actually mutually exclusive, I instructed coders to assign mobilization codes according to a hierarchy. *Offline coordination* and *Internet action* took priority, given that these are the most common forms of mobilization cited in the literature on ICTs and movements (e.g., Earl and Kimport's "e-mobilizations" corresponds to *offline mobilization* and their "e-tactics" correspond to *Internet action*). Between the two codes, if both were present, *offline coordination* was used. For example, if the author reports on the media and then invites readers to a specific location, the post would be coded as *offline coordination*.

I employed a cluster sampling method to select messages for coding. I randomly picked 100 messages and included those messages and the nine messages that followed each of them in the sample. The cluster technique helped the coder to see the messages in the context of surrounding messages. This resulted in a completed training set of 638 messages for both coders. Krippendorff's alpha, a measure of intercoder reliability, equaled 0.61, which does not exceed the minimum of 0.667 that Krippendorff (2004: 241) suggests but should suffice, given that alpha is comparatively strict measure. For the construction of the training set, I only used those messages on which the coders agreed on coding, which resulted in 428 messages.

The supervised machine learning method attempts to classify the test set of messages based on the "bag of words" associated with a particular code in the training set. Because of this, attempting to code each message with the same estimator across all languages produced more error than if there were estimators for each language taken individually. To make this clearer, words that have identical meaning could be taken to be different words by the classifier. Therefore, I chose to focus only on the two most prevalent languages, Arabic and Franco Arabic. I created a classifier for each language. Classifiers were trained only on messages in that particular language, and applied only to the corpus of text in that language. Therefore, in the final analysis I have two classifiers: mobilization-Arabic and mobilization-Franco. The disadvantage of doing this is that there is less information that we can use to classify the test set. For mobilization-Arabic, the final training set is a total of 254 messages, and for mobilization-Franco it is 138 messages. While there is no hard and fast rule for the necessary size of a training set, the machine learning algorithm requires sufficient information to assess the variety of word profiles in the rest of the corpus. Hopkins and King (2010) suggest starting with 200 documents and incrementally adding more if uncertainty is too great. The diagram in figure 5 shows the data sources for each procedure—validation and the final application—and shows how many messages were used in their respective training and test sets.

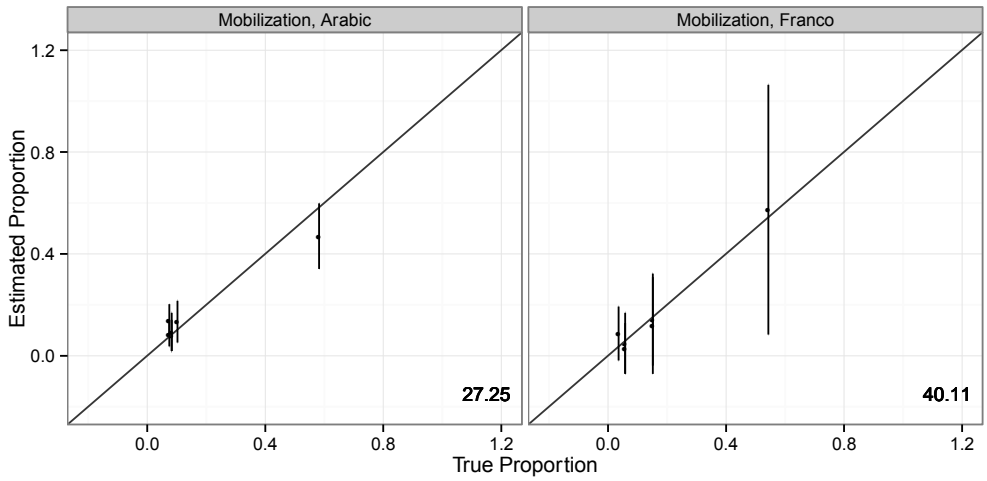
Figure 5. Diagram of the Source of Data for Validation and Application and for Each Estimator



Validation

Assessing internal validity of machine learning algorithms entails some sort of cross-validation. Cross-validation is the process by which small subsets of the data for which the true "outcome" is known are removed, and then the model is fit to see how well it predicts the out-of-sample data. This can also be used to tune parameters for model selection by choosing the model that minimizes cross-validation error. For my own purposes, I use cross-validation to assess the fit of the coded training set. Following Hopkins and King (2010), I randomly divided the training set into two parts, one serving as a training set and the other a test set. I then plotted the estimated proportions for each code generated by the machine learning method against the "known" or true proportions in the training set and calculated bootstrapped standard errors. Additionally, I calculated the mean absolute proportion error (MAPE) for each estimator. The plots and the MAPE indicate that the estimator for the mobilization-Arabic entails the least error. The confidence intervals for all parameter estimates overlap with the true proportion values.

Figure 6. Estimated Proportions Plotted Against True Proportions in the Training Set



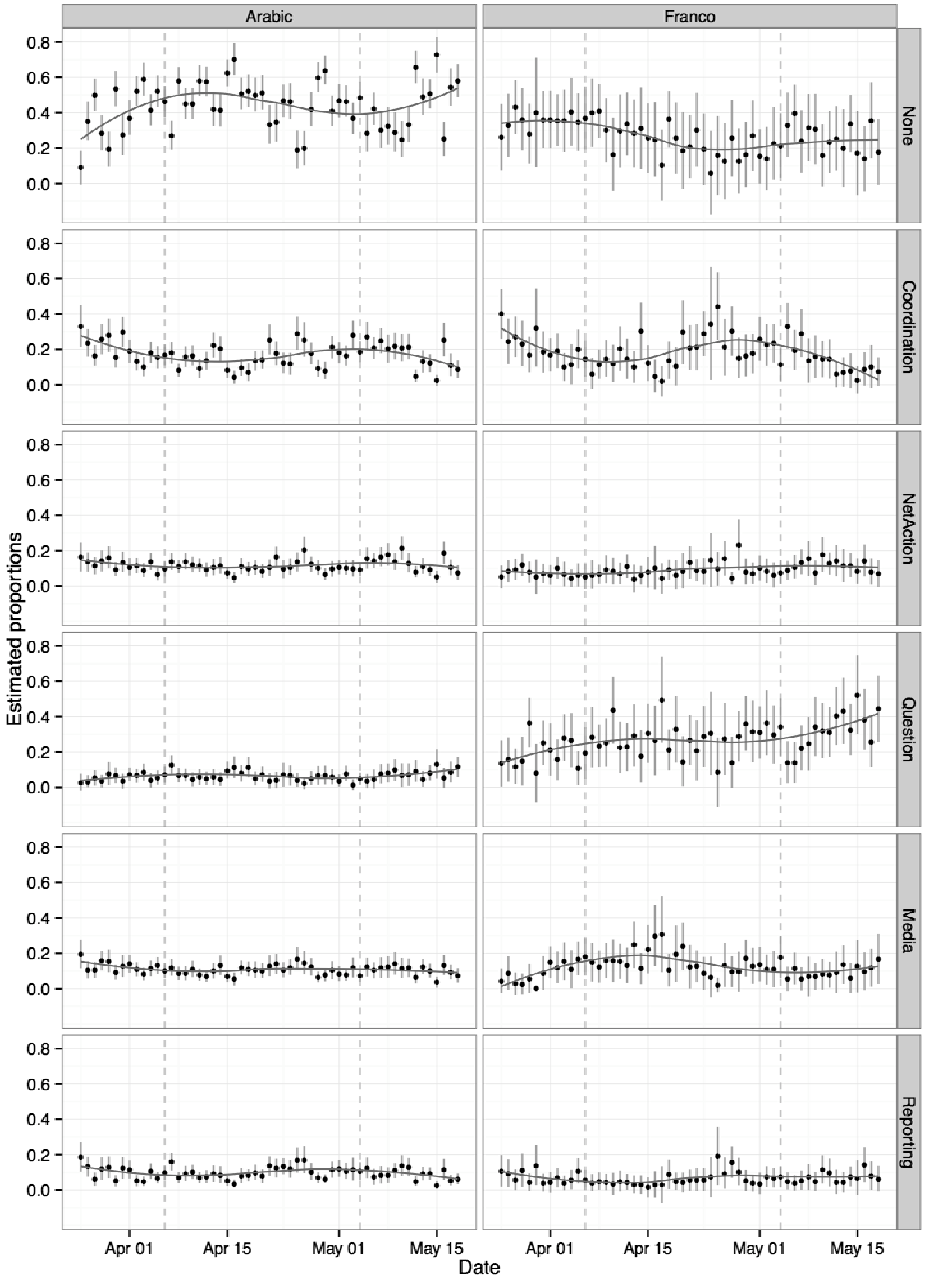
Notes: Points closer to the 45-degree line are more accurate. Error bars represent 95% confidence intervals generated by bootstrapping the standard error. Mean absolute proportion error (MAPE) reported in the bottom right corner.

Applying Across the Corpus

I applied the machine learning method to each day of the corpus, using the full training set and the messages on that day as the test set. Figure 7 displays results for the mobilization estimators in both Arabic and Franco. The most noticeable result is the high proportion of messages that have been coded as *none* for both languages. This suggests that most messages did not concern mobilization. However, some trends do emerge when paying attention to the smoothed average. With Arabic messages, fewer messages are coded as *none* early on, but that rises after April 4. Closer to May 4, the messages take another dip, but then rise after that date. However, the same patterns cannot be found with Franco messages—taking account of the confidence intervals, the messages coded as *none* are mostly constant, with a small decline before May 4.

Messages coded as *offline coordination* appear to have an inverse relationship to those coded as *none*. Both languages have higher coordination earlier on, but then dip slightly after the April 6 event. Before the May 4 event, we see a rise once again, and then a subsequent decline after May 4. This would tend to match our expectations about coordination activity—

Figure 7. Mobilization, Arabic and Franco Estimators



Note: Error bars represent 95 percent confidence intervals; curve is smoothed over time average; and vertical dotted lines are days of mobilization, April 6 and May 4.

individuals are more likely to use the Facebook group to coordinate offline activity before the actual days of action. This point should not be overstated, since the changes are limited in each direction.

Other results do not seem to reflect our expectations for a nascent mobilization. We should expect *Internet action* to match *offline coordination,* or perhaps complement it, such that in periods in which one code is low the other is high. However, for both languages this code

remains low and does not shift. Similarly, *reporting* and *media and press* ought to see spikes for time periods surrounding the days of action, but they are mostly flat. The only significant increases seem to be for the *media and press* and *reporting* codes in Franco near mid-April. These peaks would roughly correspond to the time preceding Esraa Abdel Fattah's release from prison. But it is not clear why the frequency of these codes would increase for those events and not for the days of mobilization.

Lastly, the *request for information* code (labeled "Question" in the figure) reflects divergent trends between the two languages. Estimates for this code are nearly flat for Arabic throughout the collection period, but sees a steady increase for Franco. One hypothesis for this finding has to do with technical limitations and their correlation with location. Mona, who had been studying outside of Egypt in an English-speaking country, primarily used Franco. Thus, it could be that the rate at which those outside of the country were requesting more information from others in the group increased as the group grew and reached more people. More study is needed to understand under what conditions individuals use particular languages in digitally enabled movements.

DISCUSSION

This article proposes a computer-aided content analysis program as an alternative to other modes of analysis for digitally enabled movements. In order for the results to be applicable and useful to movement studies, the researcher should proceed with a similar set of steps to assess reliability and internal and external validity. There are also a handful of considerations that need to be made about the coding process, these data, and this particular classifier.

A number of pitfalls that may result from performing this type of analysis are necessary to highlight. First, as with any content analysis project, the researcher should make theoretically informed coding choices before proceeding with the analysis. The class of methods described here requires a training set, a subset of the total messages, to be labeled by human coders. This avoids the potential error that is introduced by using only unsupervised machine learning approaches, like topic modeling, which rely on statistical co-occurrence of words or phrases to classify messages. The advantage of using human coders is that categories are defined by how humans make sense of them. I attempted to account for this tradeoff by using both deductive and inductive methods for forming categories—that is, by corroborating existing movement theory with activist interviews and by communicating frequently with the coders.

Another issue with the coding is the possibility of coder misclassification. I went through two rounds of coding—one with a different format of the coding interface, and a second round that yielded better reliability but a smaller training set. The use of multiple coders and multiple rounds would have been the ideal analysis environment, but this kind of setup is costly, defeating one of the initial impetuses for computer-aided content analysis.

Considering these data, two points are important to highlight. First, as mentioned earlier, these data were collected nearly two years after the event, which makes the data susceptible to "data decay"—users may retroactively remove their messages, or Facebook may suspend user accounts or remove messages. A recent study on the 2011 Egyptian revolution demonstrates the fragility of Twitter, YouTube, and online content, finding that ten percent of Egypt's social media content was no longer accessible one year after the revolution (SalahElDeen 2012). SalahElDeen attributes this to content owners removing their videos, users being banned from the site, or the website going offline. In an authoritarian state, users may remove messages if they feel threatened by state agents. Both Mona and Khaled reported being intimidated and threatened by police, either in person or over the Internet; this has also been the experience of several other Egyptian activists who I have interviewed. When studying movements, users may be afraid of identification and retribution by state actors and countermovements and remove their postings after a period of mobilization.[11] Second, Mona alerted me to the existence of a

message board that was unconnected to the group's Facebook wall. She described this message board as a place where people discussed contentious issues and tactics related to the strike in more detail. In this regard, we could think of the message board as a more deliberative space, compared to the Facebook wall that has much more to do with broadcasting intent and attempting to mobilize consensus. Given this sort of division, I think it is still acceptable to make claims about mobilization from the Facebook wall, since users in both of the spaces expect the same kind of audiences to read, write, and share. Research surrounding issues and persuasion may have been possible with the message board data.

The size and quality of the training set are also a possible issue. One reason that the method as applied here came up short is the sheer size of the dataset. There is no hard number of how many documents should be included in the training set. The only requirement is that it contains enough of the word combinations that exist in the rest of the corpus that it will be able to classify a document with a particular word profile. The machine learning method might work better with larger training sets. Dealing with a multilingual corpus requires that the researcher code enough documents in each language to serve as an acceptable training set.

Computer-aided content analysis methods also have not given sufficient attention to how to address multiple languages in a single corpus. There is no clear procedure for how to treat multiple languages within one corpus, not to mention a single message. For this analysis, I separated out messages based on language and generated a different estimate for each. However, usage of a particular language may indicate some other latent variable associated with factors such as perceived audience and technical ability. A Bahraini activist has told me over Twitter that she tweets in English to avoid detection by state agents. This indicates that usage of one particular language over another may not be a trivial matter, but may actually indicate fundamentally different movement behaviors. With the Arab Spring and the proliferation of data that it has produced, analysts of textual data should strongly consider how to incorporate multiple languages into their analyses.

A last issue concerns this particular content analysis method itself. This method is well suited for outcomes to be detected at the corpus level, or at the corpus level separated by time. However, there are both theoretical and methodological reasons to prefer a document-level classification method to the one used in this article. Theoretically, movement scholars are often more interested in classifying particular messages for the purpose of associating a type of behavior with particular users. Much social movement literature deals with understanding the role of leadership and the actions of a few actors who are considered critical. Attempting to gain that level of fine-grained detail may not be feasible here. Methodologically, some of the standard assessments used in the machine learning literature can be used here, such as *k*-fold cross validation, in which the corpus is divided into *k* "folds" and a model is selected based on particular features in the text. We can also apply the more familiar metrics of precision and recall, which measure relevancy and accuracy.[12]

CONCLUSION

Since the Egyptian revolution in 2011, social media has become an important object in the study of contentious politics and has gained a prominent place in the research agendas of a diverse array of scholars, including sociologists, political scientists, folklorists, philosophers, and even computer scientists.. As social movement scholars, we need to be methodologically prepared to address the large influx of data that has been associated with these movements, as part of a larger project that Lazer et al. (2010) have dubbed "computational social science." The recent rise of "big data"—data collected through social networking sites, mobile phone usage, and even clicking patterns—allows researchers to situate people in their larger networks and to analyze the content of their exchanges.

In this article, I attempt to go beyond simple analyses of digitally mediated movements by adopting a computer-aided content analytic approach to interrogate more specifically the content of movement activists' messages. I show how it is possible to code an entire corpus instead of merely sampling from it, and how results derived from simple word counts and user-level aggregations can tell an incomplete story. I worked through the process of assessing intercoder reliability and cross-validated the analysis method.

Although analysis of these data did not present any useful information surrounding issue discourses and persuasion, they did contain significant insights into how individuals mobilized in the April 6 case. I presented an analysis of how mobilization changes across time. First, most of the messages did not concern mobilization but the frequency of messages did seem to rise surrounding the days of mobilization, April 6 and May 4. Second, as expected, participants used the Facebook group more for offline coordination before the days of mobilization, presumably in an attempt to mobilize other participants. Thus, in one sense, the e-mobilization function of the Facebook group is supported by the analysis. However, other types of mobilization—such as Internet action, reporting, and media and press coverage—did not change over time or parallel the offline coordination efforts. This suggests that these efforts do not fit neatly into what we consider e-mobilization efforts. Lastly, the trend for requesting information increases for the Franco language but not for Arabic. This may be the case because those outside of the country requested information more than those inside of the country, if we assume those individuals used Franco more than those inside of it.

The data collection and content analysis process illustrates the utility of approaching large harvested data sets with computer-aided content analysis and machine learning methods. In practice, there are also a number of pitfalls that can impede the application of this type of analysis. Larger, more reliable training sets can reduce ambiguity. In the case of a multilingual corpus, there must be enough messages in each language such that the categories are adequately represented. Data decay in online text is a serious concern and may be systematically related to the vulnerability of particular movement activists or post-mobilization repression. There may also be data that the researcher does not know about and cannot retrieve after a certain time period due to technological changes.

In this era of "computational social science," social media data offer great possibilities for the study of social movements, but these data have been vastly underutilized. More and more, social movements are becoming digitally enabled at some level, giving us unprecedented access to movement workings and processes. Data from social media such as Facebook and Twitter may allow researchers to avoid the problems associated with newspaper and recall bias. Instead of only getting data from mobilization events that the press considers newsworthy, we can receive reports from activists in real time. Similarly, instead of doing retrospective interviews with movement activists, we can often observe their self-reported activity from real-time data. Obviously, these data present new biases—only activists with enough technical know-how and economic capital will be frequent users of social media for movement purposes. But these biases are known and possibly systematic enough to be addressed with methodological techniques.

Lastly, this article attempts to marry two types of empirical investigation and methodology—textual analysis and in-depth interviews—in order to better understand the intricacies and processes of mobilization in a specific case. For instance, I used qualitative data to generate coding categories. One interviewee also identified missing message board data as a potential source of bias in my analysis. Computational methods can and should be informed by other types of analysis, especially qualitative analysis. As these methods become more commonplace, we need to develop systematic strategies and guidelines for integrating qualitative methods and large-scale quantitative analysis.

NOTES

[1] In the earliest versions of Facebook, which were restricted to college campuses, collecting full network data had been possible by doing a simple breadth-first search on user profiles. However, with the platform's growth and concerns about privacy, this is no longer possible with public access.
[2] All names are pseudonyms.
[3] I thank Trey Causey for suggesting these analogies.
[4] This account relies on a number of journalistic and first-person accounts—including Carr (2012), Faris (2008, 2009), Isherwood (2008), Reese (2009), Shapiro (2009), and Wolman (2008)—personal communication with David Faris, and data from my own interviews with April 6 activists.
[5] http://7arkt6april.blogspot.com/ [Arabic].
[6] Researchers can currently rely on the various Application Programming Interfaces (API) offered by the site, which are detailed at https://developers.facebook.com/docs/reference/api/ and are subject to change.
[7] Learn more about Google Translate API at https://developers.google.com/translate/.
[8] Available at www.nltk.org/.
[9] A more detailed visual of when each user entered and exited the wall exceeds the space limitations here but can be found at http://alex-hanna.com/research/april6/.
[10] The complete formalization of the method can be found in Hopkins and King (2010: 235-38).
[11] A burgeoning literature has also begun to emerge on the ethics of online research. Ambiguity between "public" and "private," "published" and "unpublished," and "anonymous" and "identified" have made this research potentially ethically perilous. See Bos et al. (2009) for further this discussion on online research ethics.
[12] Precision can be defined as the fraction of documents correctly classified from the set of all the documents classified as that class, while recall is the fraction of documents correctly classified from the set of all documents (Manning et al. 2009). For instance, a corpus may have 20 documents that should be classified as offline coordination. Using a machine learning method, 10 documents are classified as offline coordination, but only 5 of them should actually take that code. Therefore, the precision here is 5/10 = 0.5, while the recall is 5/20 = 0.25.

REFERENCES

Barábasi, Albert-László, and Réka Albert. 1999. "Emergence of Scaling in Random Networks." *Science* 289(5439): 509-12.

Bennett, W. Lance, Kirsten Foot, and Michael Xenos. 2011. "Narratives and Network Organization: A Comparison of Fair Trade Systems in Two Nations." *Journal of Communication* 61(2): 219-45.

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly Media.

Black, Ian, and Jemima Kiss. 2009. "Facebook Launches Arabic Version." *The Guardian,* March 10. Retrieved October 28, 2013 (www.guardian.co.uk/media/2009/mar/10/facebook-launches-arabic-version).

Blei, David M., and John D. Lafferty. 2009. "Topic Models." Retrieved October 28, 2013 (www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf).

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

Bos, Nathan, Karrie Karahalios, Marcela Musgrove-Chávez, Erika Shehan Poole, John Charles Thomas, and Sarita Yardi. 2009. "Research Ethics in the Facebook Era: Privacy, Anonymity, and Oversight." Presented at the 27th International Conference on Human Factors in Computing Systems Conference, April 4-9, Boston, MA. Retrieved October 28, 2013 (http://www.cc.gatech.edu/computing/pixi/pubs/122-bos.pdf).

Caren, Neal, and Sarah Gaby. 2012. "Occupy Online: How Cute Old Men and Malcolm X Recruited 400,000 US Users to OWS on Facebook." *Social Movement Studies: Journal of Social, Cultural, and Political Protest* 11(3-4): 1-8.

Caren, Neal, Kay Jowers, and Sarah Gaby. 2012. "A Social Movement Online Community: Stormfront and the White Nationalist Movement." *Research in Social Movements, Conflicts, and Change* 33: 163-193.

Carr, Sarah. 2012. "Profile: April 6, genealogy of a youth movement." *Egypt Independent,* April 6. Retrieved October 28, 2013 (www.egyptindependent.com/news/profile-april-6-genealogy-youth-movement).

Conover, Michael, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini 2011. "Political Polarization on Twitter." Presented at the 5th Annual International AAAI Conference on Weblogs and Social Media, July 17-21, Barcelona, Spain. Retrieved October 28, 2013 (http://truthy.indiana.edu/site_media/pdfs/conover_icwsm2011_ polarization.pdf).

Darwish, Kareem. 2013. "Arabizi Detection and Conversion to Arabic." Retrieved October 28, 2013 (http://arxiv.org/abs/1306.6755v1).

Diamond, Larry. 2010. "Liberation Technology." *Journal of Democracy* 21(3): 69-83.

Earl, Jennifer. 2010. "The Dynamics of Protest-Related Diffusion on the Web." *Information, Communication & Society* 13(2): 209-25.

Earl, Jennifer, and Katrina Kimport. 2012. *Digitally Enabled Social Change.* Cambridge, MA: MIT Press.

Earl, Jennifer, Katrina Kimport, Greg Prieto, Carly Rush, and Kimberly Reynoso. 2010. "Changing the World One Webpage at a Time: Conceptualizing and Explaining Internet Activism." *Mobilization* 15(4): 425-46.

Earl, Jennifer, Andrew Martin, John D. McCarthy, and Sarah A. Soule. 2004. "The Use of Newspaper Data in the Study of Collective Action." *Annual Review of Sociology* 30: 65-80.

Earl, Jennifer, and Alan Schussman. 2004. "Cease and Desist: Repression, Strategic Voting and the 2000 Presidential Election." *Mobilization* 9(2): 181-202.

Eaton, Marc. 2010. "Manufacturing Community in an Online Activist Organization." *Information, Communication & Society* 13(2): 174-192.

Faris, David. 2008. "Revolutions Without Revolutionaries? Network Theory, Facebook, and the Egyptian Blogosphere." *Arab Media and Society* 6(Fall).

———. 2009. "The End of the Beginning: The Failure of April 6th and the Future of Electronic Activism in Egypt." *Arab Media and Society* 9(Fall).

Gaffney, Devin. 2010. "#iranElection: Quantifying Online Activism." Presented at the annual meeting of WebSci10: Extending the Frontiers of Society On-Line, April 26-27, Raleigh, SC. Retrieved October 28, 2013 (http://journal.webscience.org/295/2/websci10_submission_6.pdf).

Gamson, William A., and Andre Modigliani. 1989. "Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach." *American Journal of Sociology* 95(1): 1-37.

Garrido, Maria, and Alexander Halavais. 2003. "Mapping Networks of Support for the Zapatista Movement: Applying Social-Networks Analysis to Study Contemporary Social Movements." Pp. 165-184 in *Cyberactivism: Online Activism in Theory and Practice*, edited by Martha McCaughey and Michael D. Ayers. New York: Routledge.

Golder, Scott, and Michael Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333(6051): 1878-81.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. "The Dynamics of Protest Recruitment through an Online Network." *Scientific Reports* 1(197): 1-7.

Gurak, Lauren. 1997. *Persuasion and Privacy in Cyberspace: The Online Protests over Lotus Market-Place and the Clipper.* New Haven, CT: Yale University Press.

Hagemann, Carlo. 2002. "Participation in and Contents of Two Dutch Political Party Discussion Lists on the Internet." *Javnost/The Public* 9(2): 61-76.

Hindman, Matthew. 2008. *The Myth of Digital Democracy*. Princeton, NJ: Princeton University Press.

Hopkins, Daniel J., and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54(1): 229-47.

Hopkins, Daniel J., and Gary King. 2011. *ReadMe: Software for Automated Content Analysis*. R package version 0.99834.

Howard, Philip N., Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid. 2011. "Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?" Report, Project on Information Technology & Political Islam. Retrieved October 28, 2013 (http://pitpi.org/index.php/2011/09/11/opening-closed-regimes-what-was-the-role-of-social-media-during-the-arab-spring/).

Isherwood, Tom. 2008. "A New Direction or More of the Same?" *Arab Media and Society* 6(Fall).

Jasperson, Amy E., Dhavan V. Shah, Mark Watts, Ronald J. Faber, and David P. Fan. 1998. "Framing and the Public Agenda: Media Effects on the Importance of the Federal Budget Deficit." *Political Communication* 15(2): 205-24.

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* Upper Saddle River, NJ: Pearson Prentice Hall.

Jurgenson, Nathan. 2011. "Digital Dualism versus Augmented Reality." Cyborgology. Retrieved October 28, 2013 (http://thesocietypages.org/cyborgology/2011/02/24/digital-dualism-versus-augmented-reality).

Kidd, Dorothy. 2003. "Indymedia.org: A New Communications Commons." Pp. 47-69 in *Cyber-activism: Online Activism in Theory and Practice*, edited by Martha McCaughey and Michael D. Ayers. New York: Routledge.

King, Gary, and Will Lowe. 2003. "An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design." *International Organization* 57(3): 617-42.

Klandermans, Bert. 1992. "The Social Construction of Protest and Multiorganizational Fields." Pp. 77-103 in *Frontiers of Social Movement Theory*, edited by Aldon D. Morris and Carol M. Mueller. New Haven, CT*:* Yale University Press.

Krippendorf, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97(2): 311-31.

Lazer, David, Alex Pentland, Lada Adamic, Sinan Aral, Albert-László Barabási, Deven Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. "Computational Social Science." *Science* 323(5915): 721-23.

Leetaru, Kalev and Philip A. Schrodt. 2013. "GDELT: Global Data on Events, Location, and Tone, 1979-2012." Presented at the Annual Meeting of the International Studies Association, April 3, 2013, San Francisco, CA. Retrieved October 28, 2013 (http://eventdata.psu.edu/papers.dir/ISA. 2013.GDELT.pdf).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval.* Cambridge, UK: Cambridge University Press.

Miles, Matthew B., and A. Michael Huberman. 1994. *Qualitative Data Analysis: An Expanded Source-book,* 2nd ed. Thousand Oaks, CA: Sage.

Monroe, Burt L. and Philip A. Schrodt. 2008. "Introduction to the Special Issue: The Statistical Analysis of Political Text." *Political Analysis* 16(4): 351-55.

Ortiz, David G., Daniel J. Myers, N. Eugene Walls, and Maria-Elena D. Diaz. "Where Do We Stand with Newspaper Data?" *Mobilization* 10(3): 397-419.

Pew Internet and American Life Project. 2012. Retrieved October 28, 2013 (www.pewinternet.org/).

Porter, M. F. 1980. "An Algorithm for Suffix Stripping." *Program* 14(3): 130-137.

Reese, Aaron. 2009. "Framing April 6: Discursive dominance in the Egyptian print media." *Arab Media and Society* 8(Spring).

SalahEldeen, Hany M., and Michael L. Nelson. 2012. "Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?" Presented at *Theory and Practice of Digital Libraries (TPDL) 2012*. September 23-27, 2012, Paphos, Cyprus. Retrieved October 28, 2013 (http://arxiv. org/abs/1209.3026).

Schrodt, Philip A., Shannon A. Davis, and Judith A. Weddle. 1994. "Political Science: KEDS—A Program for the Machine Coding of Event Data." *Social Science Computer Review* 12(4): 561-87.

Shapiro, Samantha M. 2009. "Revolution, Facebook-Style." *New York Times Magazine*, January 25, 2009. Retrieved October 28, 2013 (http://www.nytimes.com/2009/01/25/magazine/25bloggers-t.html?pagewanted=all).

Taghva Kazem, Rania Elkhoury, and Jeffrey Coombs. 2005. "Arabic Stemming Without a Root Dictionary." Presented at the International Conference on Information Technology: Coding and Computing. Retrieved October 28, 2013 (http://jeffcoombs.com/isri/Taghva2005b.pdf).

Wolman, David. 2008. "Cairo Activists Use Facebook to Rattle Regime." *Wired Magazine,* October 2008. Retrieved October 28, 2013 (http://www.wired.com/techbiz/startups/magazine/16-11/ff_ facebookegypt?currentPage=all).