Information, Communication & Society

ISSN: 1369-118X (Print) 1468-4462 (Online) Journal homepage: http://www.tandfonline.com/loi/rics20

What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi Park

Christine Ogan & Onur Varol

INFORMATION, COMMUNICATION & SOCIETY, 2017
VOL. 20, NO. 8, 1220–1238
http://dx.doi.org/10.1080/1369118X.2016.1229006

Routledge
Taylor & Francis Group

What is gained and what is left to be done when content analysis is added to network analysis in the study of a social movement: Twitter use during Gezi Park

Christine Ogan[a] and Onur Varol[b]

[a]School of Informatics and Computing, The Media School, Indiana University, Bloomington, IN, USA; [b]School of Informatics and Computing, Indiana University, Bloomington, IN, USA

ABSTRACT

As social movements relying on the weak ties found in social networks have spread around the world, researchers have taken several approaches to understanding how networks function in such instances as the Arab Spring. While social scientists have primarily relied on survey or content analysis methodology, network scientists have used social network analysis. This research combines content analysis with the automated techniques of network analysis to determine the roles played by those using Twitter to communicate during the Turkish Gezi Park uprising. Based on a network analysis of nearly 2.4 million tweets and a content analysis of a subset of 5126 of those tweets, we found that information sharing was by far the most common use of the tweets and retweets, while tweets that indicated leadership of the movement constituted a small percentage of the overall number of tweets. Using automated techniques, we experimented with coded variables from content analysis to compute the most discriminative tokens and to predict values for each variable using only textual information. We achieved 0.61 precision on identifying types of shared information. Our results on detecting the position of user in the protest and purpose of the tweets achieved 0.42 and 0.33 precision, respectively, illustrating the necessity of user cooperation and the shortcomings of automated techniques. Based on annotated values of user tweets, we computed similarities between users considering their information production and consumption. User similarities are used to compute clusters of individuals with similar behaviors, and we interpreted average activities for those groups.

ARTICLE HISTORY
Received 2 December 2015
Accepted 19 August 2016

KEYWORDS
Social movements; social networks; content analysis; connective action; user behavior; Gezi Park

Social movements, particularly those that use the weak ties formed in social networks, have attracted a lot of attention among sociologists, communication scholars and network scientists in recent years. As such movements spread across the world from Occupy Wall Street to Gezi, researchers study the reasons for them to develop in so many locations, sometimes attributing their strength to the existence of online social networks (Metzger et al., 2014, p. 2).

CONTACT Christine Ogan ogan@indiana.edu School of Informatics and Computing, The Media School, Indiana University, Bloomington, IN 47408, USA

Recent movements are based on a different logic than sociologists have observed in the past, according to Bennett and Segerberg (2012). They are based on a logic of connective action and not the collective action of past movements, they claim. Connective action relies on 'loose organizational linkages, technology deployments and personal action frames' (Bennett & Segerberg, 2012, p. 757) and not on organizationally brokered networks where communication is based on collective action frames more characteristic of a time when digitally networked action was not possible (p. 743). The authors attribute this shift in late modern democracies to the involvement by younger citizens who are 'moving away from parties, broad reform movements and ideologies' (p. 760) Political organizations are also finding ways for younger participants to become politically engaged through 'micro-organizational resources in terms of personal networks, content creation and technology development skills' (p. 760). In the Gezi protests, only 21.1% of participants said that they were affiliated with any political party (Konda, 2014, p. 16). Though Bennett and Segerberg do not claim that collective action has disappeared, they do attribute increased importance to digital networked action through platforms like Twitter that may work in combination with older forms or even replace those forms (p. 760).

This study of the Gezi protests that began in Istanbul in May 2013 focuses on the nature of a movement developed in the theoretical boundaries described by Bennett and Segerberg. The individualized orientations of the participants in Gezi resulted 'in engagement with politics as an expression of personal hopes, lifestyles, and grievances' (2012, p. 744). The authors are careful to point out that though sharing, or the 'personalization that leads actions and content to be distributed widely across social networks' (2012, p. 760), is key to connective action, we do not yet know enough about the nature of the connective action formulations and 'the capacities of sustainability and effectiveness' (2012, p. 761) when this logic for approaching social change is being followed.

Our research is based on Twitter, the social network used to communicate information, persuade others to participate, spread the memes or personal action frames of the movement, and exhibit characteristics of those who played various roles. It will provide better understanding of the ways one social network was used during the most active period of Gezi. We will explore whether Gezi followed the leaderless approach of the Indignados protests in Spain in 2010–2011 (Perea, Cristancho, & Sabucedo, 2014). In an analysis of the nature of the Gezi movement, Aydintasbas asserted that 'protests at Gezi can be perceived in a way as memorandums issued but without soldiers, leaders and political parties dictating' (4 June 2013), indicators of a movement based on connective action.

Another argument for proposing a theory of connective action as a model for this research relates to the conditions in Turkish society at the time Gezi broke out. The government had imposed increasing control over the journalists and news outlets that wished to cover the demonstrations in detail. On the day of one of the largest demonstrations, for example, CNN-Turk (Turkish version of the Cable News Network) was broadcasting a documentary of penguins. In the absence of accurate and timely information from the mass media, demonstrators relied on social networks for information and direction (Oktem, 9 June 2013).

We will add to the literature on social network analysis (SNA) of such movements by combining a network analysis of nearly 2.4 million tweets exchanged during the height of the protests with a content analysis of a randomly selected subset of more than 5000 of those tweets to detect the substance of those messages.

Gezi: motivation and background of the movement

Gezi is often described as an environmental movement because it began as an effort to protect a park in central Istanbul from being converted to a shopping mall. Locating a shopping center in the park was a flash point for the environmentalists based on the recent dramatic rise in the number of shopping centers springing up all over the city. The movement may have started in this spirit, but grew to encompass a wide range of anti-government grievances.

When the police entered Taksim Park early on 30 May 2013 and violently removed the small group of protestors camped there, it prompted a full-scale demonstration by thousands of people in nearby Taksim Square. Protestors in most major cities soon joined in to express their anger with the Justice and Development party government (AK) of then-Prime Minister (and now President) Recep Tayyip Erdoğan. Protestor grievances included the ruling party's authoritarian style of leadership, violations of human rights, use of media censorship and religiously driven policies (Corke, Finkel, Kramer, Robbins, & Schenkkan, 2014). It took several weeks of harsh government crackdown with massive police intervention and excessive use of water cannons, tear gas and plastic bullets to disperse the demonstrators.

Unlike social movements in Egypt and Tunisia, Gezi did not result in overthrowing the government. The brutal actions of the police force across the major Turkish cities and the arrest of anyone associated with street demonstrations or expression of critical comments were sufficient to quell the protests if not calm the mood of the opposition.

Literature on Gezi and other recent uprisings

A theoretical approach has been adopted by some social movement scholars.

Bennett and Segerberg's logic of connective action in social movements is one example. Lim (2012), who concluded that the use of connective action led to the success of the Tunisian uprising, followed up on their work; and Caraway (2016) used it to explain the networked structure of the organization responsible for the labor actions at Walmart. Verdegem, D'heer, and DeGrove (2015) applied the theory to labor unions' response to austerity measures in Belgium. Some other research of recent social movements might be classified as descriptive or analytical (Aknur, 2014; Howard & Hussain, 2011; Wojcieszak & Smith, 2014; Zhuo, Wellman, & Yu, 2011).

Howard and Hussain (2011) claimed a powerful role for social media in the Arab Spring despite on-the-ground causes that provoked the demonstrators. Writing for *The Guardian*, Beaumant (2011) said that 'often, the contribution of social networks to the Arab uprisings has been as important as it also has been complex, contradictory and misunderstood'. Systematic research of various social movements has supported that position (see Gerbaudo, 2012; Lim, 2012; Stepanova, 2011; Tufekci & Wilson, 2012).

Empirical studies have adopted several methodologies to explore the role of social media in social movements. Using a survey methodology, Albacete, Theocharis, Lowe, and Van Deth (2013) found no relationship between online users of social media and offline participants in the Occupy Wall Street movement. Valenzuela, Arriagada, and Scherman (2012) surveyed youth to determine the relationship between their Facebook activity and protests for political change in Chile, finding a link between social media use for news and socializing but not for self-expression. Surveys conducted during the Tahir Square demonstrations in Cairo found that Facebook was used extensively to communicate information about the street protests (Tufekci & Wilson, 2012). Participants acted as citizen journalists, publishing images and organizing activity through social media.

These studies are important, in that they provide social movement participants' self-reports of their activities on social media. However, they were not corroborated in any way with the actual content of the participants' Facebook or Twitter posts. SNA is able to determine what was posted, to whom it was directed, the size of the audience for that information, and what connections to other information and people were made.

Metzger et al.'s analysis of the total number of tweets (more than 30 million from 3 million distinct users) during the Gezi protest, examined the degree of user influence (defined as the number of retweets) over the information spread through online networks (2014, p. 4). Images and tweet texts from the streets were associated with the authors' measure of influence, but that the association depended on the number of followers of the tweeter, illustrating the importance of network centrality. Other factors, such as the type of information provided, were critical to the degree of influence – while external shocks in the form of violence were not significant in the development of influence (p. 30).

In previous research, we investigated Gezi protest user roles and their evolution, the spatio-temporal cues in the Twitter discussion, and the impact of events in the streets on the online user behavior in a SNA of 10% of the tweets during the 27 days of the protests (Varol, Ferrara, Ogan, Menczer, & Flammini, 2014). We found that as the Twitter discussion spread throughout the world, trending hashtags related to Gezi Park were also worldwide in scope. Four types of user roles were exhibited during the demonstrations (common users, broadcasters, influentials and hidden influentials). Influentials were defined as users whose content was retweeted most often. Influentials were further divided into two groups – those with more followers than followees and those with fewer followers (or the hidden influentials) (Varol et al., 2014, p. 90).

A study by Theocharis, Lowe, van Deth, and García-Albacete (2014) combined the method of content analysis with limited network analysis of Twitter postings in three separate social movements to determine if the social medium was being used to change or contribute to the political communication, mobilization and organization of movements in the United States (Occupy Wall Street), in Spain (Indignados) and Greece (Aganaktismenoi). Theocharis et al. sought to 'complement existing approaches by examining how Twitter is used for political mobilization and promotion of political action from a cross-national perspective and across different movements' (2014, p. 203). Coders fluent in the tweets' main language analyzed 2000 randomly drawn tweets from each of the three movements. Network analysis of in- and out-degree centrality was also conducted (p. 207). 'The results reveal that specific uses of Twitter are largely consistent across movements (e.g., the platform is mainly used for conversation and linking information, and less so for action and organization') (p. 215). The structure and content of the Twitter exchanges varied by the national context for the movement (p. 215).

Role of Twitter in the communication ecology of the protest

Critics taking a media ecology approach to the study of protests point to the analytical fallacies of abstracting social media in general, or Twitter in particular (Segerberg & Bennett, 2011, p. 199). To do so is to isolate technologies such as Twitter from the larger technological and social contexts, they argue. However, in the case of Gezi, Twitter and other social media were the best channels for protestor communication, given media censorship or failure to publish protest information. In addition, at least 29,000 websites had been blocked by the government in 2013, according to the engelliweb.com, a site that documents this information. So for Gezi, Twitter became key to communication inside and outside the movement.

Segerberg and Bennett also argue that Twitter and similar technologies tend to be embedded in the larger complex protest spaces where a range of actors operate, and are therefore worthy of examination. In Gezi, the groups that came together in the streets and on Twitter included labor unions, community organizations, football fan clubs, lesbian, gay, bisexual and transgendered (LGBT) groups, feminists, mothers of protestors, anti-war protestors, and ethnic and religious minorities. Twitter was used to express collective grievances against the government and held groups together, if only briefly, to confront a common opponent.

Our study examines the use of Twitter during the protests as a means of understanding its role during the demonstrations as an organizational tool as well as a 'window on the larger protest ecology itself' (Segerberg & Bennett, 2011, p. 200; see also Poell, 2014, p. 717).

Social network analysis vs. content analysis

From previous research, we can see that the two methods – SNA and content analysis – generate different kinds of questions about social movements and also yield different results. In an era where it is possible to locate and analyze enormous datasets – such as those resulting from downloading social media interactions – scholars have been eager to learn as much as possible about patterns in those exchanges to develop new theories of communication. In the case of social movements, research has been able to discover the structure of the networks of the participants as well as their behavior; and identify key players in the networks and their tie strengths to other members.

Many aspects of human behavior considered impossible to address before the widespread collection and organization of information, records and communication (such as that generated by social media networks) are now commonplace. But as Boyd and Crawford (2012) have written, we should be cautious about viewing this 'socio-technical phenomenon' as the sole answer to important societal questions. Though the authors are positive about the contributions of big data to reframe 'key questions about the constitution of knowledge, the processes of research, how we should engage with information and the categorization of reality' (p. 665), they point to problems of representativeness of samples, fairness and completeness of access of social media content, limitations of tools of analysis and the ethics of use.

Network analysis is also not able to produce conclusions based on the nature of the message content. By the same token, content analysis alone is likely to produce only descriptive results because of its own shortcomings, in particular the difficulty of accessing accurate demographic information about the social network users themselves. In Twitter, researchers cannot always determine the age, gender, educational level or even the location of the users. Confined to examining the content of 140 characters, researchers may not be able to easily determine meaning. Hand coding of tweets is also limited to the number that coders can possibly manage for any given study. Subsets of the total number of tweets on a given subject may not be representative of the total and must be small enough to be manageable.

In a case study of news sourcing on Twitter, Lewis, Zamith, and Hermida (2013) combine computational and manual content analysis methods as a hybrid approach to 'preserve the strengths of traditional content analysis, with its systematic rigor and contextual sensitivity, while also maximizing the large-scale capacity of Big Data and the algorithmic accuracy of computational methods' (2013, p. 34). The number of categories and values for each variable were necessarily limited. However, this study advanced the automation of hand-coded content analysis of social network data a bit.

Another set of authors combined content analysis with computational methods to compare the mobilization patterns and the action repertoires in three different social movements – Occupy Wall Street, Indignados and Aganatismenoi (Theocharis et al., 2015). Following on Earl and Kimport (2011, p. 76), the authors point out that social networks like Twitter allow for a larger number of participants in a social movement. They selected three different movements because of their differing national locations and languages, but with potential commonalities (pp. 205–206).

By combining the results of both types of analysis, Theocharis et al. (2015) found that Twitter was primarily used for conversations and linking information rather than calling for action or organizing the movements in all three countries (p. 215):

> Although Twitter certainly supported protest communication, based on this analysis it does not seem to have altered the underlying processes that drive collective action and organization by, for example, causing a surge in online contributions that we could comfortably classify as political action. (p. 217)

Combining network and hand-coded content analysis of Twitter conversations during the Gezi protests, our study builds on the work of Theocharis et al. (2015) to determine whether a similar pattern of communication took place on this social medium in 2013.

In network analysis of social media, automated methods have been used in a wide range of application domains. Most of these applications provide insights regarding user interests (Hong, Doumith, & Davison, 2013; Kim, Jo, Moon, & Oh, 2010), detection of external events (Becker, Naaman, & Gravano, 2011; Mathioudakis & Koudas, 2010; Sakaki, Okazaki, & Matsuo, 2010) and information diffusion (Bakshy, Rosenn, Marlow, & Adamic, 2012; Baños, Borge-Holthoefer, Wang, Moreno, & González-Bailón, 2013; Ferrara, Varol, Menczer, & Flammini, 2013; Weng, Menczer, & Ahn, 2013, 2014; Weng, Ratkiewicz et al., 2013). In addition to the positive impacts, social media have the potential to threaten free speech and create a medium that disseminates misleading information. Research on social networks has also focused on detection of such problems (Ferrara, JafariAsbagh, et al., 2013; Ferrara, Varol, Menczer, & Flammini, 2016; Ferrara et al., 2014; JafariAsbagh, Ferrara, Varol, Menczer, & Flammini, 2014), as well as those pertaining to social bots (Ferrara et al., 2014; Subrahmanian et al., 2016) and fact checking (Ciampaglia et al., 2015).

The automated methods described above share one common property: identification of significantly different groups. These methods rely on features specific to the domain of interests such as network, content, sentiment and user metadata to separate the groups. However, more specific problems, such as the need to determine roles played by users, require more detailed information than can be extracted through automated methods. In our case, that information was revealed through manual content analysis. The research questions for our study were as follows:

RQ1: What was the purpose of the tweets posted during the Gezi protests?
RQ2: What roles were played by those who tweeted during the Gezi protests?
RQ2a: What types of content did they share?
RQ2b: How were they connected?
RQ3: What evidence of leadership activity emerged during the demonstrations as reflected in the Twitter conversations?
RQ4: What is the impact of language cues on the identification of the users' purpose?

Methodology

The research adopted a multi-method approach (SNA combined with content analysis) to study the posts on Twitter during the Gezi movement, from its inception in Gezi Park through the first three weeks of protest activity in the streets throughout Turkey, or 25 May—20 June 2013.

Posts from the micro-blogging platform Twitter were collected for analysis. Posts included interactions (text, URL's, external media content, etc.) between users and other participants on the platform by creating social ties (follower/followee relations), the retweeting of others' content to broadcast the same message to their friends, and mentions of other users in their posts (@somebody's name). Hashtags included in collected tweets were used as keywords to summarize a discussion topic or to convey a message in a shortened format. In this study, #direngeziparki and #occupygezi are some of common hashtags we observed (Table 1). The dataset collected for this study represents a 10% random sample of all public tweets streamed in real time.

Along with the textual information in the tweets, we also collected metadata about users such as screen names, follower/followee counts, self-reported location and more. We extracted information about geo-location of users (the latitude/longitude coordinates for 43,646 tweets) for the contents posted with Global Positioning System (GPS)-enabled devices. We adopted this subset of geo-located tweets to study the spatio-temporal nature of the protest.

To capture a topical discussion about Gezi events, we adopted a hashtag seed-expansion procedure: first, we handpicked the most popular Gezi-related hashtag (#direngeziparki)

Table 1. Examples of commonly used hashtags in different categories.

Commonly used hashtags		Local protest hashtags	Government supporters' hashtags
#direngeziparki	#bizeheryertaksim	#direnankara	#dunyaliderierdogan
#occupygezi	#gezideyim	#direnbesiktas	#seviyoruzsenierdogan
#eylemvakti	#7den77yedireniyoruz	#direnizmir	#seninleyizerdogan
#occupyturkey	#heryertaksimheryerdirenis	#direntaksim	#seninleyiztayyiperdogan
#direngezi	#korkakmedya	#direnadana	#youcantstopturkishsuccess
#tayyipistifa #bubirsivildirenis	#hukumetistifa	#direndersim	#weareerdogan
#wearegezi	#dictatorerdogan	#direnistanbul	#yedirmeyiz
	#siddetidurdurun	#direnrize	#turkiyebasbakanininyaninda

and we extracted all tweets containing this hashtag. We then built a hashtag co-occurrence list, and selected the top 100 hashtags co-occurring most frequently with our seed. We generated a final list of hashtags and collected all tweets containing at least one of these hashtags. These hashtags were manually divided into three categories: general interest hashtags, hashtags with location associations and those used by government supporters. Samples of these categories along with example hashtags can be found in Table 1.

Overall, we collected 2,361,335 tweets associated with the Gezi movement, generated by 855,616 distinct users and containing a total of 64,668 unique hashtags. Among these 2.3 million tweets, 1,475,494 were retweets and 47,163 were replies from one user to another.

In the geographical distribution of tweets, the majority were posted from Turkey as expected. Europe and the USA also had significant participation during events. Support from abroad came largely through social media. Turkish citizens living abroad also supported the protests by organizing similar demonstrations in other countries. Turkish, the preferred language of users in Twitter, reflects this multinational participation. Following Turkish, English was the second most used language, with Spanish, Portuguese, French and German being used less frequently.

We investigated unique hashtags used in conversations along with users participating in discussions and their accumulated number of frequencies. More than 60% of the users joined the conversation in the first days, and participation varied according to the occurrence of several external events. Hashtags were continuously introduced during the protests, and the number of unique hashtags increased constantly.

In order to study conversation and user roles through content analysis during the protest, we randomly selected users from our collection. In this study, 135 unique users and their 5126 tweets were coded for 18 variables. During the 27-day period, an average of 38 tweets per user were posted for this sample. In addition to the automated variables available from Twitter (number of hashtags, mentions, date (month, day and time), whether the tweet originated with the user or was retweeted, and number of times the post was retweeted), we coded tweets for actual content. Posts were coded for primary and secondary purpose alongside the types of information shared (medical, legal, location of authorities, safe places, etc.), types of questions asked and personal positions taken, if any. We also coded for the language used, whether a link was included and to what type of information, and whether one of the common protest memes was included.

One of the authors was the primary coder along with two assistants. All were able to code in Turkish and English. The other languages used in the sample were translated wherever possible. After extensive training, the three coders worked in the same space, coding simultaneously and consulting one another frequently until agreement was achieved. Since the bulk of the coding was done by the author, reliability tests were not conducted.

Findings

From content analysis

One of the affordances of Twitter is being able to spread a message widely through retweets, the use of hashtags and mentions of others within the tweet. Our analysis of the content of the sample of tweets revealed that none of these methods was effective in securing the largest number of movement supporters. The mean number of times a post was retweeted was 14, while the mean number of mentions was .31 (44.1% to another individual, 40.6% to a group and 13.3% to a mass media source), Furthermore, one-third of the messages coded were retweeted from others.

Many tweets attempted to either direct their messages to international media sources, to comment on the lack of published information, provide information reporters might not be aware of or to link to other information that could be useful to the media. These tweets were frequently posted in English. Of the 698 messages written in English, 68% shared information about the protests, while an additional 7% only provided a link to other information. We also coded a secondary purpose for a tweet, and of those, 25% shared information while an additional 31% shared only a link. The shared information provided details about the demonstrations themselves or the actions of authorities in 64% of the tweets in English.

In any of the languages used, information sharing was the top primary purpose for tweeting (43.1% providing specific information) and the top secondary purpose (24.5%), and if linking to other information is added to that total, the percentage rises to 53.9% of the primary purposes and 34.2% of secondary purposes (Table 2).

We defined tweets demonstrating leadership characteristics as those where the poster was suggesting an action to take, directing demonstrators to do something or refrain from action; or other statements suggesting the individual was suggesting a new course of action. In network analysis, leaders are often defined as 'influentials', or those who have a large number of followers and whose messages are frequently retweeted. But as our content analysis revealed, the messages of influentials do not necessarily direct action. Only 6.1% of the primary and 15.3% of the secondary purpose included any leadership

Table 2. Descriptions of available coding variables and their observed frequencies in our dataset.

Variable name	Variable values	# of tweets	# of tweets and retweets
Purpose	Sharing specific information heard about	849	1559
	Opinion statement	292	506
	General information	289	467
	Links to outside information	230	354
	Support for movement	103	239
	First person witness	130	226
	Ask for help or warn	81	209
	Provide direction	110	184
	Hashtag	86	142
	Information dissemination	62	135
	Media coverage	52	80
	Emotional statement	16	29
Position	General opinion	485	710
	Anger against govn/PM/police	244	435
	Support for movement/motivational	214	363
	Praise or support for groups/individuals	91	183
	Critical statement about people/business or organization of demonstrations	84	161
	Pro-government/police or anti-gaze opinion	59	80
Info share	Location of police Toma's, arrests, beatings, info about weapons	414	638
	Scheduled demonstration places, actions of demonstrators	373	596
	Specific info medical, legal, technical, food, safe places	161	297
	Info about specific groups, unions, gays, missing, politicians, etc.	147	215
	About media and availability	103	163

characteristics. When we selected only those tweets that did contain the influential focus, a mean of 10.7 posts were retweeted for the primary purpose and 17.7% for the secondary purpose, while about one-third of all such messages were retweets of someone else's posts. To us, this appears to be quite a small sphere of influence. Furthermore, posts reflecting leadership were distributed across a wide range of users, with only two users posting as many as 15 different messages that directed or suggested any kind of action.

For any post that was retweeted (from once to 3096 times), the most common type of retweet was information sharing in nature – general and specific information, information passed on from others as well as about events witnessed by the protestors themselves. The second most frequent type of retweet was a statement of personal opinion. A similar pattern held for the second ranked purpose for the tweet. When information was shared, the most common type was logistical – related to the police, water cannons or tear gas use or arrests (33.4%) while the second highest category was information about a scheduled demonstration or particular actions of the demonstrators in locations around the country (31.2%).

Analyzing content helps us to obtain crucial information about users who are inaccessible through network analysis. In this research, content analysis demonstrated that information sharing was the most frequent role played in this social movement. Following well behind that role were those of expressing personal opinion, encouraging others to persist in their protest activities, and finally, to direct others in specific activities.

Results from metadata and interactions

Here, we present our results on the classification of roles by using features extracted from user metadata and interaction patterns. We are considering the hand-coded content as our ground truth, or the information provided from direct observation, as opposed to that provided by inference. Our first experiments use content as features to predict categories of variables assigned by coders. This experiment aims to show the performance of simple automated method on the coding of tweets. We used the content analysis of the tweets to cluster users based on the content they created and posted. The result of the clustering task was used to assign user labels for investigating different groups of users that share similar characteristics of information production and consumption.

Language use of tweets

As mentioned above, the hand coding of tweets was conducted by three individuals – a long and tedious process. Several tests were performed on the amount of information an automated system can capture through tweets only using textual n-gram tokens.[1] We want to build dictionaries specific to the domain of interest to identify coded values. Should we be able to do that, it would speed up the hand-coding process considerably for any future protest movement in Turkey.

In Table 3, we present the most discriminative tokens for three variables (the main purpose of the tweet, types of information shared by those who posted information and personal position of the tweeter when that opinion was coded as the main purpose). We present the most discriminative tokens for each variable and its values. Most tweets were written in Turkish, so each token was translated into English in the table. The significance levels for each token were computed using a χ^2 test if a token was observed

Table 3. Significant tokens best representing categories for a given variable. Significance of tokens computed by χ^2 test and only examples of highly significant tokens are presented here.

Variable name	Variable value	Top significant tokens[a]
Purpose	Sharing specific information heard about	Detention, infirmary, gas, water canons, tear gas bombs were thrown water cannons, don't go
	Opinion statement	Tayyip, back off, resign, conscience
	General information	Koç companies, Guarantee Bank, Milan Gaz, propane gas, The people, nationality
	Links to outside information	Live broadcast, Sleepless Magazine, Anonymous
	First person witness	Gezi Park, intervention, police, tear gas bombs, water canons, they won't allow us, barricades
	Ask for help or warn	Emergency, be careful, rt, their doors, ambulance needed in Taksim
	Provide direction	Download Tor, dns, against internet censorship
	Information dissemination	time, in Ataturk Park, Culture Park, evening, we are going they are gathering, Gezi Park
Position	General opinion	Don't go, be careful , first personal safety, twitter, facebook
	Anger against govn/PM/police	Resign, Erdogan, prime minister, rte, go, Anonymous has published a report, pepper gas, disproportionate
	Support for movement/motivational	Continue the resistance, everywhere is Taksim, No to violence, Support Halk TV
	Praise or support for groups/individuals	Kudos, Thanks, You are cool, my eyes are welling up, we are children of Ataturk
	Critical statement about people/business or organization of demonstrations	Biased media, sold, we condemn-watch us, lying news
	Pro-government/police or anti-gezi opinion	It's certain who will be first in the elections, breaking up the big game, let's make history
Info share	Location of police Toma's, arrests, beatings, info about weapons	Police, tear gas, riot police, Tomas. Tear gas. Detentions, plastic
	Scheduled demonstration places, actions of demonstrators	statues, Taksim, Bursa, public square, Culture Park, fsm, veteran
	Specific info medical, legal, technical, food, safe places	Infirmary, emergency, doctor, wifi, pharmacy, facebook, medication
	Info about specific groups, unions, gays, missing, politicians, etc.	Secret, Sureyya Onder, King of Morocco, Need an ambulance, Near Benetton
	About media and availability	NTV, CNN, Ulusal, Halk TV, RTUK

[a]Tokens translated into English when necessary.

for a particular value more frequently than the expected count. Examples of those tokens show associations between the descriptions of values.

The very abbreviated format of tweets makes it difficult to determine what is exactly meant by each one. Using automated techniques, we can capture textual cues, such as significant words, co-occurrences and sequences. When manually coding tweets, deeper insights (though still not entirely accurate) may be extracted from the tweets by the coders. However, a high cost in both time and effort is required. We attempted to classify the values of three of the variables based on the tokens extracted. We extracted 3-gram tokens and computed tf-idf vectors (Jones, 1973) to account for the importance of each token for tweets used in the protests. These important values for available tokens were translated into numeric vectors. Vector representation of tweets was used to train the Random Forest classifier[2] along with the annotated variables as a labeled dataset. A classifier, with 100 estimators that use gini coefficient as a quality measure for splits, is able to learn tokens that are representative for each variable and provides discriminative power. We evaluated the performance of the classification task using 10-fold cross-validation. Our analysis with simple logistic regression models also reached similar conclusions.

Table 4. Classification experiment on predicting category of hand-coded data using three-gram textual features.

Variable	Accuracy	Precision	Recall	F1-score
Purpose	0.24	0.33	0.25	0.22
Position	0.35	0.42	0.35	0.36
Info share	0.51	0.61	0.51	0.52

For all coded variables, we evaluated performance on detecting values using tokenized tweets (Table 4). Our best performance was the detection of the 'information share' variable with 51% accuracy considered acceptable due to 70% performance increase with respect to random baseline. Detecting the 'purpose of the tweet' and the 'position of the user' was a more difficult task for the automated system because it requires the more nuanced decision-making by human coders. Therefore, we were only able to achieve 24% and 35% accuracy on those categories, respectively, which is only slight increase from baselines and classifiers performed on these tasks poorly possibly due to sparsity of the textual content. A quick read through of the top significant tokens in the table tells us that transferring these tokens to an analysis of some other social movement in the world would not be successful.

Clustering users by tweet annotations/variables

Clustering users based on their social connectivity and production of content provides an overall view of dynamics and behavior. Using information obtained through content analysis allows us to identify users by the roles they played in this social movement. In this section, we use an unsupervised clustering framework[3] to identify groups of users producing or disseminating similar content according to our content analyzed dataset.

We can study individuals by their content production and consumption preferences. Individuals share content based on their motives related to the protest participation. Analyzing content helps us to obtain crucial information about users. We compared users by coding the information in their tweets. Each user is represented by the number of posts coded for each distinct variable and its values. We also considered tweets and retweets separately to highlight the differences between information need and creation. To compare users, we computed the similarity between users by the cosine similarity of the coded distribution of their content (a type of correlation).

We considered three of the variables in the content analysis, namely primary 'purpose', 'position' and 'information share'. Combinations of those three variables were also considered.

We clustered users by the content of their tweets using hierarchical clustering. We leveraged coded similarity between tweets and retweets by user to compare the similarity of their content. In this technique, the distance between each pair of items is used to compute clusters from bottom up by agglomerating similar users in each step. In this analysis, we used complete linkage to merge clusters in the hierarchy. One of the advantages of hierarchical clustering is the use of the tuning threshold to decide the number of clusters. In Figure 1, we present clusters of users in a distance matrix and hierarchy of clusters. In this analysis, the values of three variables (information share, purpose and position) are combined for both tweets and retweets by user.

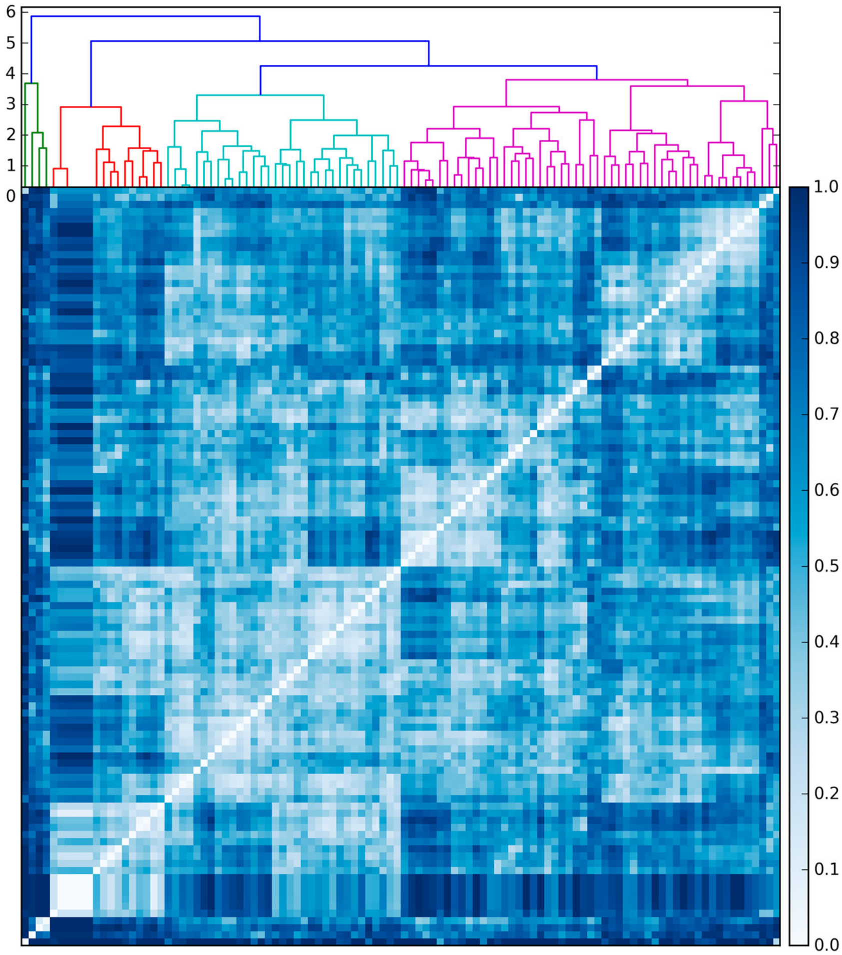

Figure 1. Hierarchical clustering of the users by using their similarities based on content annotations.

In this part of the analysis, clusters to which users belong were used as our ground-truth labels. Using our hand-coded tweet dataset, we can compute clusters in multiple ways by varying the number of clusters and variables used to compute the similarity among users. We investigated alternatives of computing similarities by using the coded tweets.

In Figure 2, we present pairwise Normalized Mutual Information scores (NMI or a method of determining how well one classification is able to predict a second classification) of clustering outcomes using different values of the variables for 10 clusters in each analysis. A higher NMI score indicates a significant amount of overlap between user assignments to groups. We observed that the correspondence between clusters was low when the values of 'purpose', 'position' and 'information sharing' were used separately. However, the contribution of 'purpose' dominated the clustering outcomes due to the higher NMI scores of comparisons involving the 'purpose' variable. The information dissemination values of the purpose variable in the tweets might discriminate users better than any other information produced or shared. Roles and group memberships also relate to the motivation of users.

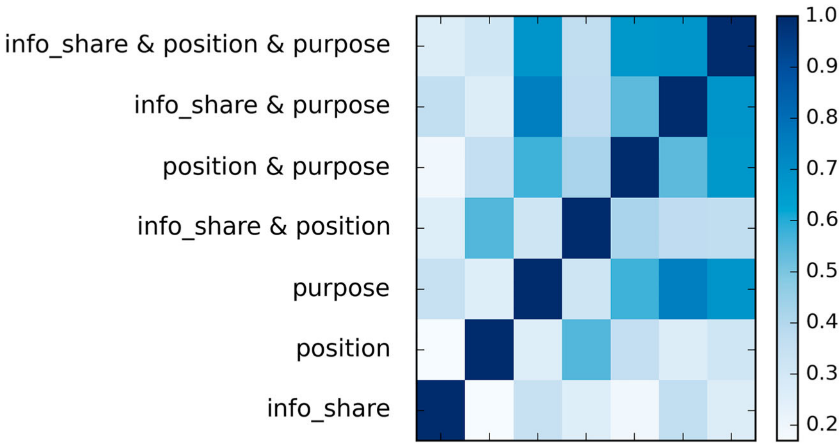

Figure 2. Consistency of clustering identified by various tweet annotations. Similarities between clustering computed by NMI.

Once we explore alternative clustering outcomes, we can select the most appropriate clustering by looking at the distance matrix and the dendrogram in hierarchical clustering shown in Figure 1. In this representation, we can choose five clusters as represented by different colors and branches in dendrogram.

We can further explore content created or shared by average users in each group. We can observe differences between average behaviors of groups. In Table 5, we summarize the most common variables for each group of users. For instance, group one has only

Table 5. Average behavior of users in each cluster. Most common five activities reported for each group along with their amount and type of share (whether retweet or tweet).

Group size	Variable	Description	Value (type: tweet or retweet)
4	Purpose	First person witness	11.75 (T)
	Info share	Location of Police Toma's, arrests, beating, info about weapons	7.0 (T)
	Info share	Scheduled demonstration places, actions of demonstrators	5.75 (T)
	Purpose	Hashtag	4.0 (T)
	Purpose	Links to outside information	1.5 (T)
16	Purpose	Others	0.75 (T)
	Position	General opinion	0.38 (T)
	Purpose	Sharing specific information heard about	0.31 (T)
	Info share	Location of Police Toma's, arrests, beating, info about weapons	0.25 (T)
	Position	General opinion	0.25 (R)
33	Position	General opinion	0.18 (T)
	Purpose	Sharing specific information heard about	0.18 (T)
	Purpose	Others	0.15 (R)
	Purpose	Sharing specific information heard about	0.15 (R)
	Info share	Scheduled demonstration places, actions of demonstrators	0.12(T)
28	Purpose	Sharing specific information heard about	0.32 (T)
	Info share	Scheduled demonstration places, actions of demonstrators	0.18 (T)
	Info share	About media and availability	0.14 (T)
	Purpose	Links to outside information	0.14 (T)
	Info share	Location of Police Toma's, arrests, beating, info about weapons	0.11 (T)
25	Purpose	Opinion statement	0.32 (T)
	Purpose	Sharing specific information heard about	0.32 (R)
	Purpose	General information	0.28 (R)
	Info share	Location of Police Toma's, arrests, beating, info about weapons	0.2 (R)
	Purpose	Support for movement	0.2 (T)

four users and represents the smallest group among the five clusters. In this small group, we observe tweets related to witnessing and information related to on-the-ground events. Users in this group were apparently quite active in the protests. The second group has other interests, but also shares its opinions and information heard from others. The last three groups are also active in disseminating information. They differ by tweet content categories either by conveying position and purpose of user or by sharing information.

We further explored content created or shared by average users by group. Differences between average behaviors of groups can be viewed in Table 5 where we summarize the most common types of content for each group. For instance, group one had only four users and represents the smallest group among five clusters. In this group, most of the tweets related to witnessing certain on-the-ground events. Users in this group appear to have been very active protesters. The difference between groups by variable and message type was calculated. Groups two and four mostly created original content, but the third and fifth groups tended to broadcast content produced by others (retweets). Content from the information share category, which detailed information about scheduled events and location of authorities, was mostly created by group four and broadcasted by group five. Groups two and three stated personal position/opinions about the protests. These groups produced or broadcasted specific types of content. There was overlap of some of this content among groups, but the amount of that content varied.

Discussion

In this study, we attempted to answer questions about the use of a social medium, Twitter, during a social uprising in Turkey. In the past, scholars have generally addressed the role played by social media in such circumstances through network analysis, content analysis or survey where scholars from separate disciplines rely on their own methodologies. Here, we combined content analysis of a subset of tweets that were extracted during the time of the social movement to learn what roles were being played by the users of the social medium. We were especially interested to see if the posts could identify individuals who displayed leadership characteristics in the uprising.

We focused on leveraging hand-coded data with automated techniques to identify distinct behavioral groups. The content analysis reflected the temporal relationships between the messages through tweets and real-world events – that is, the focus of the tweets shifted based on external events (actions or speeches by authorities, deaths of demonstrators, etc.). We also conducted an analysis of the language use for different purposes during the demonstrations. Our investigation of the textual messages through the use of extracted tokens illustrates the difficulty of trying to automate the process of hand coding. Cluster analysis of the hand-coded dataset provided clusters of users that can be used as ground-truth labels.

We have learned that the primary role for users was in information dissemination to other participants in the demonstrations, but few messages indicated any leadership role and users who tweeted directive messages did not consistently do so. This supports the research of Theocharis et al. (2015), who found that providing links to information was the primary purpose of tweeting.

The second most frequent role played was the expression of personal opinion, so those posts might have reflected opinion leadership, but many of the opinions were mere expressions of anger or frustration than ones that could clearly be identified as leadership. The various purposes in the tweets of any one user are not so hard to understand. Participants in demonstrations like that of Gezi Park were not single minded in the information they tweeted. In addition to commenting about an unplanned external event, users were also distracted by impromptu events created by other demonstrators or by graffiti written on a wall or a new song being sung about the movement.

We learned also that hand coding of tweets is both tedious and time consuming, but our automated analysis of the hand-coded tweets tells us that for at least some basic information, semi-automated data coding could be conducted through crowdsourcing methods. In our case, we illustrated that the combined variables of purpose, information sharing and opinion yielded 50% accuracy. The content analysis task could be broken down to simple questions that could be easily answered and save time and cost. Though it may be difficult to accomplish, researchers are beginning to explore this method of combining hand coding with automated analysis. We found that the hand coding of tweets was especially difficult because of the 140 character limitation. Without the context that would be found in a longer document, the abbreviated nature of the tweet can lead to ambiguous interpretations of its meaning. It can also be hard to determine the tone of the tweet to identify sarcasm or irony, for example.

Our study also shows that in a social movement like Gezi where protestors without affiliation to a political organization cannot rely on Twitter or perhaps any social medium to replace the structure of an organization to formulate and carry out goals that can lead to a successful outcome. When such a small number of participants who are distributed across a large geographical space tweet messages that attempt to direct decision-making and action in the face of a large government and police force, they are not likely to be successful. At least in the case of Gezi, we suggest that social media functioned as a tool in the connected action of the protestors, but much more was required to effect the desired outcome. It appears that some degree of collective action is still required to accompany the connected action made possible through the use of social networks in a social movement.

We are encouraged by the results of this research that refinements could be made in both hand coding and automated coding, and that researchers who specialize in content analysis and network analysis can work together to determine future directions for studies that will be able to answer a wider range of questions than previously possible. Automation of the coding process can be also mediated through an online system that can harness large datasets and recommend most ambiguous data points for expert coding. We call on others to advance this process in their work.

Notes

1. An *n*-gram is a contiguous sequence of *n* items from a given sequence. In this case, the sequence is contained in the 140-character tweet. In our analysis, a token is a grouping of words in the tweet. It could be one or a combination of contiguous words in the tweet.
2. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True (default).

See scikit-learn for more information: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

3. Clustering is the process of dividing data into subsets so that each subset shares common characteristics. When the clustering is unsupervised, the model does not have the correct results during the training. The method can be used to cluster the input data in classes based only on their statistical properties. See Ciro Donalek, 'Supervised and Unsupervised Learning.' Online: http://www.astro.caltech.edu/~george/aybi199/Donalek_Classif.pdf.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Christine Ogan is professor emerita from the Indiana University School of Informatics and Computing as well as the Media School. Currently she is an adjunct professor at Warrington College of Business, University of Florida. Her work in recent years relates to the use of social media in social movements, especially by the Turkish diaspora in Europe. She was also a member of the EU Kids Online project based at the London School of Economics.

Onur Varol is a PhD student at Indiana University, Bloomington. His work relates to detection and analysis of online manipulation and threats.

References

Aknur, M. (2014). The Gezi Park protests as a social movement in Turkey: From emergence to coalescence without bureaucratization. *Studia Universitatis Babes-Bolyai-Studia Europaea 1*, 295–320.

Albacete, G. M., Theocharis, Y., Lowe, W., & Van Deth, J. W. (2013, March 4). *Social media mobilisation as a prompt for offline participation? Analysing Occupy Wall Street Twitterers' Offline Engagement with the Movement*.

Aydintasbas, A. (2013, June 4). Gezi Park unrest uniquely Turkish. *Al Monitor*. Retrieved from http://www.al-monitor.com/pulse/politics/2013/06/gezi-park-is-uniquely-turkish.html

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012, April). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web* (pp. 519–528). Lyon: ACM.

Baños, R. A., Borge-Holthoefer, J., Wang, N., Moreno, Y., & González-Bailón, S. (2013). Diffusion dynamics with changing network composition. *Entropy*, *15*(11), 4553–4568.

Beaumant, P. (2011, February 25). The truth about Twitter, Facebook and the uprisings in the Arab world. *The Guardian*. Retrieved from https://www.theguardian.com/world/2011/feb/25/twitter-facebook-uprisings-arab-libya

Becker, H., Naaman, M., & Gravano, L. (2011). Beyond trending topics: Real-world event identification on Twitter. *ICWSM*, *11*, 438–441.

Bennett, W. L., & Segerberg, A. (2012). The logic of connective action: Digital media and the personalization of contentious politics. *Information, Communication & Society*, *15*(5), 739–768.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679.

Caraway, B. (2016). Our Walmart: A case study of connective action. *Information, Communication & Society*, *19*(7), 907–920.

Ciampaglia, G. L., Shiralkar, P., Rocha, L. M., Bollen, J., Menczer, F., & Flammini, A. (2015). Computational fact checking from knowledge networks. *PloS One*, *10*(6), e0128193.

Corke, S., Finkel, A., Kramer, D. J., Robbins, C. A., & Schenkkan, N. (2014). *Democracy in crisis: Corruption, media, and power in Turkey*. Freedom House. Retrieved from https://freedomhouse.org/sites/default/files/Turkey%20Report%20-%202-3-14.pdf

Earl, J., & Kimport, K. (2011). *Digitally enabled social change: Activism in the internet age*. Cambridge, MA: MIT Press.

Ferrara, E., JafariAsbagh, M., Varol, O., Qazvinian, V., Menczer, F., & Flammini, A. (2013, August). Clustering memes in social media. In *Advances in social networks analysis and mining (ASONAM), 2013 IEEE/ACM international conference on* (pp. 548–555). Niagra: IEEE.

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2013, October). Traveling trends: Social butterflies or frequent fliers? In *Proceedings of the first ACM conference on online social networks* (pp. 213–222). Boston: ACM.

Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2016, March). *Detection of promoted social media campaigns*. In tenth international AAAI conference on web and social media, March 2016, Cologne, Germany.

Gerbaudo, P. (2012). *Tweets and the streets: Social media and contemporary activism*. London: Pluto Press.

Hong, L., Doumith, A. S., & Davison, B. D. (2013, February). Co-factorization machines: Modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on web search and data mining* (pp. 557–566). Rome: ACM.

Howard, P. N., & Hussain, M. M. (2011). The role of digital media. *Journal of Democracy*, *22*(3), 35–48.

JafariAsbagh, M., Ferrara, E., Varol, O., Menczer, F., & Flammini, A. (2014). Clustering memes in social media streams. *Social Network Analysis and Mining*, *4*(1), 1–13.

Jones, K. S. (1973). Index term weighting. *Information Storage and Retrieval*, *9*(11), 619–633.

Kim, D., Jo, Y., Moon, I. C., & Oh, A. (2010, April). *Analysis of Twitter lists as a potential source for discovering latent characteristics of users*. ACM CHI workshop on microblogging, Daejeon, Republic of Korea .

Konda. (2014, June 5). *Public perception of the 'Gezi Protests': Who were the people at Gezi Park*. Retrieved from http://konda.com.tr/en/raporlar/KONDA_Gezi_Report.pdf

Lewis, S. C., Zamith, R., & Hermida, A. (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.

Lim, M. (2012). Clicks, cabs, and coffee houses: Social media and oppositional movements in Egypt, 2004–2011. *Journal of Communication*, *62*(2), 231–248.

Mathioudakis, M., & Koudas, N. (2010, June). Twittermonitor: Trend detection over the twitter stream. In *Proceedings of the 2010 ACM SIGMOD international conference on management of data* (pp. 1155–1158). Bloomington, IN: ACM.

Metzger, M., Penfold-Brown, D., Barberá, P., Bonneau, R., Jost, J., Nagler, J., & Tucker, J. (2014). *Dynamics of influence in online protest networks: Evidence from the 2013 Turkish protests*. Annual meeting of the Midwest Political Science Association, Chicago.

Oktem, K. (2013, June). Why Turkey's mainstream media chose to show penguins rather than protestors. *The Guardian*. Retrieved from http://www.theguardian.com/commentisfree/2013/jun/09/turkey-mainstream-media-penguins-protests

Perea, E. A., Cristancho, C., & Sabucedo, J. M. (2014). Mobilization through online social networks: The political protest of the indignados in Spain. *Information Communication and Society*, *17*(6), 750–764.

Poell, T. (2014). Social media and the transformation of activist communication: Exploring the social media ecology of the 2010 Toronto G20 protests. *Information, Communication, & Society*, *17*(6), 716–731.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web* (pp. 851–860). Raleigh, NC: ACM.

Segerberg, A. & Bennett, W. L. (2011). Social media and the organization of collective action: Using twitter to explore the ecologies of two climate change protests. *The Communication Review*, *14* (3), 197–215.
Stepanova, E. (2011). The role of information communication technologies in the 'Arab spring'. *Ponars Eurasia*, *15*, 1–6.
Subrahmanian, V. S., Azaria, A., Durst, S., Kagan, V., Galstyan, A., Lerman, K., … Menczer, F. (2016). The DARPA Twitter Bot challenge. *Computer*, *49*(6), 8–46.
Theocharis, Y., Lowe, W., van Deth, J. W., & García-Albacete, G. (2015). Using Twitter to mobilize protest action: Online mobilization patterns and action repertoires in the Occupy Wall Street, Indignados, and Aganaktismenoi movements. *Information, Communication & Society*, *18*(2), 202–220.
Tufekci, Z., & Wilson, C. (2012). Social media and the decision to participate in political protest: Observations from Tahir Square. *Journal of Communication*, *62*(2), 363–379.
Valenzuela, S., Arriagada, A., & Scherman, A. (2012). The social media basis of youth protest behavior: The case of Chile. *Journal of Communication*, *62*(2), 299–314.
Varol, O., Ferrara, E., Ogan, C., Menczer, C., & Flammini, A. (2014, June). Evolution of online user behavior during a social upheaval. In *WebSci'14 proceedings of the 2014 ACM conference on web science* (pp. 81–90). Bloomington, IN: ACM.
Verdegem, P., D'heer, E., & De Grove, F. (2015). *Social media in times of neoliberalism: Connective action or polarization of the public debate?* Montreal: IAMCR.
Weng, L., Menczer, F., & Ahn, Y. Y. (2013). Virality prediction and community structure in social networks. *Scientific Reports*, *3*. doi:10.1038/srep02522
Weng, L., Menczer, F., & Ahn, Y. Y. (2014). Predicting successful memes using network and community structure. In *Eighth International AAAI Conference on Weblogs and Social Media*, Ann Harbor.
Weng, L., Ratkiewicz, J., Perra, N., Gonçalves, B., Castillo, C., Bonchi, F., … Flammini, A. (2013, August). The role of information diffusion in the evolution of social networks. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 356–364). Raleigh, NC: ACM.
Wojcieszak, M., & Smith, B. (2014). Will politics be tweeted? New media use by Iranian youth in 2011. *New Media & Society*, *16*(1), 91–109.
Zhuo, X., Wellman, B., & Yu, J. (2011). Egypt: The first internet revolt? *Peace Magazine*, *27*(3), 6–10.