High Dimensional Network Analytics:

Mapping Topic Networks in Twitter Data During the Arab Spring

Kathleen M. Carley, Wei Wei and Kenneth Joseph

Abstract

Social change is often reflected in social talk. The ability to track who is talking about what, where and with whom, as well as changes in the topics of concern by region, may provide insight into emerging crises and provide guidance on how to mitigate other crises. Network analytics have proven successful at analyzing such data. However, such talk is increasingly carried out in social media at dramatically higher volumes than previously analyzed. A high-dimensional network approach for assessing this talk and identifying not just what is being talked amount, but the locality and change in that talk and the associated groups and their structure is presented. This approach is applied to data captured with respect to the Arab Spring. The results provide insight into the co-evolution of topics and groups across the region during this period of dramatic social change.

Network Analytic of Twitter Data for Mapping Topic Networks

Introduction

The wave of revolutions in the Arab world, commonly referred to as the Arab Spring, was a period of major social change. As protests and demonstrations broke out in country after country, questions arose as to what mechanisms supported the diffusion of ideas and actions, promoting or inhibiting violence, and thus enabling successful regime change. New communication technologies and social media were touted as critical to these revolutions. The belief in the power of the Internet was such that in some cases embattled leaders turned off access, e.g., Egypt and Syria [1]. In all cases, as these countries moved from a pre-revolutionary to a revolutionary state the "talk" changed. Where Wikileaks and sports were topics of interest prior to the onset of the protests, discussion moved towards issues such as liberation, government overthrow and insurgency once the revolution began. At the same time, groups formed and disbanded, and alliances among diverse actors altered the way they went about their activities.

Throughout the Arab Spring, discussion of the transition and issues potentially related to the transition, such as economic conditions, injustices, and civil rights were discussed in the traditional and social media. Various actors, purportedly, used these media to engage discussions to foment or counter rebellion. These media contain information about both the set of actors, the set of topics, and the connections among actors and topics. A geo-temporal assessment of this information should provide insight into the ways in which actors and topics coalesce and disperse during periods of social change. Our key concern is to understand the geo-temporal distribution of topics and groups, and the extent to which these are global or state specific, temporally invariant or transient. Social media data from the Arab Spring, specifically Twitter data, provides a corpus of interest ideal for studying the geo-temporal dynamics of social and topic networks.

Social network analysis (SNA) supports the understanding of groups using graph theoretic and statistical approaches for assessing the connections among actors. SNA has historically been used to understand how the structure of society, the patterns of connection among actors, influences behavior. The traditional social network analytic approach, however, is limited vis-a-vie its utility for understanding massive social change, particularly when the data source is media based. There are several critical limitations: 1) many of the metrics do not scale well to massive data such as those based on shortest path calculations; 2) social media sites often alter the network structure of the data e.g., Twitter does not provide the true retweet network but rather connects all retweets only to the original tweet; 3) geo-temporal factors are not accounted for or easily assessed; and 4) typical approaches use only one type of network such as the actor-to-actor network rather than the high dimensional network data available.

In contrast, dynamic network analysis (DNA) overcomes these limitations [2]. Herein, a DNA assessment of actors and topic networks through the Arab world over the course of the Arab Spring is conducted using Twitter data. Using a high-dimensional network representation, referred to as the meta-network, complex systems can be represented. We employ this representation to look at two specific questions. First, we study a meta-network of actors, topics, and the sub-networks of actor-actor, topic-topic, and actor-topic in a geo-temporal context. We then consider only the actor-actor network and study the evolution of groups within this network over time. We employ a combination of methods based on techniques from machine learning and statistical network analytics to understand results.

The networks of interest are derived from Twitter data collected for multiple countries over the course of the Arab Spring. Using this corpus, actors - the users, and topics – the critical concepts/hashtags discussed, and the networks connecting these are extracted per tweet. Temporal information and as possible geospatial information are also captured. The result is a set of large high dimensional geo-temporal networks. These network data are "big" due to their high dimensionality, and the large number of time periods.

Arab Spring

Beginning in December 2010 a large number of protests, riots, and demonstrations began in country after country in the Middle East. These events are generally referred to as "the Arab Spring." In some cases, e.g., Libya, these protests turned into an insurgency and civil war. In many cases, e.g., Egypt, the current leader was over-thrown. When the Arab Spring ended, or whether it has ended, is a point of contention.

One of the key elements of the Arab Spring is that it occurred in a region fraught with conflict, revolution [3] and change [4] where the political dialogue since at least the 1920s has been one of identity [5]. Prior to the onset of the revolutionary protests, there was a rise in the number of young educated people with low job prospects, increased urbanization, changing economic basis, and changes in the presence of and the integration of terror groups into the local communities. Numerous topics were emerging as points of dissension some such as polygamy were associated with Sharia law, while others such as soccer were associated with general past-times.

Social media played a critical role in the Arab Spring [6]–[8], social media, debates about freedom, civil liberty and democracy raged. While not everyone in these countries used Twitter, it nevertheless is thought to provide a good window into the digital conversation. However, the data needs to be used with caution as the users are both within and outside the affected countries and the dialogue is carried in English and in Arabic [9] with some, albeit limited, overlap. Moreover, Twitter appears to be used differently by, and different memes appear to be preferred by protesters at the site and by remote observers [10].

The Arab Spring, and the Twitter usage associated with this event presents an ideal venue for studying at scale the geo-temporal distribution of, and co-evolution of, topics and groups from a network perspective. These prior studies suggest that some topics will be more local and others global, and that there may be greater locality in the topics expressed in Arabic. These prior studies also suggest that the topics will change over the course of the events. We further ask, to what extent are these changes geographically as well as temporally local?

General Background

The number of studies of Twitter data has exploded in recent years. Key reasons include the relative ease of collection, the fact that the data is held under creative common license, and the interest in large scale networks. These studies demonstrate that such data can provide early indications of change. Twitter ties are generally predicted by being within the same metropolitan region, being nearby, sharing a common border, sharing the same language, and the frequency of airline flights between the sites [11]. While strong social ties thus exist on Twitter and can be algorithmically uncovered with reasonable accuracy [12], Twitter networks are not necessarily reflective of actual social networks [13].

Despite the breadth of study, the movement of ideas and groups as reflected in Twitter is still poorly understood. To understand groups, retweet, mentions and reply networks are often extracted from the meta-data and then assessed. In many cases these networks are fairly sparse. Understanding the movement of ideas is less straightforward. Media studies, whether using traditional or social media, often turn to sentiment analysis to interpret the flow of information. Recent studies have shown that it is possible, to make "predictions" albeit retrospectively. For example, Leetaru [14], using simple sentiment (positive/negative) and geo-location was able to show that the level of sentiment expressed in traditional media in Egypt went to an all-time low (considering last 30 years) prior to Mubarack's resignation. More detailed content analytics that look at the key concepts have also been used to provide general predictions of revolution and violence using diffusion modeling [15].

From the perspective of this study, the point here is that most studies of media focus only on identifying sentiment or identifying what are the most frequently used concepts. In contrast, our concern is with topics where a topic can be thought of as a general idea or issue around which a number of diverse words and sentiments might coalesce. Text mining algorithms are generally used to extract topics from texts – the most popular examples are latent Dirichlet allocation [16] (LDA) and latent semantic analysis [17] (LSA). Such algorithms generate a set of "latent" topics for a given text corpora, where each word is associated probabilistically with each topic.

Data

For the purposes of the present work, we have extracted information on the time, textual content, geo-coordinates and social interactions (retweets, mentions, and replies) enclosed within a corpora of tweets related to the Arab Spring. The data collected consists of approximately 95M tweets gathered from two sources from April 2009 to November 2013. The first source was collected by tracking a manually curated set of keywords, users and geo-boxes related to the Arab Spring using the Twitter Streaming API, which returns a maximum of around 1% of the full set of tweets at any given time[1]. Parameters used to search the Streaming API focused mostly on events surrounding Egypt, Libya, Syria, Tunisia and Yemen, though certain parameters did apply to the entire region associated with the Arab Spring. The second way in which data about tweets was obtained was from an outside researcher who provided us with geo-tagged tweets from a 10% sample of the full set of tweets during this same time period[2]. Information about geo-tagged tweets was obtained only for the set of countries studied by [15] on prior work for the Arab Spring.

A high level statistical overview of this data by country is shown in Figure 1. Information on Tweets (which includes retweets), Users (which are those who tweeted), and Terms (which are the isolated sets of three or more characters including hashtags in the textual content of the tweet). The data includes information about both Arabic and non-Arabic Tweets. In many countries, the statistical profile of the Arabic and non-Arabic tweets are similar.

Figure 1 presents boxplots for three statistics for each country. In Figure 1 and in the analyses below, a tweet was considered to be associated with a country if it 1) was geo-tagged and sent from within that nation's borders; 2) contained the name of the country in English or in Arabic; or 3) contained the name of any of the three largest cities within that country in English. We included the three largest cities in determining the countries associated with a given tweet after noting that discussions around certain important events, such as the Tahrir Square Protests, only mentioned a city (Cairo) as opposed to the encompassing nation. Note that via this methodology, a tweet could be associated with more than one country and thus no straightforward statistical comparisons can be made across countries comparing the values in Figure 1. Also note that all plots in Figure 1 are log-scaled.

---

[1] For more details on the collection of part of this data, we refer the reader to [18]. For details on the Twitter Streaming API, see [19]. This data was collected by our Minerva research team.

[2] These tweets were collected through the Language Technology Institute at Carnegie Mellon University under the direction of Brendan O'Connor under an agreement that allowed all CMU researchers to make use of this data.

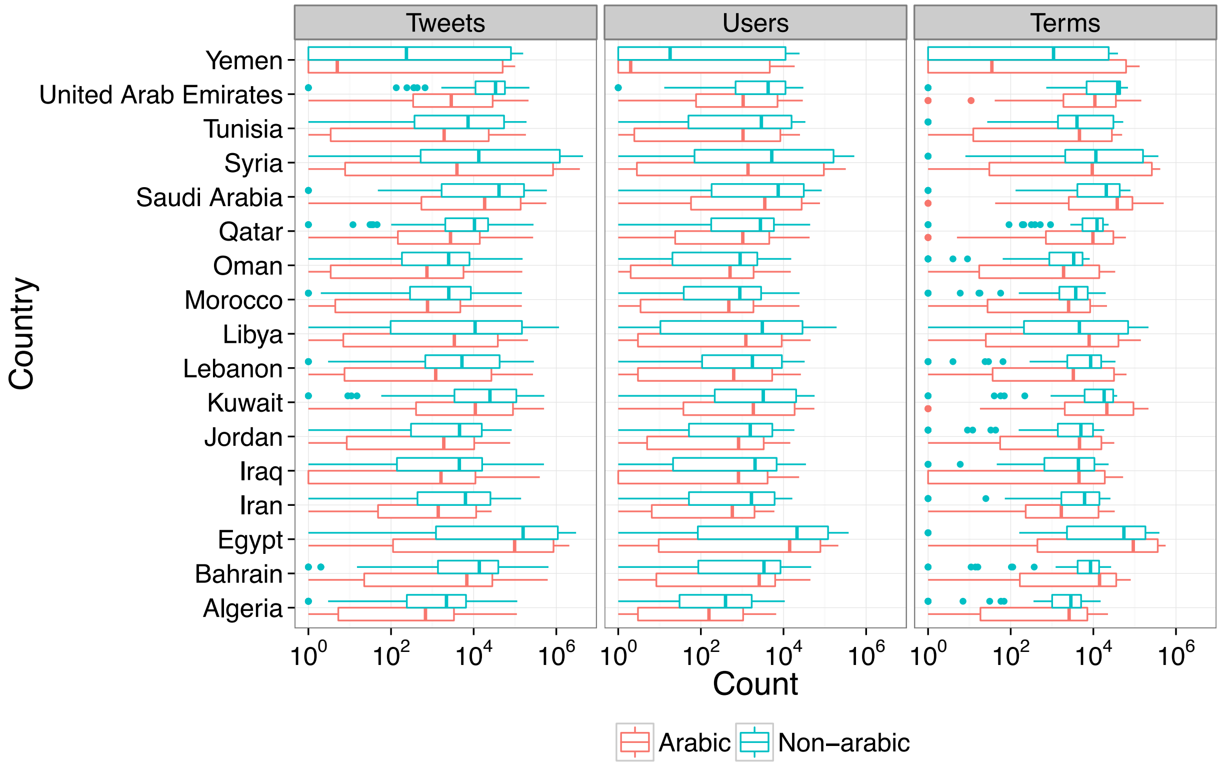


Figure 1 Mean and Standard Deviations of Tweets, Users and Terms from Different Countries

The leftmost plot in Figure 1 presents of the number of tweets for each country studied for both the set of all tweets and the set of tweets that contained Arabic terms. We see that on average, countries saw in the tens of hundreds to tens of thousands of tweets per month. However, our dataset contains months where Egypt, Syria and Libya each saw several hundred thousand tweets in a single month. We also see evidence that tweets containing Arabic accounted for a non-negligible portion, and in several cases the majority, of all tweets within a particular country.

The middle subfigure of Figure 1 shows the distribution of the number of unique users for each country for both all tweets and Arabic tweets, where a user is included in the Arabic count if that user ever used an Arabic term. Again, we see that the number of users who tweeted in Arabic in our dataset approached the number of non-Arabic speakers in each nation. This result furthers the point that the Arabic speaking population played a prominent role in the discussion of the Arab Spring on Twitter, and thus, as implied by [18] that English-only analyses of the events that transpired may be a biased representation of the discussions occurring online.

The final subplot in Figure 1 displays the number of non-Arabic and the number of Arabic *terms* in the data. A term was any string in any tweet with greater than three characters. A general stop-word list was applied to remove common terms from the topic list, and tokenization was performed using the widely accepted (e.g. [20]) tokenizer from [21]. This final subfigure shows that in some cases, the number of Arabic terms was greater than the number of English terms used. While this may be in part due to the fact that tokenization of Arabic terminology is unique from English and thus may provide slightly inflated values, the finding serves as additional validation that Arabic and non-Arabic tweets and discussions were both prominent, and that it was not simply users utilizing a few choice Arabic words to, for example, emphasize an Arabic identity.

In addition to considering these statistics, it is also interesting to examine the proportion of tweets about each country that are geo-tagged. This is particularly important in the present work, as our geo-temporal analysis of topics is run on only geo-tagged Tweets. Figure 2 shows boxplots of the percentage of tweets that were geo-tagged for each country in each month. While there are several outliers, these tended to be months where data was sparse and thus percentage estimates were highly variable. In sum, the figure shows that on average geo-tagged tweets were somewhere between 5-15% of the tweets in any given month across all countries in the corpus used.

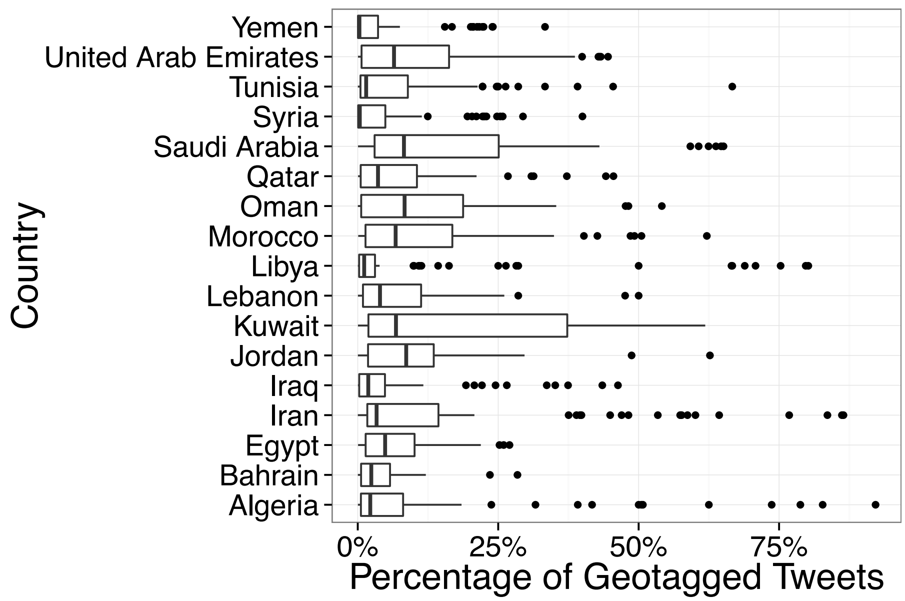


Figure 2 - Number of Geo-tagged Tweets in each country

While our dataset represents a large portion of tweets related to the Arab Spring, it is important to note that this overview provides only a description of our dataset and thus should not be taken as a definitive overview of what the full collection of tweets relating to the Arab Spring looked like. As noted by [22], such considerations are important in cases where samples of tweets are already biased by the search criterion used. In particular, while we believe that results on our dataset may in many cases generalize to the overall sentiment that surrounded the Arab Spring, our emphasis on geo-tagged data from Arabic countries may suggest our results over-represent the general level of discussion of the Arab Spring that occurred globally in Arabic over the allotted time period.

The Social Pulse: Geo-Temporal Trends in Twitter Topics and Users

Methodology

To garner a better understanding of the interrelationships between actors, the topics they discussed and spatial location, we utilize LDA. In order to do this, we first aggregate all the tweet text of by user and treat this aggregated text as a single document, an approach has been adopted by several recent works on Twitter [23], [24]. Given a set of users and the terms associated with them, LDA will extract a number of "latent" topics based on a Bayesian probabilistic model, which assumes that each user discusses a subset of all possible topics. In the model, each latent topic is described by some subset of all terms that tend to be used frequently by the same user. Users are then evaluated using the mined topics, giving an indication of the relevance between each topics and the given user. Note that the number of topics in LDA is specified by the researcher- In the present work, we estimate the model with 100 topics, noting that larger numbers of topics tend to fair better in recovering important latent information [25].

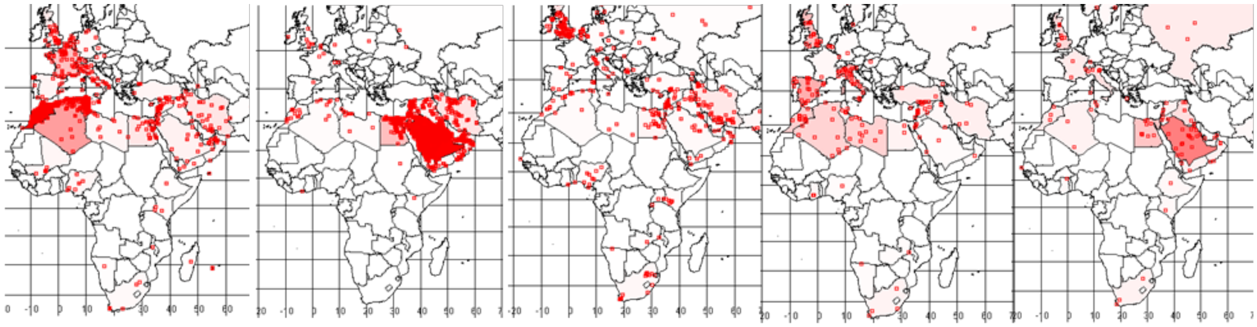
While LDA allows us to associate users with topics, we are also interested in two additional pieces of information, both of which can be inferred using the posterior distribution given by the model. First, we are interested in associating particular tweets (as opposed to users) with each topic. Under the (reasonable) assumption that each tweet is concerned with only one topic, we can determine the topic of a given tweet by selecting the topic that the terms in the tweet are best associated with. Second, we are interested in connecting users by the similarity of their topical discussion. This leads to the formation of a "co-topic" network, which is formed by comparing the topic scores between two users. Each user in the data is associated with a topic score vector. The mod score (explained below) between an arbitrary pair of users in the data set is evaluated. If that similarity is larger than a preset threshold, a link will be generated between those two users with the mod score as the tie strength. In our network, we have eliminated self-links so that no node is pointing to itself.

After running LDA on our data, we first explored how the topics clustered in different geographical locations. Here, we analyze in more detail the five top topics uncovered by the model, considering the terms that best represented these topics and the spatial distribution of tweets relevant to the topic. The top five topics were determined by selecting the five topics which had the highest likelihood of occurring across all users. Experimental results showed that topics present high locality and differed significantly from country to country.

Topic Overview

Figure 3 presents information on the locality of 5 topics. For each topic the five most representative terms for each of the top five topics uncovered in our data. The top terms in each topic were either entirely in English or entirely in Arabic, thus we give the language of the topic next to its title. For English terms, this meant the five terms that had the highest likelihood in the posterior distribution of the topic. For Arabic terms (only the English translation is provided for clarity) we show the five terms with the highest likelihood in the posterior that we could also satisfactorily translate using Google Translate. Figure 3 shows that, at least amongst the top five terms for these topics, the foci of discussion were unique. For example, topic 46 talks about Beirut (the capital of Lebanon) Jordan and America, which were all nations only tangentially involved in the events of the Arab Spring; and, topic 98 is focused on sports and possible sports medicine. In contrast, Topic 91 consists of Arabic words directly related to the Arab Spring region and includes the name Ali and the term tyrant, most likely referring to former Tunisian President Zine El Abidine Ben Ali, ousted during the Arab Spring. Quite interestingly, in contrast to the negative sentiment in Topic 91, we instead find a set of positive words such as good, peace, rose and possible characterizing Topic 92. Thus, our results suggest that notions of peace and tyranny tended to come from distinct segments of the Twittersphere, a claim that would be interesting to substantiate further in future work.

Figure 3 also presents a geo-visualization of all the tweets related to each specific topic. Here we see that certain topics present strong localities: topic 40 which talks about African, families generally concentrated in Morocco, Algeria, Tunisia and south west Europe. Topic 46 which talks about American activity in Beirut and Jordan in particular, and the middle east more generally, is concentrated in Saudi Arabia and Egypt. The rest of the topics have a greater span across the entirety of the Arab world. The examination of these plots suggests that, in accord with what we would expect, general concepts like “peace” and “tyranny” spread throughout the Arab world, while local topics (e.g. those mentioning a specific location) tended to stay within the confines of certain spatial regions.



Topic 40 (English)	Topic 46 (English)	Topic 91 (Arabic)	Topic 92 (Arabic)	Topic 98 (English)
social	American	Arab	good	follow
happening	east	Spring	peace	sporting
african	Beirut	Gulf	roses	mosh
families	Jordan	Ali	special	coast
meridien	middle	tyrant	possible	liam
Morocco Algeria Tunisia	***Saudi Arabia Egypt***	***Iran Jordan***	***Algeria Libya Morocco Syria***	***Suadi Arabia Egypt***

Figure 3 Geo Visualizations of top 5 topics, language, and key associated terms.

Over Time analysis

Apart from the geo-spatial distributions of topics, the temporal distribution is also important. In this analysis, we aggregated the topic scores of each tweet assigned by LDA algorithm and picked up only the topic with highest aggregated score over the whole data set and generated a global top topic. Table 1 shows the top topic and associated terms in our data set calculated by month. Note that the top topic moved from cry looking for solutions to wonderment over the revolution, to more specific discussions of key issues – the role of the Americans and the concern with Morsi (who was removed from office in July 2013). It is not clear what topic 74 refers to, although one possibility is that it the associated tweets may contain excerpts from a song.

Table 1. Temporal dynamics of topics

Time Period	10/2010 1/2011	2/2011	3/2011 4/2011	5/2011 9/2012	10/2012 11/2013
Topic	94 (Arabic)	74 (Arabic)	41 (Arabic)	46 (English)	62 (Arabic)
Term	people	UaC	Arabs	american	Egypt
Term	god	Elly	people	east	Morsi
Term	life	Quaoui	country	information	Head
Term	solutions	Pak	beloved	Beirut	people
Term	even	OiYai	what is happening	Jordan	President

We can see that over time, the topics changes from 2010 to 2013, generally in a way that is related to the political movement in the areas where the tweets are being sent out. For example, in 2010, the most prevalent topic is are prayers for solutions. This corresponds to the beginning of the Arab Spring movement, which spreads over the whole Arab world that involves revolutionary wave of demonstrations and protests. This topic is not localized. In the beginning of 2011, topic 41 came to the fore – and within a cry asking what was happening to their countries. This is followed by a year-long debate on the role of Amrican's in the middle east – see the associated keywords directly related to key locations such as Beirut and Jordan of Arab Spring. Interestingly, this is the only time in which an English topic dominated the discussion. At the end of 2012, tweet topics moved to political events that are related to the reign and overthrow of the fifth president of Egypt, Morsi. This trend lasts until the end of the data set.

Several important themes underlie the over time analysis. First, the dominant discussion topic moved over time from general topics – to specific topics. Second, the dominant discussion topics moved from being geographically broad to geographically narrow. Third, the dominant discussion topics moved from being a-political to political. Thus the topics from those to which there was general universal accord and that while specific to the Arab Spring was geographically relevant to the entire region to those that were politically charged and most relevant to a small set of countries. This indicates the qualitatively assumed but, as to our knowledge not quantitatively shown, assumption that the Arab Spring region moved over the course of the last few years to a more political focus that surrounded the events of the Arab Spring. Second, we note that the dominant form of discussion was in English for only the period around the most intense actions surrounding the Arab Spring. Naturally, this suggests that the English-speaking world was interested during the height of the conflict but rapidly moved to other topics, while the Arabic speaking world was (and still may be) predominantly focused on the political events transpiring in the region.

Characterization of User-Topic Similarity Network

Next we consider the relations of users to the geo-temporal distribution of topics. This required constructing networks of users based on whether or not they both tweeted on a topic. This network is based off the user by topic network where the topics were those previously identified and the links were the number of tweets by that user associated with that topic. Although most topics tend to be associated with a single topic, most users are associated with most topics. The resultant user by topic network is then used to define links between users based on shared topics or similarity in topic usage. Most approaches to generating a user to user network based on tweeting about the same topics generates networks that are too dense for most network algorithms to run metrics on in a reasonable amount of time. Since most users are associated with most topics the simple number of topic shared tends to result in a very dense, and non-discriminating network that under-represents the focal interest of the user. Therefore, instead of shared topic counts we use a similarity index that weights the topics by focus. It is important to note that cosine similarity is the generally accepted solution to this problem; however, it is too slow for the size of our data and which makes it unrealistic for generating the user-topic similarity network in an acceptable running time. We used an alternative more efficient method to calculate the similarity between the topic usage vectors based on mod scores. First we define a vector as the real valued score for that user on all topics. Given two vectors $v_1 \in \mathfrak{R}^K$ and $v_2 \in \mathfrak{R}^K$ that are both real valued vectors in the k dimensional space, the *mod score* between those two vectors is defined as:

$$Sim_{i,j} = \frac{\min(|v_1|, |v_2|)}{\max(|v_1|, |v_2|)}$$

Since both $|v_1|$ and $|v_2|$ can be calculated in advance before the generation of the network, the magnitude similarity can be calculated fairly efficient. We then define a link between two users to be 1 if the magnitude of the similarity of the two users is larger than or equal to 0.99, else 0. As a result, we obtain an undirected, binary user topic similarity. This network can be interpreted as showing those users who have a strong focus on the same topics. For each country, we generate a separate user topic similarity based on all the tweets associated with that country.

For each country, for each use topic network, standard network level statistics are calculated – see Table 2. First, we consider the number of non-isolated nodes in the network, which is the number of unique users that that have strong topical similarity to at least one other user. Note that this is only a very small subset of the data set since the high threshold filtered out the majority of the users. Second, we consider the number of edges in the user topic similarity network which is the number of dyads that have strongly similar topic foci. Finally, we look at the density of the user topic similarity network after the isolates (those users who were not strongly tied to any other user) are removed. This provides insight into the overall structure of connectivity among the users. For contrast we also show these same statistics for just those tweets that are non-Arabic.

We can see that the majority of the countries have a network density of roughly 0.01, which indicates that only 1 out of 100 users in this strong similarity network share similar topical distributions, and that on average, each user has a high degree of topic similarity with about 1% of the other users in the network. Some countries, such as Tunisia, have significantly lower densities, indicating the potential for a less homogeneity of topics in the discussion in these areas. Of all countries, Yemen has the highest network density, which indicates that more of the users in that country tend to discuss similar topics on twitter. The number of nodes in the networks indicated that there are far more active twitters talking about dominate topics in Egypt, Saudi Arabia, Syria and UAE than other countries. Among these countries, Saudi Arabia has an especially high number of links in the network because of the number of twitter users talking about the same topics.

Table 2. Network Statistics of the user topic similarity network by country in the original data set

	All Tweets			Non Arabic Tweets		
Country	Non Isolates	Edges	Density No Isolates	Non Isolates	Edges	Density No Isolates
Bahrain	4559	206642	0.012	8698	149612	0.004
Qatar	6948	378262	0.016	10721	230981	0.004
Iraq	1852	42257	0.024	3295	23008	0.004
Iran	975	6304	0.013	1344	4998	0.006
Libya	4394	110910	0.011	5259	88827	0.006
Algeria	780	5913	0.019	955	5134	0.011
Egypt	42060	9490034	0.011	62653	7964548	0.004
Kuwait	19713	6087116	0.031	45955	4273476	0.004
Lebanon	6687	226560	0.010	7722	171573	0.006
Morocco	5612	258507	0.016	6689	157733	0.007
Jordan	3711	79486	0.012	4887	61438	0.005
Saudi Arabia	33663	35921282	0.063	136543	46843301	0.005
Oman	2193	45820	0.019	4491	71297	0.007
Syria	40625	8603652	0.010	53350	7042616	0.005
Yemen	1109	84280	0.137	6000	131767	0.007
United Arab Emirates	24417	3542578	0.012	33448	3155592	0.006
Tunisia	3692	63728	0.009	4253	49105	0.005

On the right in Table 2, the network statistics of for the non-Arabic user-topic similarity network is shown. More users have a high topic similarity with at least one other user in this non-Arabic discussion network; however, these users are on average connected to fewer other users. That is, in general, the density of the non-Arabic user topic similarity networks are lower than the overall user topic similarity network. This is because people tweeting not in Arabic tend to focus on a wider variety of and different topics. Most countries that have high node count in the overall data set also have a high node count in the Non-Arabic only data set, which is not surprising since those countries have a large number of Twitter users in general.

The differences in densities for the overall data and just the non-Arabic data has some interesting implications. Consider Bahrain. Overall, the density implies that among those users who are strongly tied to at least one other, the average user is strongly tied to about 55 others, but in just the non-Arabic realm only to about 35 others. Whereas, in Saudi Arabia the values are 2,121 users overall and 683 in the non-Arabic. This also implies that the in the Arabic tweeting part of this network there is substantially more homogeneity in shared topics and more of the Arabic tweeting actors have higher similarity to each other in their topical focus. This could indicate some transference of topics between Arabic and non-Arabic speakers. However, part of the difference is due to the fact that the 100 topics, when assessed overall are much broader and less discriminatory then the 100 topics for just the non-Arabic tweet content.

Social Interaction Overview: The Reply Network

We now consider the social relations among the actors in our dataset. While social relationships in Twitter data require a degree of nuance in interpretation due to the technological affordances of the media [22], [26], if one is careful insights can nonetheless be gained. In general, most network analytics focus on either the retweet, mentions, or the reply network. We focus here on the reply network. The reply network can be identified whether the sender hit reply when sending the tweet.

This network changes dramatically over time as new users join Twitter, and as users move between topics and so groups. In Table 3 summary statistics describing the reply network are shown. As there is substantial variation by month the results shown are the averages across the months. In other words, the reply network was constructed for each month for each country and then the months averaged by country. Examining this information we see that the sheer volume of users replying to others tweets, and the density of the tweet network does not correlate with revolutionary activity. There is high country variability. For example, Iran shows a small dense community with very fast information flows (low characteristic path length). It is possible this network is dominated by expatriates. Whereas for Saudi Arabia there is a larger, sparse, community with more distinct clusters with users often needing 5 to 6 steps to move information.

Table 3. Network Statistics of the reply network by country in the original data set

Country	*Mean Nodes*	*Mean Edges*	*Mean Density*	*Mean Clustering Coefficient*	*Mean Characteristic Path Length*
Bahrain	804.212	690.115	0.010	0.017	5.749
Qatar	1120.818	959.891	0.010	0.008	5.173
Iraq	306.852	214.111	0.034	0.001	2.305
Iran	253.945	204.182	0.029	0.003	2.924
Libya	1114.686	1158.686	0.053	0.002	2.719
Algeria	132.714	100.653	0.020	0.006	2.015

Egypt	6120.964	6712.636	0.008	0.004	5.209
Kuwait	4745.909	4824.418	0.013	0.014	5.577
Lebanon	693.731	568.923	0.019	0.005	3.982
Morocco	380.472	324.679	0.034	0.003	2.927
Jordan	456.824	388.569	0.016	0.010	3.553
Saudi Arabia	9025.873	9460.709	0.004	0.007	5.305
Oman	316.462	270.346	0.030	0.009	3.348
Syria	3712.906	3717.962	0.016	0.003	3.246
Yemen	502.628	444.581	0.049	0.000	2.493
United Arab Emirates	3785.018	3677.291	0.002	0.014	5.888
Tunisia	395.224	322.061	0.008	0.005	3.531

Characterization of Group Structure

The reply network is not a uniform or random network of connections. Rather, it has a very sparse multi-component structure that changes over time. Figure 4 displays various metrics calculated on the reply networks over time, where each point on the line represents the network for a given month. In each subplot, there are three lines – the red represents results for tweets containing one or more Arabic terms, the blue for tweets that did not contain any Arabic, and the green for the full dataset. The top row of plots, from the left, displays the number of nodes in the network, the number of edges, and the size of the largest strongly connected component (LSCC), defined as the number of nodes in the largest portion of the directed reply graph where each node is reachable (via following directed lines) from each other node in that portion. The bottom row displays (from the left) the size of the largest *weakly* connected component (LWCC), which considers connectivity assuming that the reply network is an *un*directed network, the number of weak components and the percentage of actors in the LWCC.

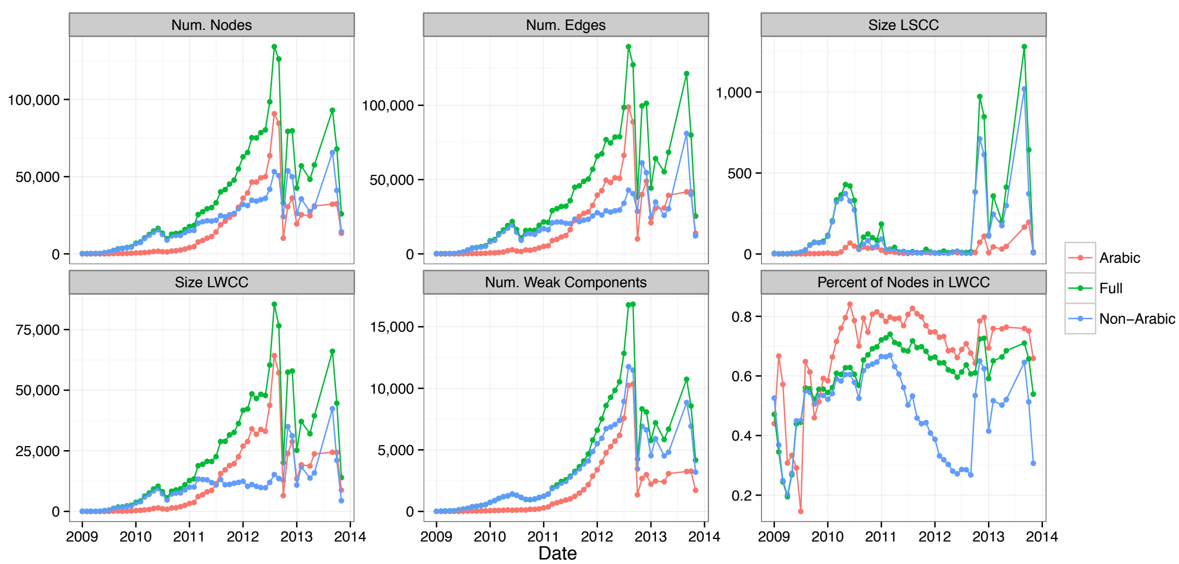


Figure 4. Temporal change in the reply network

From Figure 4, several points of interest can be ascertained. First, as we would expect, as the number of nodes increases the size of the LWCC and LSCC, as well as the number of weak components, steadily increases. Interestingly, however, the percent of nodes in the LWCC is highest in early 2011, when things were just starting to flair up in Egypt and Libya. This suggests that actors may have been more invested in obtaining new information from Twitter during this time as opposed to from traditional media sources. Second, we observe that size of the LSCC is much, much smaller than the size of the LWCC. Because of the one-way directionality of interaction on Twitter, this is to be somewhat expected. However, it also suggests that there may have been little reciprocity in the core of the network, where certain actors were being replied to but were not replying to others who directed communication at them. This may simply be a result of on the way in which the reply network is constructed (recall that a reply implies a response and we do not have information on who was mentioned in the initial tweet), but it also suggests that the use of gatekeeping [8] on Twitter strongly structures the resulting network.

Finally, and perhaps most interestingly, we see that while once the Arabic component of the reply network reached a high proportion of nodes in the LWCC this proportion stayed above approximately 60%, the number of nodes in the LWCC of the all-English tweets dipped significantly through 2012-2013. This observation matches the intuition proposed in the sections above that Arabic users were much more invested in the events of the Arab Spring throughout the past four years, while English-speaking (or at least non-Arabic speaking) users tended only to be drawn in to the discussion at dramatic turns in the events. This intuition can be further qualified by suggesting that not only were Arabic speakers more focused on the discussion, they also were more engaged in networked discussions with each other. This point is strengthened by noting that in the user topic similarity network the non-Arabic tweeters show less strong connectivity to each other than do the users overall and those tweeting in Arabic.

Of course, just being together in the same component does not necessarily mean that these actors were interacting with each other. Indeed, within a single network component, there are still groups of actors who discuss particular topics and interact almost exclusively by themselves. To determine the extent to which this is true ideally one would correlate the various networks. This will be done in future research. For now, we take a faster approach to assess the extent to which there are these clumps of users and topics linked together, by taking the LWCC and running a network grouping algorithm on it. Figure 5 presents various metrics from the network grouping that results from taking the LWCC for each month and using the Louvain clustering method on it [27]. From left to right, the subplots depict the modularity [28] of the graph, which often is used as a measure of the quality of the clustering (i.e. the degree to which the graph is separable), the number of groups into which the clustering split the graph, the mean size of those groups and finally, the percentage of actors in the largest found group.

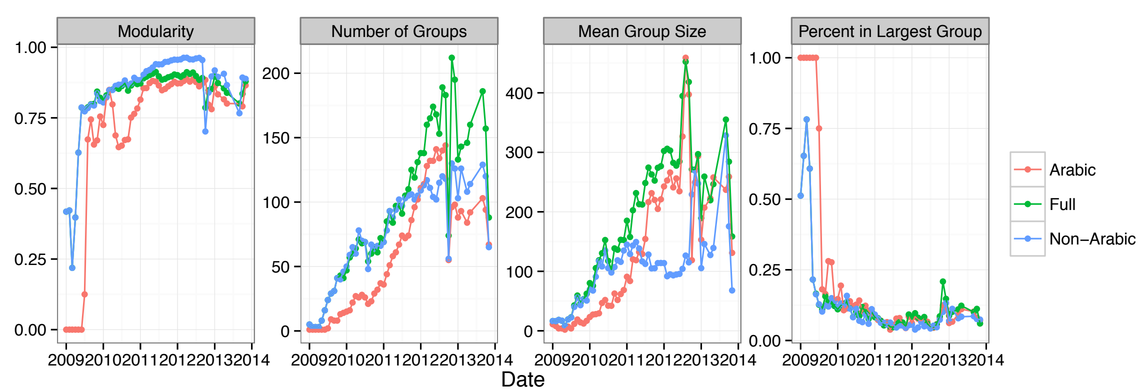


Figure 5. Temporal change in groups in the Twitter reply network

As we can see from Figure 5, modularity was very high across all months, thus the network was reasonably separable into groups even within the LWCC. Second, the number of groups and the average size of each group increased steadily over time, indicating that both more communities were being added to the network and that existing communities were growing. However, the size of the largest connected component stayed relatively stable, suggesting that while communities increased in size and number, there never became a global social community that infiltrated across the Arab Spring. Thus, one would imagine that brokers of information across communities existed and thus that there existed a select number of individuals that may have shaped interactions across groups.

Key Actors

The final question we address is the typical network question – who are the key actors? For the reply network, the top 6 users in total degree centrality are shown. These are users who reply or are replied to the most. These users are distinct by country suggesting that the country networks may have little interconnection. These users are not newsagencies suggesting that though newsagencies are a dominant presence in the Twittersphere, they are not central to the sub-groups trying to build consensus. Rather, the interaction is done at the grass roots level among general users. Some of these users, however, representing the extant government e.g. in Syria one of the most central users in the replay network is SyriaParliament or are freelance journalists such as DubaiWriter. In some sense, this examination of key actors raises more questions than it addresses. Most of these users are connected to most of the topics and the influence of these users vis-a-vie the topics is unclear. Future work should consider other social interaction relations such as the retweet and mentions network, and consider the relative standing of verified, news, and government users. The relation of these users to the topics and the relevant opinion leader for each topic identified.

Table 3. Network Statistics of the reply network by country in the original data set

Country	*1*	*2*	*3*	*4*	*5*
Bahrain	sasbahrain	YasiQannati	Farahfraidoon	bucheeri	SaroooLi
Qatar	Gadgod_	septboog	HEYitsSAL	JamesBryanBG	salkhulaifi
Iraq	yarab14	erdgnhsn	GeorgetteA	cerenationnext	gabitamatos
Iran	matuzalem	Mehrdad	tiagodvaz	dsantamaria	strasboorg
Libya	MaybeLaser	bulltas	Ben_Mussa	FairuzJumain	Sniggah
Algeria	Boubled	Hakim_3i	AmanIrh	Falqallaf	HuskyDaz
Egypt	Betsy_Mo	Hazem_Azim	Monasosh	OFree_zyIV	ZOGHBYZO
Kuwait	BuBarrak	DCiawy	iL3uBiD	HassanALSherazi	m7amdalnasser
Lebanon	iJoePopSlap	myrrnzz	sam_lb	AbirGhattas	Arabear
Morocco	HakimKhadija	neishatorres	yucefj	ravfm	rajk971
Jordan	pminttt	samihtoukan	Ibrahimmbl	h_alkhafaji	OmarBiltaji
Saudi Arabia	faisalmeshari	indiesaudi	MohamadAlarefe	Sara_wolf	battalalgoos
Oman	bijoyjacobk	raideraid	MusaabK	Degoat82	samiasmi
Syria	nasermaya	SyriaParliament	monsternotfan	SubhanAksa	resifahma
Yemen	bimbie07	Cyndaquillian	mussoO_	renytacitra	Bilal_ALhamzee
United Arab Emirates	wildpeeta	AskAli	ylootah	binmugahid	DubaiWriter
Tunisia	archcindymonica	iheb911	kcyam5	NaymaMC	Raahma_

Discussion

On-line media, and social media in particular, generates a wealth of geo-temporal data that can be used to gain insight into the pulse of a population. Extracting meaning from this data, and using this data to answer research and policy questions can be a daunting task (see Carley, forthcoming). Data collection, cleaning, translation, present challenges over and above analysis and visualization. We focused on what can be learned from social media, after the collection, cleaning and translation using a network approach. The data was segmented by country and time period and networks connecting users and topics were developed for each country and time period. We used a meta-network strategy in which users were connected to each other in a retweet network, a replies to network, and a similarity in topic usage network. In addition we have a user by topic network and a co-topic (i.e. topic by topic) network. Many additional analyses can be done with this encoded data – such as determining the extent to which the replies network predicts the similarity network and determining the paths by which topics change over time and how that relates to the changes in opinion leaders.

On the surface, the data analyzed is big, over 90m tweets. It covers a large range, 15 countries. And it covers a multi-year time span. However, there are limitations to this data. First, it is not a comprehensive account of all twitter activity in these countries during this time frame. In some countries, where there is low twitter usage, it may be close. Second, the data that is not geo-tagged may be coming from outside of the middle east. Thus, it is possible that the topics identified are mixing internal and external concern. These issues should be examined more in the future.

We recognize that big data increases statistical confidence, not accuracy (Silver, 2012). Accurate prediction with big data requires triangulation and the identification of patterns; thus, future work should use multiple types of data and triangulation techniques to generate predictions. For social media, a fair amount of triangulation can be afforded by doing broader international studies such as we did for the Arab Spring by comparing countries and comparing Arabic versus non Arabic data. In general, such comparative work will need the support of data publishers like Twitter who control access to the wealth of data. An alternative form of triangulation is possible by comparing and contrasting results from multiple media. This, future work should compare, at scale, distinct sources such as social media and traditional media.

The analyses that were run, were all relatively scalable for big data. This means that data processing short cuts were sometimes used which may have impacted the results. For example, topic identification was done at the term level which means that n-grams were segmented. A good example here is middle east which appears as separate terms "middle" and "east" in topics, rather than as a single concept. This speeded analysis, increased the number of terms and decreased interpretability. Future work should add common n-gram detectors. Another simplification was that translation was done after analysis and at the term level. A good example here is the set of top terms such as Elly associated with topic 74. This approach speeded analysis, but decreased interpretability. Future work should consider alternative translation, or analysis without translation, options such as using LSA.

The analyses described increased in complexity, and scalability. As we moved from one level to the next, the analytics took longer, but the number of results and the operational utility and research insight afforded by those results increased. The first analyses focused on counts. Counts are relatively simple and fast with big data. Unfortunately, counts provide little insight. The most one could learn here was where and when Arabic was more dominant, and changes in the sheer volume of communication. The second level of analysis focused on the clusters in the data. This was done through topic identification. This led to additional insight concerning what was being talked about and how it changed geo-temporally. The third level of analysis focused on the networks themselves. This led to additional insight about major sources of influence and geo-temporal change in that influence. As we move through these levels of analysis new insights possible, but the scalability of the process somewhat decreases. The issue is not that the network metrics don't scale well – indeed many of them scale as N. Rather, the issue is that the data as you move from level one to three additional data processing is needed to create the right data structures. Creating the data structures and storing them is, in and of itself, at this point, a time consuming process. Tools that facilitate network construction and the associated data cleaning would, and automated workflows would streamline this process.

As we move to the future there are a number of additional challenges that must be addressed when one is interested in network analytics and big data. A general discussion of these challenges appears in National Research Council Report (2013) and a more detailed review with specific relevance to crisis management and social media appears in Carley (forthcoming). For data such as the Arab Spring Twitter data, some of the major challenges we have encountered include:

- Understanding how the data collection filters bias the results.
- Inferring location for non geo-tagged data.
- Improving the scalability of statistical network tools, such as MRQAP for regression on networks.
- Incremental and approximation techniques for path-based network metrics
- Automated techniques for network extraction.

Conclusion

Social media is increasingly becoming a major source of information for populations. However, the grass-roots nature of social media is changing. The majority of news agencies, e.g., BBC, CNN and al-Jazeera use Twitter and Facebook to spread breaking news. Social media is also a major outlet for citizens to express their concerns. For example, in the recent Benghazi consulate attack, while the majority of tweets were from individuals, the top "tweeters" were news agencies and the Libyan Youth group (Carley et al, 2013). Within the tweet network individuals and news-agencies play different roles and have different geo-temporal tags. Who follows whom, the retweet network, the cyber-norms, the use of hashtags, and incorporated videos or images appear to be different for corporate, group, and individual users. Although social media is a major source of information, so too is traditional media. The information carried via social media is not completely distinct from traditional media (Pfeffer & Carley, 2012). Moreover, the information in social media is not always more timely than that in traditional media (Pfeffer & Carley, 2012). As more organizations and news agencies turn to the use of social media the relative impacts of social media and traditional media on social change become more complex, as does their role in governance. We have found that news agencies are among some of the most frequent tweeters, and are often re-tweeted within this data; e.g., there are approximately 10^5 tweets by BBC world in our data. Future work should consider the relative role of news agencies and other users relative to the change in topics over the course of crisis events.

Throughout the Arab Spring, social media had a presence. Twitter in particular was used both to provide and get information. In assessing data extracted for 15 countries over three years we identified key trends in topics and users. Against a backdrop of increasing Twitter usage, and country level shut downs in access, topics and groups emerge and fade. We found high levels of Arabic and non Arabic content, but with relatively little overlap. In general the Arabic Twitter network and non Arabic Twitter network seem to have little connection. Topics identified tend to be predominantly either Arabic or non Arabic. We found geo-temporal trends in topics. Specifically, temporally topics moved from expressions of concern to detailed political discussions. Geographically, highly localized topics tended to be narrower such as focusing on specific leaders; whereas, geographically dispersed topics tended to be more general. This suggests that in general, the more generic a topic the broader it's geographic and temporal footprint. Our results also suggest that the progress to revolution is one involving the incitement of concern and the transition to political specificity. We found that the user community and its connectivity increased over the course of the Arab Spring. Yet, this community remained fairly fragmented, held together largely by local opinion leaders.

The strength of these results is due, in part, to the fact that they span a wide geo-temporal swath and are not dependent on the vagaries of specific twitter users. The strength of the analysis is also due to the co-examination of both topics and users. The strength, however, points to a significant limitation in our ability to assess such large networks, and that is the ability to identify "topic-groups", i.e., those sets of users and topics that are tightly linked such as the set of users who only talk about particular terror activity or a specific soccer game. Advances are needed to support the rapid assessment of users and topics together to determine how these communities are evolving. Even without such methodological tools, the foregoing analysis does demonstrate the by considering both users and topics from a network perspective, and applying scalable network techniques, results in critical insight into social change. The combination of complex analytical techniques and high dimensional network data provides the analyst with the tools necessary to go beyond simple trend and sentiment analysis to an improved understanding of the way in which different sub-groups are interacting in the Twittersphere.

Acknowledgements

The authors would like to thank Dr. Huan Liu, for comments on an earlier draft of this work, and to thank Fred Morstatter, and Brendan O'Connor for insights into the data format. This work was supported in part by the Office of Naval Research (ONR) through a MURI N00014081186 on adversarial reasoning and through MINERVA N000141310835 on State Stability. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. government.

References

[1] "Internet 'cut off across Syria,'" *BBC News*. [Online]. Available: http://www.bbc.co.uk/news/technology-20546302. [Accessed: 21-Apr-2014].

[2] K. M. Carley, "Dynamic network analysis," in *Dynamic social network modeling and analysis: Workshop summary and papers*, 2003, pp. 133–145.

[3] W. R. Louis and R. Owen, *A revolutionary year: the Middle East in 1958*. IB Tauris, 2002.

[4] K. Selvik and S. Stenslie, *Stability and change in the modern Middle East*. IB Tauris, 2011.

[5] M. N. Barnett and E. Goldberg, "Dialogues in Arab Politics," *Comp. Polit. Stud.*, vol. 33, no. 2, pp. 271–272, 2000.

[6] P. N. Howard and M. R. Parks, "Social media and political change: Capacity, constraint, and consequence," *J. Commun.*, vol. 62, no. 2, pp. 359–362, 2012.

[7] G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and D. Boyd, "The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions," *Int. J. Commun.*, vol. 5, pp. 1375–1405, 2011.

[8] S. Meraz and Z. Papacharissi, "Networked Gatekeeping and Networked Framing on #Egypt," *Int. J. Press.*, vol. 18, no. 2, pp. 138–166, Apr. 2013.

[9] A. Bruns, T. Highfield, and J. Burgess, "The Arab Spring and Social Media Audiences English and Arabic Twitter Users and Their Networks," *Am. Behav. Sci.*, vol. 57, no. 7, pp. 871–898, 2013.

[10] K. Starbird and L. Palen, “(How) will the revolution be retweeted?: information diffusion and the 2011 Egyptian uprising,” in *Proceedings of the acm 2012 conference on computer supported cooperative work*, 2012, pp. 7–16.

[11] Y. Takhteyev, A. Gruzd, and B. Wellman, “Geography of Twitter networks,” *Soc. Netw.*, vol. 34, no. 1, pp. 73–81, Jan. 2012.

[12] E. Gilbert, “Predicting tie strength in a new medium,” in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, New York, NY, USA, 2012, pp. 1047–1056.

[13] H. Kwak, C. Lee, H. Park, and S. Moon, “What is Twitter, a social network or a news media?,” in *Proceedings of the 19th international conference on World wide web*, New York, NY, USA, 2010, pp. 591–600.

[14] K. Leetaru, “Culturomics 2.0: Forecasting large-scale human behavior using global news media tone in time and space,” *First Monday*, vol. 16, no. 9, 2011.

[15] K. Joseph, K. M. Carley, D. Filonuk, G. P. Morgan, and J. Pfeffer, “Arab Spring: from newspaper data to forecasting,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–17, Dec. 2014.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J Mach Learn Res*, vol. 3, pp. 993–1022, Mar. 2003.

[17] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, “Indexing by latent semantic analysis,” *JASIS*, vol. 41, no. 6, pp. 391–407, 1990.

[18] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, “Is the Sample Good Enough? Comparing Data from Twitter’s Streaming API with Twitter’s Firehose,” in *The 7th International Conference on Weblogs and Social Media (ICWSM-13), Boston, MA. Retrieved from http://www. public. asu. edu/~fmorstat/paperpdfs/icwsm2013. pdf*, 2013.

[19] K. Joseph, P. M. Landwehr, and K. M. Carley, “Two 1%s Don’t Make a Whole: Comparing Simultaneous Samples from Twitter’s Streaming API,” in *Social Computing, Behavioral-Cultural Modeling and Prediction*, W. G. Kennedy, N. Agarwal, and S. J. Yang, Eds. Springer International Publishing, 2014, pp. 75–83.

[20] A. Ritter, S. Clark, and O. Etzioni, “Named entity recognition in tweets: an experimental study,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1524–1534.

[21] J. Eisenstein, B. O’Connor, N. A. Smith, and E. P. Xing, “A latent variable model for geographic lexical variation,” in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2010, pp. 1277–1287.

[22] Z. Tufekci, “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls,” *ArXiv Prepr. ArXiv14037400*, 2014.

[23] L. Hong and B. D. Davison, “Empirical study of topic modeling in twitter,” in *Proceedings of the First Workshop on Social Media Analytics*, 2010, pp. 80–88.

[24] D. Ramage, S. Dumais, and D. Liebling, “Characterizing Microblogs with Topic Models,” in *ICWSM*, 2010.

[25] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, “Evaluation methods for topic models,” in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, pp. 1105–1112.

[26] D. Boyd and K. Crawford, “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon,” *Inf. Commun. Soc.*, vol. 15, no. 5, pp. 662–679, 2012.
[27] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *J. Stat. Mech. Theory Exp.*, vol. 2008, no. 10, p. P10008, 2008.
[28] M. E. J. Newman, “Modularity and community structure in networks,” *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006.