Supplementary Materials for "Spontaneous Collective Action"

Zachary Steinert-Threlkeld

August 15, 2015

1 Collection of Twitter Data

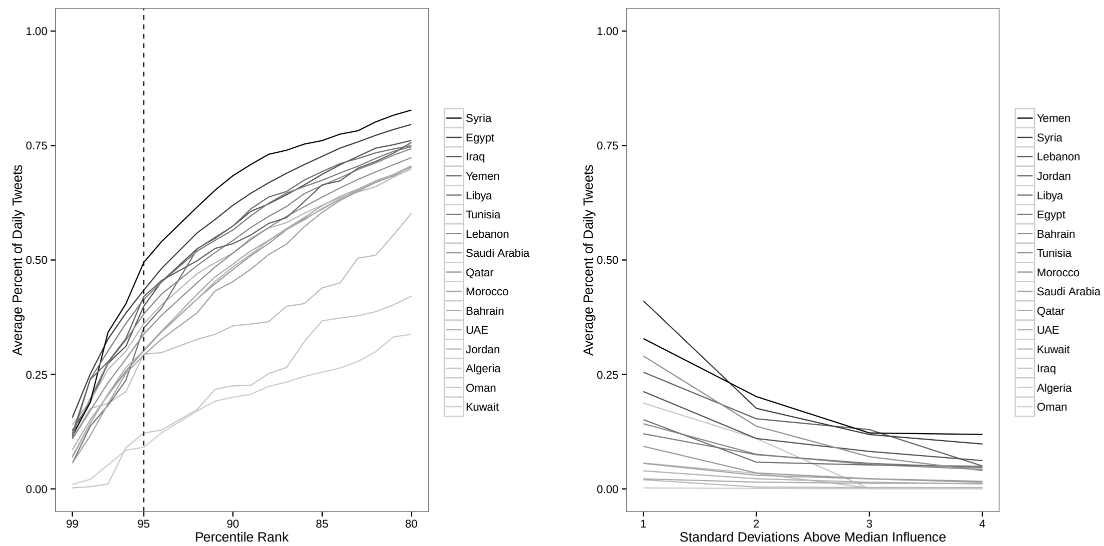
The tweets involved in this analysis were extracted from Twitter's 10% API, an unbiased sample of 10% of all public activity on the platform. There are two ways in which country of origin was identified. First, if a Twitter user has enabled location sharing, the tweet will have GPS coordinates, and those coordinates are used to assign country-location for that tweet. If the country-location is one of the 16, the tweet is saved. Second, users can report their location as part of their profile, and that location is reported as metadata with each tweet. The user-reported location is then compared to a dictionary of cities and country names to assign each tweet to a city or country. For more detail, see the Materials and Methods section of Mocanu et. al 2013.

2 Alternate Measures of Elites

The following figures show how tweet production and follower count vary depending on the threshold used.

Figure 1 shows the average percent of each country's daily tweets which come from popular users, based on two measures. Figure 1a shows users in the 99th percentile, 98th percentile, and so on to the 80th percentile. The vertical dashed line represents the 95th percentile, the threshold used in the paper's regressions. Figure 2b uses standard deviations above each country's median number of followers as its threshold. The lines now decrease because each 1 unit increase in standard deviation creates a smaller group, so there have to be fewer tweets in that group than the one immediately proceeding. The legends are ordered based on the 80th percentile: the country with the highest average percentage of daily tweets coming from users at or above the 80th percentile is Syria, then Egypt, so on down to Kuwait.

Figure 1: Daily Average Tweet Production by Popularity Threshold



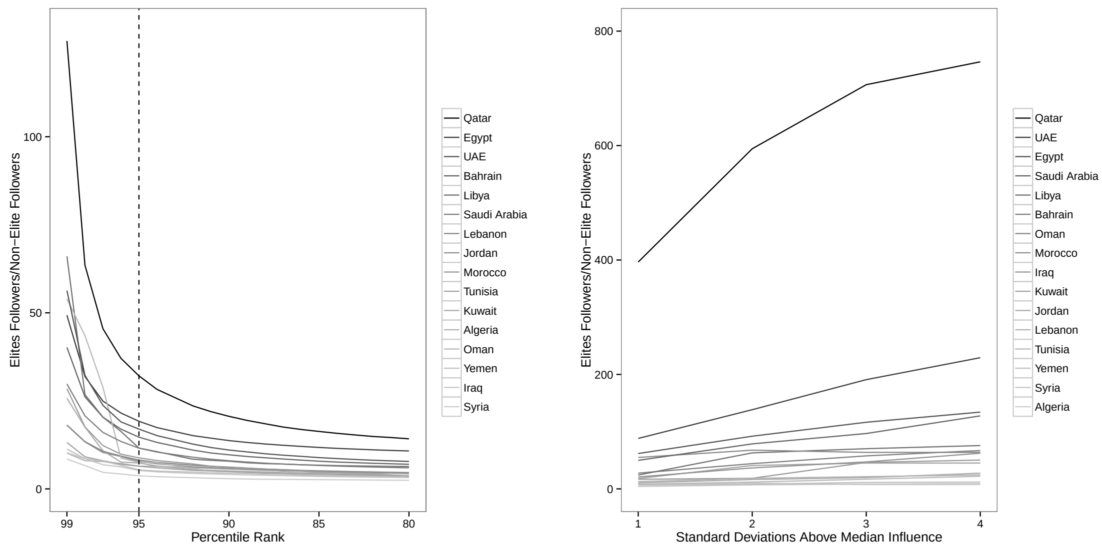
(a) Percentile

(b) Standard Deviation Rank

Figure 2 is similar to Figure 1 except the outcome measure is the ratio of followers for core users to followers for peripheral ones. Figure 2a sorts users based on percentiles, while Figure 2b sorts on standard deviations above median number of followers. In Qatar, users in the 99th percentile based on number of followers have approximately 150 times as many followers as a Qatari with the median number of followers; users with 1 standard deviation more followers than the median user have approximately 400 times as many followers as the median Qatari. As in Figure 1, the legend for each subfigure is ordered from highest to lowest.

The takeaway from these charts is that there is no inexpensive way to identify core members that will find the same kind of accounts across countries. The optimal approach to validating the chosen measure is therefore to run the main model separately as many times as there are core threshold measures. The results from this robustness check are reported in the paper's body text.

Figure 2: Elite Follower Ratio by Popularity Threshold



(a) Percentile

(b) Standard Deviation Rank

3 Exploring Upper Tail of Core Distribution

Table **??** shows descriptive statistics for core users. It replicates the presentation in Table 1 of the main paper. A few interesting trends emerge from Table **??**. First, notice how few accounts are in the very upper tail (99^{th} percentile or above, or 6 standard deviations above the median number of followers) of each country. Many countries have fewer than 10 users in these bins, and those users often appear in the dataset only a few times; see Oman, Yemen, and Algeria are in this bin. Second, there exists a steep descent in the number of followers from the 99.9^{th} percentile to the 99^{th}, and the descent is almost as steep down to the 95^{th} percentile. Third, notice that the 99^{th} and 99.9^{th} percentiles use hashtags and links at much higher rates than those even slightly below them in their country's followers' distribution. Fourth, and finally, notice that some accounts identified as being Tunisian or Egyptian appear in other countries as well, such as Lebanon, Morocco, and Qatar. This finding was unexpected and worth future investigation.

Table 1: Core Threshold Descriptive Statistics Across Countries

Country	Group	Accounts	Followers Avg.	Tweet Avg.	Mention %	Retweet %	Hashtag %	Link %
Algeria	1 SD Above Med.	15	1621.01	93.33	0.27	0.43	0.59	0.79
Algeria	6 SD Above Med.	2	6786.50	1.50	0.33	0.00	0.33	1.00
Algeria	99.9^{th} Percentile	1	9511.00	1.00	1.00	0.00	1.00	1.00

Table 1: Core Threshold Descriptive Statistics Across Countries

Country	Group	Accounts	Followers Avg.	Tweet Avg.	Mention %	Retweet %	Hashtag %	Link %
Algeria	99^{th} Percentile	5	3582.28	164.20	0.10	0.72	0.73	0.97
Algeria	Core 95 Percentile	23	1183.08	95.61	0.35	0.28	0.39	0.60
Bahrain	1 SD Above Med.	365	4445.40	412.00	0.36	0.04	0.44	0.36
Bahrain	6 SD Above Med.	64	14938.11	523.58	0.17	0.03	0.56	0.50
Bahrain	99.9^{th} Percentile	21	25657.88	824.91	0.09	0.02	0.58	0.58
Bahrain	99^{th} Percentile	201	6873.44	449.57	0.29	0.03	0.50	0.42
Bahrain	Core 95 Percentile	1005	1995.99	319.24	0.42	0.05	0.38	0.28
Egypt	Mainstream Media	1	103927.00	5281.00	0.00	0.00	0.74	0.70
Egypt	Non-Media Org.	2	23877.40	457.50	0.32	0.07	0.22	0.53
Egypt	MSM Employee	9	22463.50	650.22	0.41	0.01	0.21	0.21
Egypt	Blogger	15	8394.17	1070.67	0.52	0.08	0.33	0.22
Egypt	Activist	10	8036.60	703.40	0.42	0.07	0.28	0.33
Egypt	1 SD Above Med.	543	9585.76	830.09	0.43	0.06	0.27	0.34
Egypt	6 SD Above Med.	104	31542.21	1295.48	0.26	0.05	0.37	0.57
Egypt	99.9^{th} Percentile	80	37001.28	924.69	0.33	0.02	0.39	0.44
Egypt	99^{th} Percentile	793	7104.31	736.38	0.46	0.05	0.27	0.32
Egypt	Core 95 Percentile	3962	1868.80	409.94	0.45	0.04	0.25	0.30
Iraq	1 SD Above Med.	17	19710.78	485.24	0.62	0.06	0.39	0.09
Iraq	6 SD Above Med.	1	236675.00	1.00	0.00	0.00	0.00	1.00
Iraq	99.9^{th} Percentile	5	57288.40	59.60	0.16	0.06	0.75	0.11
Iraq	99^{th} Percentile	45	8576.75	414.38	0.45	0.20	0.25	0.10
Iraq	Core 95 Percentile	224	2254.30	260.17	0.40	0.14	0.20	0.20
Jordan	1 SD Above Med.	251	2117.48	277.47	0.37	0.06	0.24	0.51
Jordan	6 SD Above Med.	24	9621.04	470.92	0.19	0.03	0.35	0.73
Jordan	99.9^{th} Percentile	9	15371.72	282.11	0.44	0.03	0.12	0.44
Jordan	99^{th} Percentile	86	4342.62	446.14	0.24	0.03	0.20	0.71
Jordan	Core 95 Percentile	426	1465.35	231.61	0.39	0.05	0.24	0.49
Kuwait	1 SD Above Med.	56	7332.00	10.64	0.59	0.01	0.08	0.14
Kuwait	6 SD Above Med.	10	25240.53	7.00	0.40	0.07	0.07	0.30
Kuwait	99.9^{th} Percentile	4	43236.25	6.75	0.78	0.00	0.07	0.04
Kuwait	99^{th} Percentile	32	11216.00	9.09	0.47	0.02	0.10	0.13
Kuwait	Core 95 Percentile	158	3344.44	17.13	0.65	0.01	0.09	0.09
Lebanon	MSM Employee	1	56753.60	5.00	0.60	0.00	0.00	0.40
Lebanon	Blogger	2	7311.00	1.00	0.00	0.00	0.50	0.00
Lebanon	Activist	1	6681.62	170.00	0.74	0.01	0.11	0.15
Lebanon	1 SD Above Med.	368	2381.27	302.15	0.37	0.10	0.27	0.41
Lebanon	6 SD Above Med.	42	9173.22	458.29	0.18	0.02	0.25	0.67
Lebanon	99.9^{th} Percentile	18	14186.87	385.39	0.19	0.01	0.36	0.51
Lebanon	99^{th} Percentile	173	3886.31	340.76	0.28	0.12	0.29	0.52

Table 1: Core Threshold Descriptive Statistics Across Countries

Country	Group	Accounts	Followers Avg.	Tweet Avg.	Mention %	Retweet %	Hashtag %	Link %
Lebanon	Core 95 Percentile	863	1306.22	231.43	0.41	0.09	0.28	0.37
Libya	Blogger	1	2147.00	2.00	0.00	0.00	1.00	0.00
Libya	1 SD Above Med.	79	4910.78	162.68	0.47	0.04	0.36	0.19
Libya	6 SD Above Med.	10	18325.41	293.90	0.39	0.05	0.60	0.26
Libya	99.9^{th} Percentile	4	31026.75	369.00	0.20	0.07	0.80	0.34
Libya	99^{th} Percentile	37	8482.10	129.11	0.43	0.06	0.47	0.23
Libya	Core 95 Percentile	182	2540.17	164.10	0.35	0.19	0.31	0.38
Morocco	Blogger	1	1671.32	244.00	0.53	0.38	0.23	0.36
Morocco	1 SD Above Med.	142	6061.31	196.80	0.50	0.13	0.22	0.37
Morocco	6 SD Above Med.	22	23346.91	212.50	0.45	0.03	0.28	0.36
Morocco	99.9^{th} Percentile	10	37980.35	159.40	0.29	0.01	0.39	0.57
Morocco	99^{th} Percentile	98	8022.34	172.94	0.43	0.11	0.22	0.45
Morocco	Core 95 Percentile	489	2380.80	179.68	0.48	0.09	0.24	0.35
Oman	1 SD Above Med.	6	47661.06	3.00	0.33	0.00	0.44	0.17
Oman	6 SD Above Med.	2	122889.50	1.50	0.677	0.00	0.00	0.33
Oman	99.9^{th} Percentile	1	130487.00	2.00	1.00	0.00	0.00	0.00
Oman	99^{th} Percentile	7	41600.91	2.71	0.32	0.00	0.42	0.21
Oman	Core 95 Percentile	33	10328.22	31.42	0.55	0.03	0.11	0.32
Qatar	Mainstream Media	2	281861.00	1381.00	0.08	0.00	0.50	0.91
Qatar	MSM Employee	3	6360.29	836.67	0.65	0.01	0.27	0.19
Qatar	Blogger	1	2838.00	1.00	1.00	0.00	1.00	0.00
Qatar	1 SD Above Med.	58	26021.46	417.91	0.41	0.06	0.30	0.34
Qatar	6 SD Above Med.	8	127181.55	812.63	0.15	0.00	0.40	0.62
Qatar	99.9^{th} Percentile	22	56340.09	634.95	0.36	0.05	0.29	0.39
Qatar	99^{th} Percentile	217	8590.11	349.83	0.43	0.05	0.21	0.21
Qatar	Core 95 Percentile	1083	2218.35	307.21	0.46	0.06	0.16	0.12
Saudi Arabia	MSM Employee	2	53017.31	767.00	0.79	0.00	0.08	0.08
Saudi Arabia	1 SD Above Med.	425	9345.20	580.17	0.56	0.03	0.24	0.28
Saudi Arabia	6 SD Above Med.	74	32760.58	631.07	0.67	0.02	0.10	0.18
Saudi Arabia	99.9^{th} Percentile	89	28922.06	576.03	0.67	0.02	0.11	0.18
Saudi Arabia	99^{th} Percentile	885	5243.44	544.22	0.57	0.03	0.21	0.24
Saudi Arabia	Core 95 Percentile	4425	1448.05	337.76	0.58	0.04	0.14	0.17
Syria	1 SD Above Med.	148	1295.07	637.03	0.16	0.05	0.59	0.73
Syria	6 SD Above Med.	13	4891.79	598.00	0.15	0.02	0.44	0.72
Syria	99.9^{th} Percentile	4	1003.41	690.00	0.13	0.01	0.56	0.74
Syria	99^{th} Percentile	37	2927.40	719.65	0.21	0.02	0.56	0.67
Syria	Core 95 Percentile	182	1134.64	623.71	0.15	0.05	0.62	0.75
Tunisia	Mainstream Media	2	5604.50	741.00	0.16	0.12	0.77	0.79
Tunisia	MSM Employee	1	52503.00	1.00	0.00	0.00	0.00	1.00

Table 1: Core Threshold Descriptive Statistics Across Countries

Country	Group	Accounts	Followers Avg.	Tweet Avg.	Mention %	Retweet %	Hashtag %	Link %
Tunisia	Blogger	3	1910.77	258.33	0.57	0.13	0.30	0.20
Tunisia	Activist	4	2496.28	57.00	0.60	0.09	0.36	0.29
Tunisia	1 SD Above Med.	172	2499.44	385.69	0.45	0.13	0.30	0.41
Tunisia	6 SD Above Med.	15	11473.11	311.87	0.51	0.06	0.16	0.51
Tunisia	99.9^{th} Percentile	7	17749.31	206.71	0.25	0.06	0.22	0.68
Tunisia	99^{th} Percentile	62	4880.14	410.92	0.37	0.06	0.28	0.55
Tunisia	Core 95 Percentile	307	1681.87	308.22	0.47	0.11	0.31	0.38
UAE	1 SD Above Med.	240	15377.37	251.14	0.33	0.05	0.31	0.41
UAE	6 SD Above Med.	30	71908.66	342.30	0.27	0.06	0.42	0.32
UAE	99.9^{th} Percentile	48	50588.01	343.21	0.32	0.06	0.37	0.31
UAE	99^{th} Percentile	471	9018.08	232.85	0.39	0.07	0.26	0.37
Yemen	1 SD Above Med.	55	1217.52	367.22	0.13	0.03	0.51	0.67
Yemen	6 SD Above Med.	7	4405.43	280.00	0.06	0.01	0.33	0.72
Yemen	99.9^{th} Percentile	2	6694.44	452.00	0.01	0.00	0.01	0.83
Yemen	99^{th} Percentile	14	2868.15	532.21	0.10	0.02	0.60	0.69
Yemen	Core 95 Percentile	69	1039.71	374.57	0.13	0.03	0.52	0.70

4 Potential Model Misspecification

The main model may not capture the correct data generating process, so different specifications, in addition to those in the verification section of the main paper, are presented below. Table 2 shows these and verifies that likely confounds have not driven the results. In the first column, every variable is lagged by 3 time periods, though only results for the one-day lags are shown. $Coordination_{i,t-1}$ is still positive and significant but not 2 or 3 days before a protest. None of the variables of core members' activity, even the lags, are significant, further showing that their behavior does not affect subsequent protest.

Table 2: Robust to Model Specification

	$Protest_{i,t}$		
	3 Lags	Only Arabic	Friday Fixed Effects
	(1)	(2)	(3)
$\text{Coordination}_{i,t-1}$	1.454**	2.324***	2.518***
	(0.593)	(0.502)	(0.657)
Hashtag $\%_{i,t-1}$	0.399	0.877**	0.541
	(0.692)	(0.379)	(0.638)
Retweet $\%_{i,t-1}$	−0.062	1.422***	−0.214
	(1.077)	(0.506)	(1.027)
Link $\%_{i,t-1}$	−0.656	−0.083	−0.768**
	(0.444)	(0.424)	(0.382)
Mention $\%_{i,t-1}$	−0.551	−0.315	−0.872**
	(0.481)	(0.515)	(0.390)
$\text{Repression}_{i,t-1}$	0.004	0.023*	0.019*
	(0.010)	(0.012)	(0.011)
$\text{Protest}_{i,t-1}$	0.099***	0.115***	0.119***
	(0.008)	(0.010)	(0.010)
Core Hashtag $\%_{i,t-1}$	0.626	0.580	0.930***
	(0.504)	(0.377)	(0.340)
Core Retweet $\%_{i,t-1}$	0.233	−0.153	0.142
	(0.375)	(0.251)	(0.385)
Core Link $\%_{i,t-1}$	0.187	0.176	0.685
	(0.521)	(0.565)	(0.577)
Core Mention $\%_{i,t-1}$	0.206	0.037	−0.240
	(0.347)	(0.201)	(0.268)
$\text{Friday}_{i,t}$			0.539***
			(0.182)
$\text{Coordination}_{i,t-1}$*Core Hashtag $\%_{i,t-1}$	−1.004	−1.136	−1.740*
	(1.391)	(0.842)	(0.980)
Intercept	−2.225***	−1.210***	−1.080***
	(0.589)	(0.330)	(0.207)
Country FE	Yes	Yes	Yes
N	6,468	6,397	6,620
Log Likelihood	−8,083.642	−8,074.174	−8,244.873

*p < .1; **p < .05; ***p < .01

It is also possible that English tweets drive the results. Twitter did not introduce an Arabic interface until March 2012 (though it did accept Arabic input for tweets), so users in the study may have needed a passing familiarity with English. Moreover, cellphones, especially smart phones, were not widespread in 2011, and Twitter use was relegated to an even smaller percentage of the population (Mourtada & Salem 2011, Tufekci & Wilson 2012, *International Telecommunications Union Statistics* 2014). Appealing to international actors is a common tactic of activists (Keck & Sikkink 1998) and was a prominent use of Twitter by Arab Spring activists (Howard & Hussain 2011, Howard, Duffy, Freelon, Hussain, Mari & Mazaid 2011, Aday, Freelon, Farrell, Lynch & Sides 2012, Starbird & Palen 2012). Since individuals in the protest countries are most likely not going to consume protest information in English, it is possible that international attention-seeking drives the results. To rule out this possibility, the main model is rerun but ignoring English tweets.[1] Replicating the main model, Column 2 in Table 2, using only Arabic tweets confirms the importance of peripherals' coordination and impotence of those in the core.

The main model does not account for potential periodicity in the data: Friday, the main religious day for Muslims, also saw surges in protest. In Egypt, for example, January 25^{th} was chosen as the first protest day because it coincided with National Police Day, a national holiday commemorating a massacre of Egyptian police by the British in 1952. January 25^{th} was a Tuesday, and many people returned home that night and to work the next day. Momentum then built for another major protest on January 28^{th}, a Friday, when individuals surged to Tahrir Square after the end of prayers (Ghonim 2012).[2] To make sure that $Coordination_{i,t-1}$ does not capture religious effects, a fixed effect for Fridays is added. The model with Friday fixed effects, shown in Column 3 of Table 2, shows that coordination occurs independent of Friday effects, though Fridays do have more protests than non-Fridays. Column 3 also suggests some effect for the percent of hashtags originating in the core, though coordination from the core is not statistically significant and negative.

Table 3 shows that the results hold when countries at the extreme of the protest and tweet distribution are dropped. To identity countries to drop, each country's average number of protests and tweets per capita over the entire sample were calculated. Model 1 drops the 5 countries with the fewest protests per capita, Model 2 drops the 5 with the most. Models 3 and 4 do the same

[1] A Python implementation of Google's Compact Language Detector was used to identify each tweet's language.

[2] It is not a coincidence that the state shut-down all internet and phone service (save the internet connection for the Cairo stock exchange) late at night on the 27^{th}.

but with tweets per capita. In all models, peripheral coordination is positive and statistically significant while core coordination is negative and significant. The only behavior from the core that may be significant and positive is Core Hashtag $\%_{i,t-1}$, though the effect is out weighted by core coordination.

Table 3: Robust to Removing Countries

	$Protest_{i,t}$			
	Top 10 Protests	Bottom 10 Protests	Top 10 Tweets	Bottom 10 Tweets
	(1)	(2)	(3)	(4)
$\text{Coordination}_{i,t-1}$	2.908^{***}	1.861^{**}	3.373^{***}	2.470^{***}
	(0.627)	(0.815)	(0.802)	(0.554)
Hashtag $\%_{i,t-1}$	0.682	0.090	0.898	0.452
	(0.658)	(0.715)	(0.598)	(0.615)
Retweet $\%_{i,t-1}$	−0.071	−0.328	−1.785	−0.385
	(1.326)	(1.359)	(1.464)	(1.051)
Link $\%_{i,t-1}$	-1.048^{**}	−0.257	-1.403^{**}	-0.860^{**}
	(0.450)	(0.329)	(0.560)	(0.430)
Mention $\%_{i,t-1}$	−0.933	-1.098^{***}	-1.335^{**}	-1.011^{**}
	(0.570)	(0.381)	(0.640)	(0.400)
$\text{Repression}_{i,t-1}$	0.018	−0.005	0.013	0.027^{**}
	(0.013)	(0.011)	(0.009)	(0.011)
$\text{Protest}_{i,t-1}$	0.106^{***}	0.116^{***}	0.090^{***}	0.106^{***}
	(0.010)	(0.015)	(0.010)	(0.010)
Core Hashtag $\%_{i,t-1}$	0.708^{*}	0.974^{*}	1.595	0.907^{**}
	(0.427)	(0.497)	(1.052)	(0.375)
Core Retweet $\%_{i,t-1}$	0.060	0.758	−0.390	0.122
	(0.479)	(0.473)	(0.587)	(0.440)
Core Link $\%_{i,t-1}$	1.138^{*}	0.658	0.716	1.059^{**}
	(0.602)	(0.723)	(0.815)	(0.517)
Core Mention $\%_{i,t-1}$	0.007	-0.475^{*}	0.190	−0.317
	(0.348)	(0.259)	(0.560)	(0.228)
$\text{Coordination}_{i,t-1}$*Core Hashtag $\%_{i,t-1}$	-2.170^{**}	-2.809^{*}	-1.744^{*}	-1.969^{*}
	(1.016)	(1.644)	(0.969)	(1.005)
Intercept	-1.600^{***}	-1.229^{***}	-1.748^{***}	-0.978^{***}
	(0.331)	(0.263)	(0.503)	(0.240)
Country FE	Yes	Yes	Yes	Yes
N	4,551	4,545	4,559	4,545
Log Likelihood	−6,900.976	−4,944.046	−5,694.607	−6,674.396

$^{*}p < .1$; $^{**}p < .05$; $^{***}p < .01$

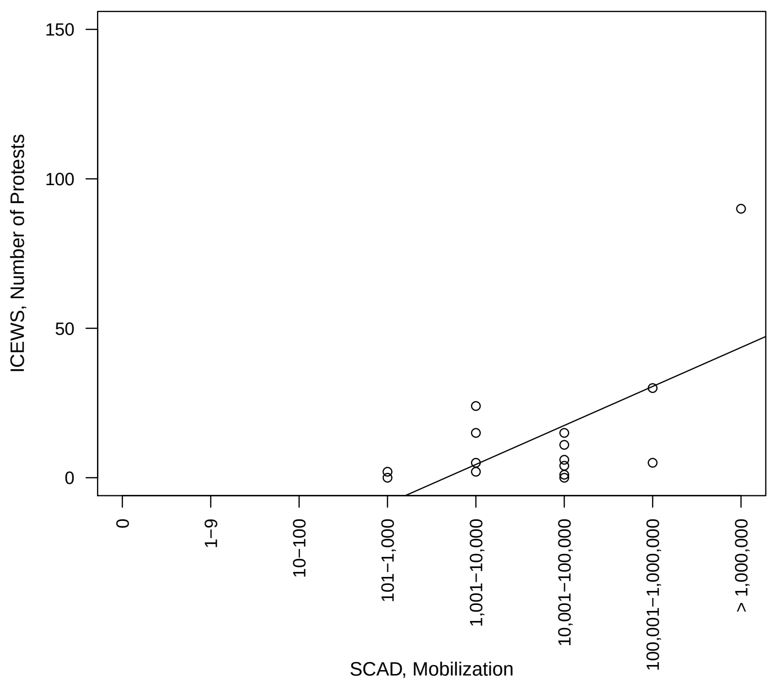
5 Events Data

The overriding concern with using ICEWS, and machine-coded events data more generally, is measurement validity. This concern has three subcomponents. First, such data may capture news coverage of events more than it captures actual events, e.g. a large number of "protests" may actually mean a large number of articles about a protest. Second, news sources overreport novel events and underreport persistent ones (such as civil war or ongoing protests), biasing event counts even if duplication is not a problem (Davenport & Ball 2002). ICEWS' noise, though very low, may therefore be biased. Third, GDELT may not record the true number of protests, as it is known to have poor deduplication (Caren 2014), so the recorded number of protests might be inflated.

To the first concern: ICEWS correlates strongly with ACLED, a handcoded dataset, and GDELT, another machine coded dataset, and this correlation matches previous analyses. While no correlation statistics are given, time series figures and tables provided in an analysis from Michael Ward's team, the first academic users of ICEWS, shows that GDELT captures the same changes in activity as ICEWS (Ward, Beger, Cutler, Dorff & Radford 2013). Another analysis comparing the two shows that ICEWS correlates highly with GDELT and does best on conflictual events (Arva, Beieler, Fisher, Lara, Schrodt, Song, Sowell & Stehle 2013). See the main paper for a discussion of that finding.

Hand-coded datasets have the highest levels of deduplication, as the dedepulication rules can be as flexible as necessary because of human oversight. As with machine-coded datasets, GDELT correlates positively with hand-coded datasets. A comparison of GDELT to two hand-coded datasets of violence in Africa finds that GDELT correlates with ACELD at .64, the Geo-Referenced Event Dataset at .33 (Hammond & Weidmann 2014). The Social Conflict in Africa Dataset tracks protest, among other events, across Africa (and Mexico, Central America, and the Caribbean) from 1990-2013 using handcoded articles from the *Associated Press* and *Agence France Presse* (Hendrix, Hamner, Case, Linebarger, Stull & Williams 2012). To compare SCAD to ICEWS, all events that were types 1 (organized demonstration), 2 (spontaneous demonstration), 3 (organized riot), or 4 (spontaneous riot) from Algeria, Egypt, Libya, Morocco, and Tunisia were selected. Those events are matched to their corresponding days in the ICEWS data, and only those events labeled as nationwide are kept. Because SCAD aggregates events which occur simultaneously in different parts of a country that are of the same type and topic, "nationwide" encapsulates more than one (but how many is unknown) event. The estimated number of participants, which is reported on a log scale, is then taken to see if it correlates with the number of protests recorded by ICEWS for that country-day. The results are shown in Figure 3. The more attendees at nationwide protests, the more protests ICEWS records.[3]

Figure 3: More Protestors Correlate with More GDELT Protests



GDELT correlates with many other event datasets that have been designed to account for news coverage bias. An analysis comparing GDELT's report of violent events in Syria with Syria Tracker's, a crowdsourced project to track violence in Syria, finds a correlation of .53 between the two (Masad 2013).[4] Similar comparisons do not exist for ICEWS, but ICEWS' correlation with GDELT suggests the results would be the same.

[3]This relationship is less strong when the analysis is disaggregated to the city level because ICEWS, and machine-coded events data more generally, is imprecise on subnational geolocation.

[4]The correlation varies depending on which governate one analyses. Geocoding events from news reports is even harder than coding events from news reports, so that the correlations are less reliable than nationwide aggregates is not surprising. The difficulty of geocoding is the primary reason this paper kept analysis at the country-day level.

The second concern, the most serious one, is that machine-coded events data may simply replicate bias inherent in news coverage. ICEWS probably over-reports events when an event is new or under-reports an event if it has lasted for a long time. Reporting is known to spike when an event is new but then tire or reporting on that event after some time. The event therefore will seem to be more widespread than it actually is when it is new but less widespread than it actually is when it has existed for awhile. This problem is especially acute in conflict reporting: what bleeds leads, but continued bleeding is still considered boring. This bias has most recently been shown to apply in the context of the Syrian civil war (Masad 2013).

Before proceeding, two caveats should be noted. First, the overall effect of news coverage bias is not clear. While it may lead to spurious positive findings at the start of protests, one is less likely to find a correlation, and may find a negative one, if the bias underreports subsequent protests. That is, after n months of protests, individuals still use Twitter at the same levels as before and for the same purpose, meaning any change in Twitter measures reflect true changes in the variable of interest. But these true changes in coordination are then compared to changes in protest counts that may undercount protests if newspapers fatigue of reporting protests. Second, the negative bias in machine-coded data will also exist in handcoded datasets that also rely on newspaper coverage. If a newspaper does not cover an event, then no amount of scholarly intervention will be able to insert that protest into a dataset. Any bias that ICEWS exhibits is therefore a reflection of bias in news reporting, but whether the net effect of the two biases cancel out is an open question. Because this study covers time before, during, and after the protest events, the net bias effect could be positive, negative, or neutral.

While there is no ground truth against which to compare ICEWS, analysis of the protest patterns revealed in ICEWS suggest news coverage bias does not drive results.[5] ICEWS clearly records protests when they are known to have happened, and it records them at magnitudes which vary according to commonly held beliefs on how widespread those protests were. For example, ICEWS records many more protests for Egypt than it does for Qatar, and it records more protests on January 25^{th}, the first day of protests, January 28^{th}, and the day Mubarak left power than in the summer, though the summer still has higher levels of protest than before January 24^{th}, 2011.

[5] Indeed, ICEWS is the gold standard in machine-coded events data and the closest such dataset scholars have to a ground truth.

Figures 4 and 5 show this behavior in Egypt and Bahrain, high-protest countries, and Morocco and Qatar, low-protest countries.

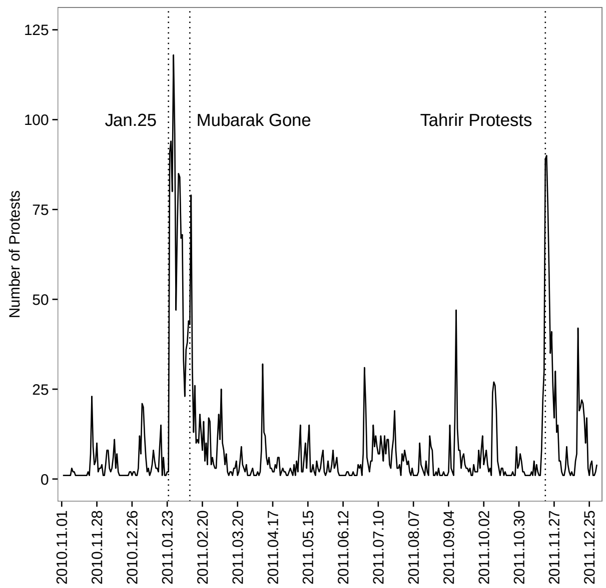
Figure 4 shows that counts of protest vary and vary in ways that accord with subjective understandings of protests in Egypt and Bahrain. Figures 4a and 4b show total number of protests for Egypt and Bahrain, respectively. They show a marked increase in protests on January 25^{th} and February 14^{th}, with protest continuing throughout the year. Both countries experience sustained protest throughout 2011 and had their protest movements start on January 25^{th} (Egypt) and February 14^{th} (Bahrain).

While the absolute number of protests tracks actual protests, it is possible that it only does so because it is driven by news coverage. In other words, the number of protests shows spurious correlation because it is a function of how much news coverage a country receives. To control for how much news coverage a country receives, each day's count of protests is divided by the number of events, a proxy for news coverage, reported in that country; these results are shown in Figures 4c and 4d. Controlling for news coverage, the same patterns hold. There is greater protest intensity (more protests per events) on relevant days, and protest intensity overall is greater after the start of protests than before.

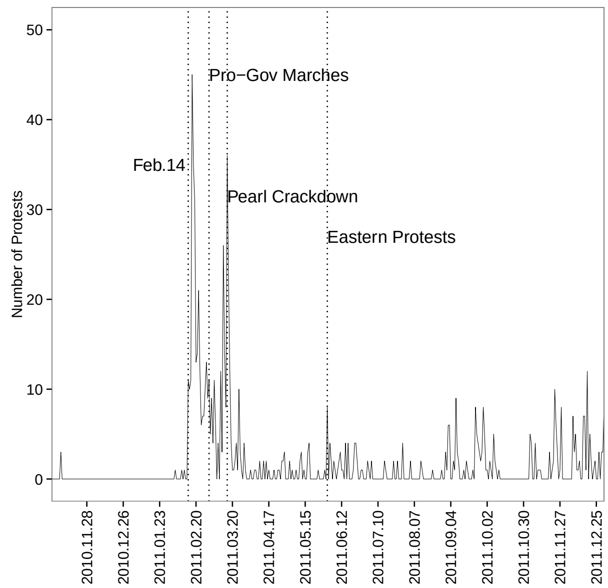
Figure 5 is the same as Figure 4 except it looks at two countries understood to have experienced little (Morocco) to no (Qatar) disruption during the Arab Spring. Right away, it is clear that Morocco and Qatar experienced fewer protests, which is expected since they are smaller countries. Both countries seem to have random variance in the number of protests recorded, with Morocco showing distinctive increases on February 20^{th}, the first day of organized protests there, and one month later. Qatar's most distinctive increase occurs on January 25^{th}; as no news reports discuss protests there then, that day is probably a spurious correlation with Egypt's protests. The visualization in Figures 4 and 5 corroborate the regression model created with the protest rate as the dependent variable.

Unexpectedly, Morocco has much higher levels of protest intensity than Egypt or Bahrain. This finding suggests one of two possibilities. First, Morocco may have actually had more protest than is commonly realized. If that is the case, it is interesting and poses no threat to the paper's conclusions. Second, it may be that Morocco receives little attention from news sources, so the high levels of protest intensity are because it is reported on most often when protests occur. If that

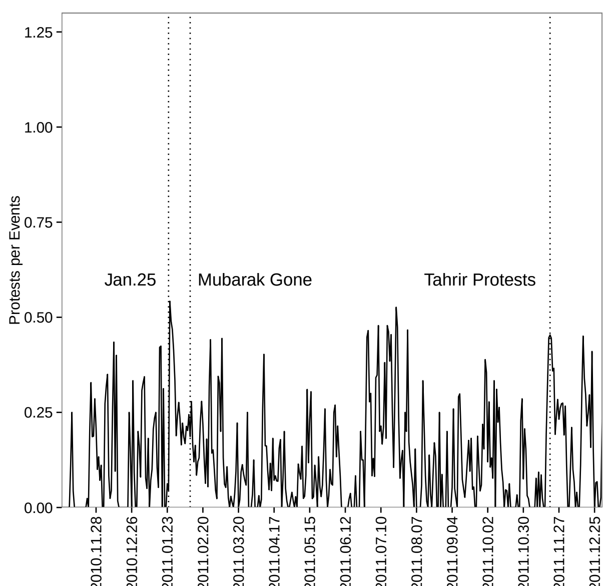
Figure 4: Countries with High Levels of Protest have High Levels of Protest in GDELT



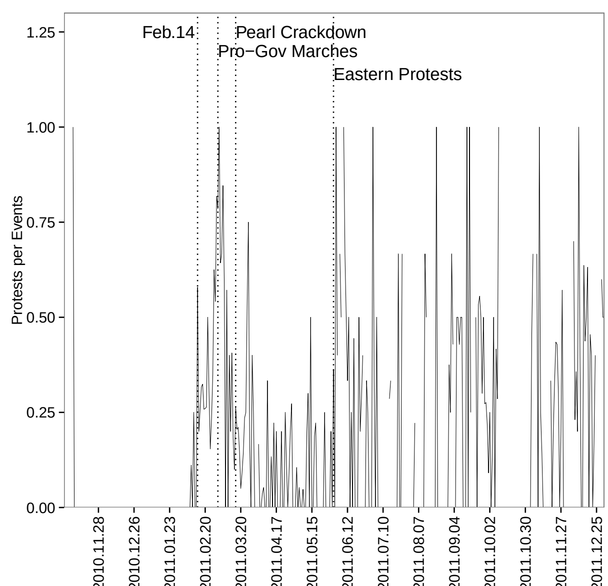
(a) Egypt, Total Protests



(b) Bahrain, Total Protests

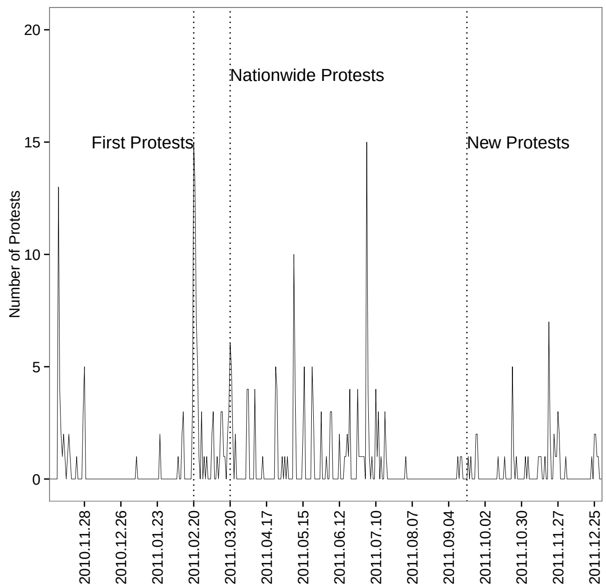


(c) Egypt, Protest per Event

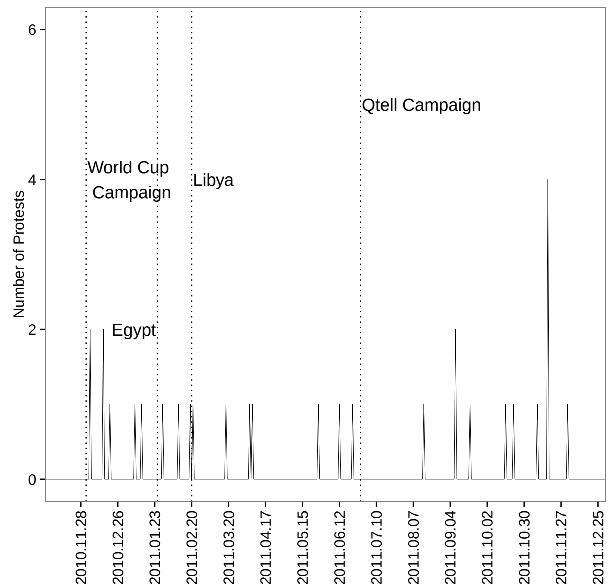


(d) Bahrain, Protest per Event

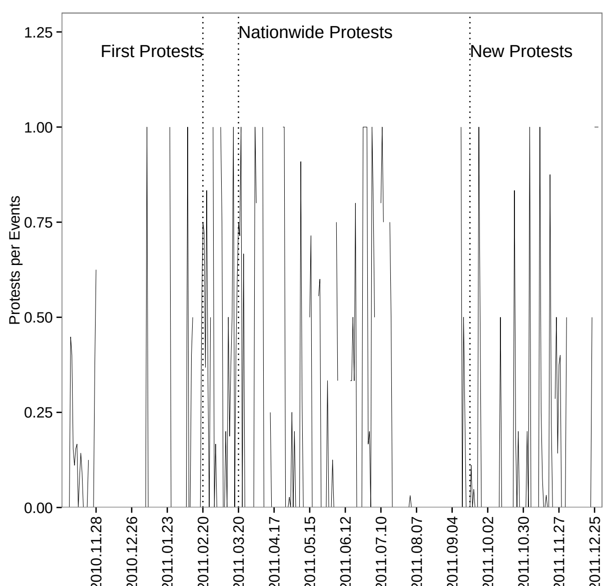
Figure 5: Countries with Low Levels of Protest have Low Levels of Protest in GDELT



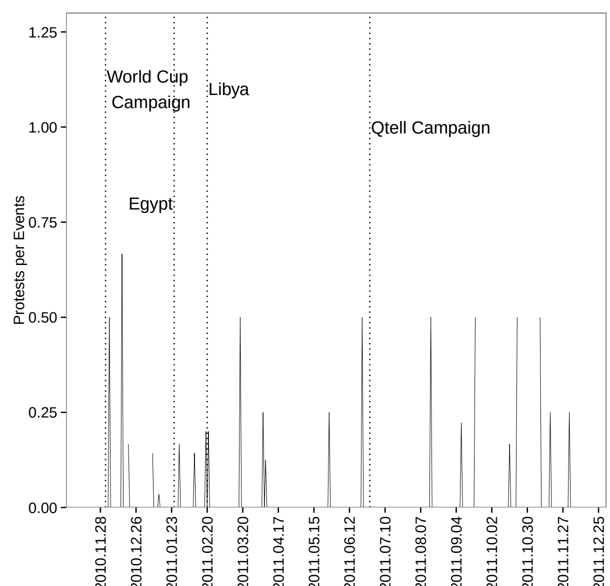
(a) Morocco, Total Protests



(b) Qatar, Total Protests



(c) Morocco, Protest per Event



(d) Qatar, Protest per Event

is the case, the recorded numbers of protest may have higher variance than they should, as protests that occur but do not attract attention are not reported while those that do attract coverage are covered extensively.

Bias in news reporting is also controlled for with two models. First, the main model is recreated without any day with an inordinate amount of protests reported, as these are days likely to suffer from novelty bias. This model is shown in Table 3, and the results do not differ from when the days are included. Second, protests from ACLED are used as the dependent variable in almost the same model. The two differences are that ACLED does not measure state repression and a zero-inflated negative binomial is used under the assumption that ACLED underreports some protest events.[6] These results are reported in Table 4. $Coordination\_i, t-1$ is still positive but no longer statistically significant, while core coordination is still not significant. A positive effect is found for the percent of tweets with hashtags that are from the core, while a larger negative effect is found for tweets from the core that mention another user. These results should be weighted very slightly, as ACLED only records 5 countries from the study's main sample.

The third concern is epistemological and correct, but it does not change the statistical interpretation of the model. If the noise is simply a scalar, then the problem is only one of effect size but not direction. For example, if ICEWS multiplies the true number of protests by 5 each time a protest is observed, than the effect found for any variable will be times larger than it actually is, but its direction and relative magnitude will not matter. If the noise of ICEWS is therefore uniform at each moment of time and space, then the noise is a nuisance but does not bias inference.

The noise can threaten inference if it does not vary with a variable of interest or is biased. If it is does not vary with a variable of interest, such as protest, then it can hide any effect of protest that does actually exist. While this outcome would be unfortunate, it increases the risk of a false negative (Type II error). False negatives are unfortunate, but a null finding is not what is being tested, meaning such an effect biases against any finding.[7]

[6] Zeroes are modeled with only an intercept, though rerunning the model with zeroes as a function of an intercept and country fixed-effects does not change results.

[7] It creates the risk of a false positive if some other noise component drives the reported number of protests but is not actually driven by the real world number of protests. Whether that is the case or not cannot be known.

Table 4: ACLED Protests

	Protest
$\text{Coordination}_{i,t-1}$	0.464
	(1.471)
Hashtag $\%_{i,t-1}$	1.112
	(1.902)
Retweet $\%_{i,t-1}$	−1.024
	(1.986)
Link $\%_{i,t-1}$	0.371
	(0.669)
Mention $\%_{i,t-1}$	−0.257
	(0.828)
$\text{Protest}_{i,t-1}$	0.304^{***}
	(0.082)
Core Hashtag $\%_{i,t-1}$	2.030^{**}
	(0.811)
Core Retweet $\%_{i,t-1}$	0.081
	(0.563)
Core Link $\%_{i,t-1}$	−0.640
	(1.213)
Core Mention $\%_{i,t-1}$	-2.138^{*}
	(1.258)
$\text{Coordination}_{i,t-1}$*Elite Reachout $\%_{i,t-1}$	4.885
	(3.525)
Intercept	-2.862^{***}
	(0.514)
Country FE	Yes
N	2,069
Log Likelihood	−1,252.761

*p < .1; **p < .05; ***p < .01

6 Codebook

Please see `TweetTranslation_Codebook_v3.docx` for detail on coding categories, sample tweets for each category, and individual coders.

7 Topic Model Detail

The primary problem with relying on hashtags is that the researcher has to surmise meaning from the tag, and the tag can be attached to texts with wide ranging meaning; only for the most specific hashtags, such as #postegyptianrevolutionsocialtrends can one safely assume a strong correlation between the hashtag and the meaning of its content. To more precisely measure meaning, one has to create a topic model. A topic model is a statistical algorithm that determines how features of a document words, sets of words, syntax, etcetera correspond to the topic of the document.[8]

To create the supervised topic models, 3,000 tweets from Egypt and Bahrain (6,000 total) were handcoded. The coding for Egypt was performed by a team of 3 undergraduates, all native Arabic speakers, who were asked to assign a tweet to any of 40 categories; a tweet could belong to multiple categories. The coders agreed on 95.09% of the tweet-categories. Of the categories directly related to protests, agreement ranged from 66.05% (if a tweet was political but not about protests) to 97.35% (if a tweet was about economic security. Intercoder reliability was 84.00% for tweets about protest coordination and 90.51% for tweets providing common knowledge. Because of financial constraints, the coding for Bahrain was performed by the author and a colleague. The author coded 1,500 English tweets, the colleague, who conducts fieldwork in the Middle East, 1,500 Arabic ones. See the Supplementary Materials for an explanation of these categories.

Upon completion of coding, a supervised topic model was created. A supervised approach was chosen, for 3 reasons. First, the level of interpretation required of unsupervised approaches leaves results of those models difficult to interpret. In the unsupervised approach, one takes a collection of documents, tells the computer to how many categories the documents belong, and the computer sorts the documents into those categories depending on a loss-minimization criterion. The number of categories is arbitrary, and the researcher has to test different numbers to find which appears

[8] "Document" means the textual unit of analysis. In this study, the document is the tweet, but it can be any text: a speech, a magazine article, a collection of articles, a Facebook post &c.

to best divide the documents into natural categories. The researcher then has to interpret the sorting of the documents to understand what real-world topic the groupings represent. Second, even if the interpretation of each category is not contentious, the number of categories is. There is no clear rule to distinguish between 5, 10, or 100 categories. While the supervised approach also relies on choosing a number of topics to which each document could belong, that decision is driven by theory, the researcher's contextual knowledge, and an iterative reading of the documents. Third, supervised learning allows the researcher to define the categories in which one is interested. Unsupervised approaches require the researcher to fit a collection of documents to a grouping that may correspond to a real-world category, whereas the supervised approach fits real-world categories to a document that may or may not actually be about that topic. The latter is therefore best when one knows for what one is looking, such as tweets coordinating protest or talking about a state's repressive response. See Grimmer 2013 and Lucas 2015 for more detail on text analysis.

After deciding upon a supervised topic model, a grid search was performed to tune model parameters. Specifically, a support vector machine and 3 varieties of Naive Bayes classifiers were tested. Each were tried with varying numbers of document features, with the features defined 1, 2, or 3 n-grams. Those combinations were then trained on 95% of the coded tweets and tested on the remaining 5; the specific 95% was chosen at random with replacement. The training was repeated a variable number of times, from once to fifty, and a tweet is classified as belonging to a category if more than half of the models assign it to that category. This process is known in machine learning as a bagged ensemble, is equivalent to bootstrapping, and lowers the variance of a model. This entire process was performed 4 times: once on protest coordination tweets in Egypt and Bahrain, and then again for common knowledge ones in each country. The model with the highest F1 score was chosen and used to generate predictions for the out of sample data. The final classification parameters are shown in Table 5.

Table 5: Topic Models' Parameters

Country	Topic	Classifier	Features	N-grams	Bags	F1
Bahrain	Coordination	Bernoulli Naïve Bayes	450	3	30	.65
Bahrain	Common Knowledge	Bernoulli Naïve Bayes	600	1	15	.62
Egypt	Coordination	Bernoulli Naïve Bayes	800	2	25	.64
Egypt	Common Knowledge	Bernoulli Naïve Bayes	900	3	10	.59

References

Aday, Sean, Deen Freelon, Henry Farrell, Marc Lynch & John Sides. 2012. New Media and Conflict After the Arab Spring. Technical report United States Institute of Peace Washington D.C.: .

Arva, Bryan, John Beieler, Ben Fisher, Gustavo Lara, Philip A Schrodt, Wonjun Song, Marsha Sowell & Sam Stehle. 2013. "Improving Forecasts of International Events of Interest.".

Caren, Neal. 2014. "It is time to get rid of the E in GDELT.".
URL: *http://badhessian.org/2014/05/it-is-time-to-get-rid-of-the-e-in-gdelt/*

Davenport, Christian & Patrick Ball. 2002. "Views to a Kill: Exploring the Implications of Source Selection in the Case of Guatemalan State Terror, 1977-1995." *Journal of Conflict Resolution* 46(3):427–450.

Ghonim, Wael. 2012. *Revolution 2.0: The Power of the People is Greater Than the People in Power: A Memoir*. Houghton Mifflin Co.

Grimmer, Justin & Brandom M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267–297.

Hammond, J. & N. B. Weidmann. 2014. "Using machine-coded event data for the micro-level study of political violence." *Research & Politics* 1(2):1–8.

Hanna, Alexander. 2013. "Computer-aided Content Analysis of Digitally Enabled Movements." *Mobilization: An International Quarterly* 18(4):367–388.

Hendrix, Cullen, Jesse Hamner, Christina Case, Chris Linebarger, Emily Stull & Jennifer Williams. 2012. "Social Conflict in Africa: A New Database." *International Interactions* 38(4):503–511.

Howard, Philip N., Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari & Marwa Mazaid. 2011. Opening Close Regimes: What Was the Role of Social Media During the Arab Spring? Technical report Project on Information Technology & Political Islam.

Howard, Philip N. & Muzammil M. Hussain. 2011. "The Role of Digital Media." *Journal of Democracy* 22(3):35–48.

International Telecommunications Union Statistics. 2014.

URL: *http://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx*

Keck, Margaret E. & Kathryn Sikkink. 1998. *Activists Beyond Borders: Advocacy Networks in International Politics*. Ithaca: Cornell University Press.

Masad, David. 2013. "Studying the Syrian Civil War with GDELT.".

Mocanu, Delia, Andrea Baronchelli, Nicola Perra, Alessandro Vespignani, Bruno Goncalves & Qian Zhang. 2013. "The Twitter of Babel: Mapping World Languages through Microblogging Platforms." *PLOS One* 8(4):e61981.

Mourtada, Racha & Fadi Salem. 2011. Civil Movements: The Impact of Facebook and Twitter. Technical Report 2 Dubai Scool of Government.

Starbird, Kate & Leysia Palen. 2012. (How) Will the Revolution be Retweeted? Information Diffusion and the 2011 Egyptian Uprising. In *CSCW*. Seattle: .

Tufekci, Zeynep & Christopher Wilson. 2012. "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square." *Journal of Communication* 62(2):363–379.

Ward, Michael D., Andreas Beger, Josh Cutler, Cassy Dorff & Ben Radford. 2013. "Comparing GDELT and ICEWS Event Data.".