The Decline of Social Media Censorship and the Rise of Self-Censorship after the 2016 Failed Turkish Coup

Rima S. Tanash
Rice University

Zhouhan Chen
Rice Universityy

Melissa Marschall
Rice University

Dan S. Wallach
Rice University

Abstract

In this work, we examine the effect of the 2016 Turkish coup on social media censorship, both by the government ordering Twitter to conduct censorship and as well by people removing their own tweets. We compared 5.5M tweets collected from Turkey pre-coup to 8.5M tweets collected post-coup. Although self-censorship of the press is not a novel practice following past military coups in Turkey, in this work we examine and quantify social media self-censorship, and empirically compare its effect relative to government-implemented censorship of social media.

Our measurements following the coup show a 72% decline in publicly identifiable government-censored tweets. We attribute this, in part, to an estimated 43% decline in overall Twitter usage in Turkey and in part to users' self-censorship. Supporting this theory, we detected that 41% of all users in our pre-coup dataset voluntarily removed 18% of their old tweets by either switching their accounts to protected mode, deleting their accounts, or deleting some tweets. Using NLP and graph metrics, we identify a new focus of Turkish government censorship on the *Gülen movement*. Our analysis show pro-Gülen tweets being widely self-censored. Additionally, we detected 40% more publicly-accessible anti-Gülen tweets. Unlike activists who regularly tweet political content, and are more likely to be censored by the government, we found that self-censoring users appear to be more typical users who normally post neutral tweets, and only 6% political tweets on average.

1 Introduction

The short lived Turkish military coup attempt on July 15, 2016, left the streets of Ankara in distress, with over 265 people killed [24]. The Turkish president, Recep Tayyip Erdoğan, was quick to blame Fethullah Gülen and his followers for the coup-plotter [7]. Fethullah Gülen is a former political figure who was previously close to President Erdoğan [11], and the founder of the Islamic movement *Hizmet* (a.k.a., the *Gülen movement*) [1], also known for his support for science, and thousands of schools and universities around the world [10]. Amid the coup, 140 journalists were arrested over ties to the *Gülen-movement*, 29 news outlets were shut down despite some being anti-Gülenist, and 21,000 academics were fired [4, 21, 18].

This coup is not unique in Turkish history. Five other coups were carried between 1960 and 1997 [8, 5]. Although self-censorship and censorship of the press have become commonplace following previous Turkish coups [3], self-censorship of social media is a new twist on an old story. Many Turkish Twitter users started *self-censoring* their Twitter accounts by switching to protected mode, presumably to avoid punishment for the public expression of prohibited content [7]. Considering this coup as a turning point in Turkish politics, we expect its effect to be reflected in our data. Our study is first to examine and quantify users' self-censorship of Twitter, using a systematic approach to label and process millions of Turkish tweets and users accounts. Additionally, we empirically analyze the coup's impact on government-censored topics, and the volume of government censorship, by comparing posts after the coup to posts we collected before the coup during the Turkish general election of 2015.

In our work, we try to answer the following questions:

1. Because of the political instability in Turkey after the failed coup, and the fear of government punishment or retribution, will we detect fewer government-censored tweets generated from inside of Turkey due to users' self-censorship?

2. Are there new government-censored topics post-coup? If so, are they related to the coup?

3. Can we systematically classify user accounts in our

dataset based on self or government censorship?

2 Related work: Censorship in Turkey

Tanash et al. [17] detected government-censored tweets by scraping Twitter for tweets containing a *withheld_in_countries* field in Twitter's API responses. They notably showed an order of magnitude higher censorship than disclosed in Twitter's "Transparency Reports."

Tanash et al. [16] also proposed methods for precisely extracting topics from censored Twitter data, using metrics on the social graph (e.g., user out-degree) as well as standard machine learning topic clustering algorithms. Applying their methods to censored tweets collected between late-2014 and mid-2015, they found that the Turkish government's most censored topics related to *Kurdish issues* and *government corruption*. In our work, we borrow from these methodologies to collect and identify censored tweets, then extract community based topics.

There are no known studies of social media self-censorship in Turkey, aside from some news reports [7], however, Arsan [3] studied censorship and self-censorship of the Turkish press. In his survey, he found that 96% of respondents agreed that journalists apply censorship when reporting on general interest topics, mainly due to internal pressure, and media owners' financial interest. As a result, 55% of the time, journalist do not convey important information that concern the general public.

3 Data

Twitter's public APIs allow external crawlers to follow specific users, search within geographic constraints, follow popular hashtags, etc. While Twitter makes it difficult to extract a full feed of *every* tweet, a more constrained search, such as in our work in Turkey, allows us to gather something much closer to a complete sample.

Pre-coup data: Previously, we collected 5.6M tweets during the June 2015 Turkish general election, for a duration of 24 days starting on 3 June 2015 with geo-parameters set to three major provinces in Turkey: Ankara, Istanbul, and Izmir. Using methodology similar to Tanash et al. [17], we expanded our sample by identifying users with at least one censored tweet who are not withheld-accounts, then crawling their timelines and their friends' timelines. Our overall sample contains 513,719 censored tweets.

Post-coup data: We streamed Twitter using the same geo-parameters collecting roughly 8.5M tweets posted by 342,650 distinct users across for 75 days. We began collecting this data roughly three hours after hearing news of the coup attempt on 15 July 2016.

3.1 Identifying Censored Tweets

To determine if a tweet is censored, we use the Tanash et al. [17] methodology. After collecting our tweets, we use the Twitter REST API to re-collect the same tweets by "id", then we examine if the *"withheld_in_countries"* field is present in the JSON structure returned by the API call. If the field is present, we check if its value equals *"TR"*, indicating censorship in Turkey. For the purposes of this study, we do not include tweets from *Withheld-Accounts* who are automatically censored by Twitter, as we could not distinguish which individual tweets were selected for censorship on targeted topics, and thus could not easily include them in our subsequent measurements and topics analysis.

4 Hypotheses

In this section, we propose several hypotheses concerning Turkish Twitter censorship.

4.1 Censorship Size

Hypothesis 1*: The dynamics of censorship shifted after the failed coup. We hypothesize that we will detect less censored tweets generated from inside of Turkey, perhaps because users will be less likely to tweet "sensitive" topics.*

Coups, by their nature, can be violent affairs. We would expect people to fear government retribution for tweeting opinions contrary to the ruling party. Therefore, many users might then switch their Twitter accounts to private mode, or deleted tweets or accounts entirely, resulting in a reduced volume of public tweets, and thus less content to be censored by the government.

Prior to the coup in 2015, we detected 513,719 censored tweets from non-withheld users by processing 5.6M tweets collected from Turkey. To test Hypothesis 1, we applied the same methods to our 8.5M tweets collected post-coup, and finding 142,492 distinct censored tweets from non-withheld users, which is **72% fewer censored tweets post-coup, compared to pre-coup**. Does this mean that people are tweeting less? Note that the 5,644,284 pre-coup tweets were streamed over 24 days, compared to the 8,543,856 post-coup tweets streamed over 75 days, as shown in Figure 1. Normalizing the counts into 24-days bins, we see 51% fewer streamed tweets during the first 24 days post-coup, and an estimated 43% decline in the overall collection relative to the Twitter volume we observed in 2015. We also noticed an incremental decline in the streamed volume in each consecutive periods; 2% decline in August, and 5% in September.

Investigating possible causes for this decline, we ruled-out the possibility of Twitter being throttled by the Turkish ISPs [6, 23, 19]. Network-level throttling apparently occurred during the initial days of the coup. Such network censorship, as well as the use of Tor to work around it, can be observed with Tor's "Directly Connecting Users Tor metric"[1], presented in Figures 2 and 3. These graphs show the volume of Turkish Tor users before and after the coup, with a clear *decline* in Tor usage afterward. This data is more consistent with users curtailing their use of Tor and Twitter, than with being thwarted from using them. Note that other popular censorship circumvention tools include Psiphon and Lantern, however, neither website provides accessible user metrics as Tor does, and as a result we decided to focus our attention on Tor.

Perhaps there was more Twitter data that we just did not see? In our collection, we identified only one rate-limited tweet[2] in our post-coup data. As best we can tell, we collected virtually every Turkish Twitter post from our post-coup time period.

Deutsche Welle [7] reported that the government made several arrests over social media posts praising the coup, and that "people are scared," and are "self-censuring themselves," fearing government punishment. Is the decline in our streamed data caused by users' self-censorship? To test this, we must classify all 342,650 users accounts in our post-coup data as either protected, deleted, suspended, or active with deleted-tweets.

1. We first identify tweets in our post-coup data that are no longer public. We do this by revisiting each tweet using the REST API[3]. From 8,543,856 tweets, we found that 19% of tweets, from 139,656 users, are no longer reachable by the API. This means that these tweets are either individually deleted, or their author's account is protected, deleted, or suspended.

2. Next, we label the 139,656 users accounts. We first check the status-code value by requesting *twitter.com/intent/user?user_id=xxx*, if the code is 400, the account is deleted by the user, otherwise we call the REST API by user-id, determining if the account is suspended, or active-protected.

Our results are summarized in Table 1. We found that 41% of users in our post-coup data, excluding suspended users, voluntarily removed 18% of all tweets either by switching their accounts to protected, deleting tweets, or deleting their accounts entirely[4]. Note that the largest set are "Active users" with some deleted tweets, followed by "Protected." We conclude that there are fewer government-censored tweets from inside of Turkey after the coup. This finding is consistent with Turkish people being afraid to speak in public, and consequently taking steps to hide prior speech and self-censor future speech.

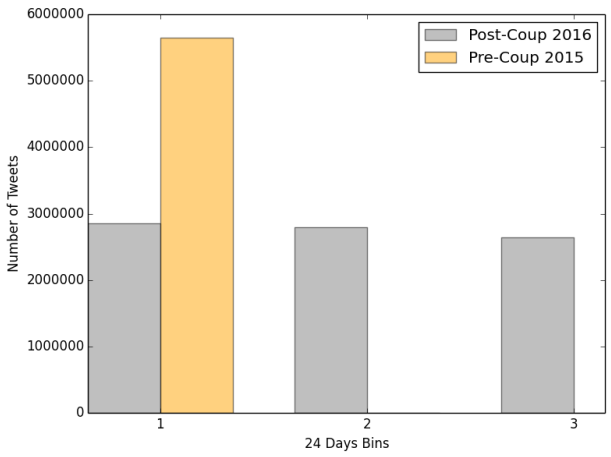

Figure 1: Volume of streamed Tweets post-coup 2015 vs. pre-coup 2016, 24 days bins.

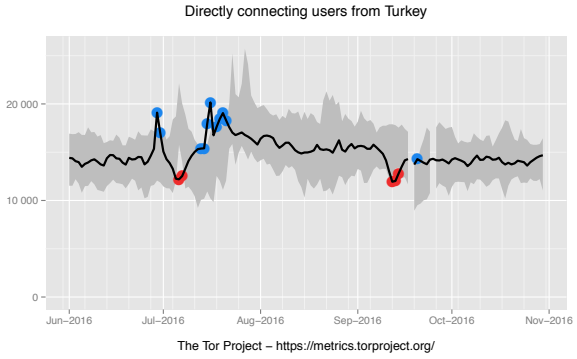

Figure 2: Daily Tor connection in Turkey - July 2016.

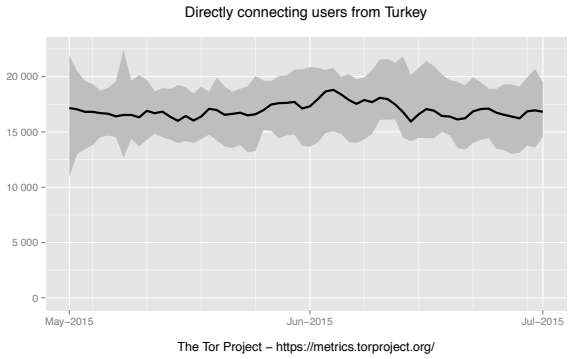

Figure 3: Daily Tor connection in Turkey - June 2015.

4.2 Censored Topics

In this section, we examine which topics are censored by the Turkish government post-coup. We start by proposing the following hypothesis.

	Tweets	%	Users	%
Total streamed tweets	8,543,856	-	342,650	-
Total unreachable tweets	1,582,632	19%	139,656	41%
Suspended users	29,971	2%	614	0.4%
Deleted users	200,523	13%	6,893	5%
Protected users	662,319	**42**%	30,612	22%
Active users with deleted tweets	689,819	**44**%	101,537	73%
Changed by user	1,552,661	18%	139,042	**40.6**%

Table 1: Post-coup user-accounts labeling results

Hypothesis 2*: Censored tweets post-coup will contain new topics related to the coup.*

Tanash et. al [16] established that the Turkish government censored two main topics; government corruption and Kurdish issues. To identify newer topics and influential users, we convert our 2016 censored tweets to a *Data-Flow-Influence* graph (as in Tanash et al. [16]), representing data communication between users. We define influential users as users with the highest out-degree who generate most original content, and whose tweets reached most users in their social graph, which is a reasonable proxy for a user's influence. Applying a graph modularity metric, a widely used metric for extracting network structures [22, 12], we extract user's communities, then manually examine the top influential users from each community. Figure 4 illustrates the results of our users-communities clustering and influential users.

Figure 4 shows two users' clusters. We asked a native Turkish speaker to manually examine the top influential users from each cluster. We found that the right clusters' (pink and red) influential users are mainly Gülen supporters, such as *"Yagizefe"*, and the left cluster's influential users are Kurdish users, such as *"ANF_TURKCE_ANF"*. To extract popular topics from each cluster, we first apply several steps to remove noise from our data and achieve more precise topics extraction, as in Tanash et al. [16]. This includes removing tweets from users with bot-like behavior, stripping URLs and stop-words, and then using the widely-used *zemberek* Turkish-language stemmer [2]. After this data tuning, we apply *tf-idf* and *NMF* [9], both standard machine learning algorithms for topic retrieval from documents and tweets [13, 17].

Our results show that the topics from each community cluster correspond to the profiles of the top influential users. The cluster with the Gülen supporters contained topics with Islamic references, references to popular Twitter Gülen supporters and Erdoğan critics, as well as specific terms that are clearly anti-government, e.g., *"torture", "human rights", "dictator", "arrests", "military", "Erdogan soldiers"* (originally Turkish, shown here in English). Our other cluster is clearly Kurdish activists. Several topics contained the string *"kurd"*, references to popular Kurdish Twitter usernames, including "red hack": a hactivist group. We similarly saw words like *"guns"* and *"fascism"*.

The Red Hack group of hactivists has drawn a lot of government attention. They're clearly visible in Figure 4 as "theRedHack". In September 2016, the Turkish government blocked Github, Google Drive, and other content sharing sites to specifically prevent the Red Hack group from leaking government secrets [14], which they threatened to do if Turkey did not release imprisoned Kurdish politicians [15].

In summary, our findings confirm our second hypothesis. The Gülen movement was not a topic of censorship prior to the coup, but afterward became a clear focus of Turkish government attention, both online and offline.

4.3 Self-censorship of Gülen topics

Turkish police began arresting people with ties to the Gülen movement, including social media posts praising the coup. Naturally, to protect themselves, we would expect people to both self-censor their old postings by voluntarily removing their old tweets via switching their accounts to protected mode, deleting their accounts, or deleting some tweets, as well as making strong public performances of their loyalty [7]:

Hypothesis 3*: Pro-Gülen tweets from before the coup will be self-censored.*

Hypothesis 4*: Public anti-Gülen tweets will occur more often post-coup.*

To test these, we conducted two experiments:

1. Topic Clustering: We first identified Gülen-related tweets using a bag-of-words approach against both the public and the unreachable tweets in our 2016 sample, excluding users who are suspended by Twitter. The results are summarized in Table 2. Notably, the unreachable rate for Gülen-related tweets is twice the background rate. Next, we extract popular topics from each set using *tf-idf* and *NMF*, asking a native Turkish speaker to label each topic's sentiment as either pro-Gülen, or anti-Gülen. We found **zero** pro-Gülen topics in the pub-

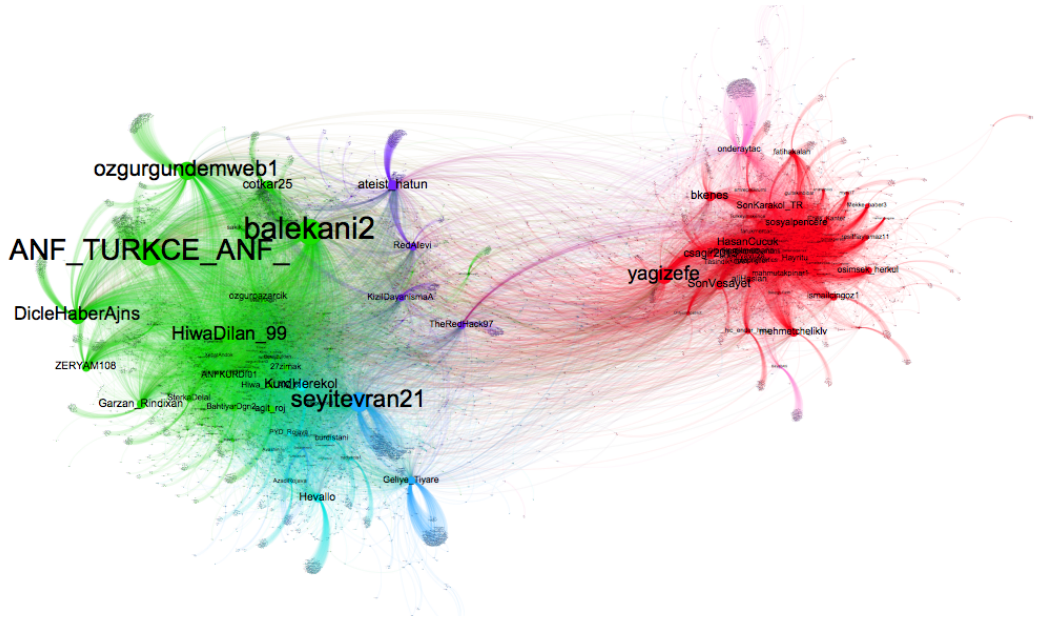

Figure 4: Users communities and influential users of censored tweets post-coup.

lic tweets, supporting Hypothesis 3. Conversely, we found 70% of the unreachable Gülen tweets were pro-Gülen, supporting Hypothesis 4.

Status	Unreachable	Public	Unreachable%
All tweets	1,582,632	6,961,224	23%
Gülen tweets	3,599	25,538	14%

Table 2: Gülen-related vs. all tweets, collected post-coup, counting how many remain public vs. unreachable.

2. Sample Labeling: We randomly sampled 40 tweets from each set, exclusively from the first three days post-coup, expecting to find more political content closer to the coup event. We asked a native Turkish speaker to classify the tweets as pro-Gülen, anti-Gülen, or neutral/unrelated. We summarize the results in Figure 5. As above, we found support for Hypothesis 3: pro-Gülen tweets appear **only** in the unreachable set, with none in the public set. We also found 40% more anti-Gülen tweets in the public set, supporting Hypothesis 4.

At this point, we have strong support for Hypotheses 3 and 4. What is less clear is whether some of the anti-Gülen discussion is "natural" or is a sort of public performance intended to defend the poster against otherwise baseless pro-Gülen accusations.

4.4 Self-Censoring Users

Self-censoring users are clearly responding to an external stimulus. They don't want to get caught up in the Turkish government's anti-Gülen witch-hunt, so

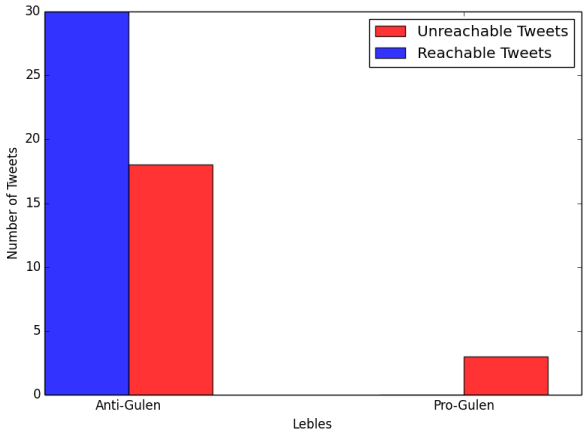

Figure 5: Labeling of Gülen related Tweets from public vs. unreachable tweets.

it's sensible that they would take steps to polish their public Twitter timeline. We might expect this sort of self-censorship to be performed broadly across all Twitter users, as opposed to just being the province of dedicated activists, who presumably would have a harder time hiding obscuring their political speech, whether online or elsewhere. This lead us to our final hypothesis.

***Hypothesis 5**: Self-censoring users' tweets are more likely to be politically neutral, for the most part, with only a few "sensitive" tweets.*

To test this hypothesis, we extract popular topics from all tweets posted by users who voluntarily changed their profiles status, or removed some of their old tweets. Examining the top 10 topics, we found that 9 of 10 topics contained neutral language, such as *"I love it very much"*. The only political topic contained two political terms: *"soldier"* and *"democracy"*, which clearly relate to tweets posted during the first week after the coup. Repeating the same experiment using only tweets from the active-users set who deleted some tweets, while keeping their profiles public, accounting for 44% of total unreachable tweet, we found similar results.

Next, we conducted a comparative analysis to quantify political and non-political tweets per user for each user category, and found that on average, active, protected, and deleted users, tweeted only 6%, 5%, and 6% political tweets, respectively. Figures 6 show this distribution for the top 200 users sorted based on the highest number of political tweets a user generated in our dataset for each of the user groups. All of these graphs demonstrate that these users largely do not discuss Gülen or other political topics, supporting Hypothesis 5.

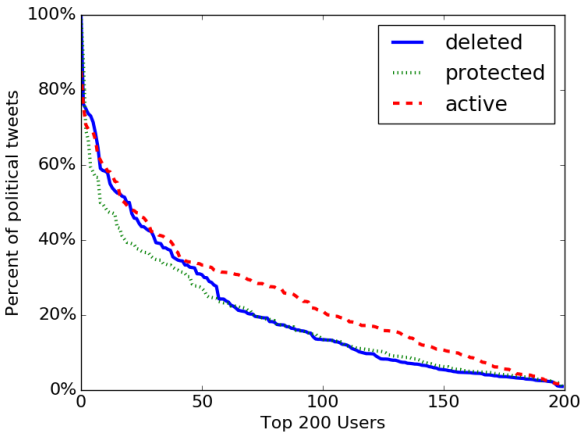

Figure 6: Percent of political tweets of the top 200 users, from the highest percent to the lowest

5 Twitter Transparency Report

In 2015, Tanash et al. [17] observed one order of magnitude more censored tweets than the data reported by Twitter's Transparency Reports. Similarly, we tried to compare the number of censored tweet in our post-coup dataset to the number reported in Twitter's July 1, 2016 – December 31, 2016 Transparency Report [20], in which Twitter reported 489 censored tweets in Turkey from non-withheld accounts. Our 142,492 censored post-coup tweets were posted between July 15 through November 1, 2016 from non-withheld accounts, of which 96% are retweets, as reported by Twitter's JSON metadata. Deduplicating these retweets, we identified 6,402 unique censored tweets, which contrasts with the 489 tweets reported by Twitter. As with Tanash, we find an order of magnitude more censored tweets than Twitter reports. Consequently, we caution other researchers from treating Twitter's reporting as a reliable source.

6 Conclusion

The 2016 coup attempt in Turkey provided us with an unusual opportunity to measure the impact of a singular event like this on both government-driven censorship as well as self-censorship. We were able to compare a dataset of 8.5 million tweets, collected in 2015, with a new dataset of 5.6 million tweets, collected in the immediate aftermath of the coup. Our data shows clear evidence that users are self-censoring their post-coup posts, particularly anything they might have said positively about Gülen, accused by the government of masterminding the coup. Similarly, they are limiting what they write going forward, with some evidence of users even deliberately writing anti-Gülen tweets, perhaps as a public performance of loyalty to the government.

Going forward, we note that social networks change in popularity over time, and Twitter may not always represent a reliable barometer of public opinion. Twitter is valuable for conducting this sort of research because it's easy to scrape and most content is public. Conducting similar research on Facebook or elsewhere, where user's default security settings limit their posts' visibility to their friends, would represent a significantly greater challenge, particularly without the social network's cooperation. If that cooperation cannot be assured, then other tactics, ranging from browser plugins to server-side apps, with their own issues of low user adoption, may become necessary to understand and measure this sort of application-level censorship as it occurs.

7 Acknowledgments

We would like to acknowledge an individual who prefers to remain anonymous, for his Turkish translation effort and helpful discussions. We also thank our reviewers for their comments. This work is supported in part by NSF grants CNS-1409401 and CNS-1314492.

References

[1] Gülen Movement – Wikipedia, the Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=G%C3%BClen_movement&oldid=776204051, 2017.

[2] A. A. Akın and M. D. Akın. Zemberek, an open source NLP framework for Turkic languages. *Structure*, 2007.

[3] E. Arsan. Killing me softly with his words: Censorship and self-censorship from the perspective of Turkish journalists. *Turkish Studies*, 2013. http://www.tandfonline.com/doi/abs/10.1080/14683849.2013.833017.

[4] J. Bohannon. Turkish academics targeted as government reacts to failed coup. *Science Magazine*, 2016. http://www.sciencemag.org/news/2016/07/turkish-academics-targeted-government-reacts-failed-coup.

[5] M. Burke. Turkey has a history of military coups. *USA Today*, 2016. https://www.usatoday.com/story/news/world/2016/07/15/turkey-military-coup-history/87153106/.

[6] C. Daileda. Twitter throttled in Turkey amid attempted coup. *Mashable*, 2016. http://mashable.com/2016/07/15/turkey-facebook-twitter-youtube-blocked-attempted-coup/#5qXZJ6HVokqC.

[7] E. Grenier. Erdogan and social media: Use and abuse. *Deutsche Welle*, 2016. http://www.dw.com/en/erdogan-and-social-media-use-and-abuse/a-19413205.

[8] M. Gurcan and M. Gisclon. From autonomy to full-fledged civilian control: The changing nature of Turkish civil-military relations after July 15. *IPC-Mercator Policy Brief*, 2016. http://ipc.sabanciuniv.edu/wp-content/uploads/2016/08/IPM_PolicyBrief15July_12.08.16_web.pdf.

[9] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 1972.

[10] P. McGeough. Who is Fethullah Gülen? *City Journal*, 2012. https://www.city-journal.org/html/who-fethullah-g%C3%BClen-13504.html.

[11] P. McGeough. Turkey coup: Who is Fethullah Gulen and why is Recep Tayyip Erdogan obsessed with him? *The Sydney Morning Herald*, 2016. http://www.smh.com.au/world/turkey-coup-who-is-fethullah-gulen-and-why-is-recep-tayyip-erdogan-obsessed-with-him-20160716-gq76m7.html.

[12] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 2006. http://www.pnas.org/content/103/23/8577.short.

[13] F. Å. Nielsen. Clustering of scientific citations in Wikipedia. *arXiv preprint arXiv:0805.1154*, 2008. https://arxiv.org/abs/0805.1154.

[14] D. Pauli. Turkey blocks Drive, GitHub, OneDrive in bid to kill RedHack leaks. https://www.theregister.co.uk/2016/10/10/turkey_blocks_drive_github_onedrive_in_bid_to_kill_redhack_leaks/, 2016.

[15] E. K. Sozeri. RedHack leaks reveal the rise of Turkey's pro-government Twitter trolls. *Daily Dot*, 2016. https://www.dailydot.com/layer8/redhack-turkey-albayrak-censorship/.

[16] R. S. Tanash, A. Aydogan, Z. Chen, D. S. Wallach, M. Marschall, D. Subramanian, and C. Bronk. Detecting influential users and communities in censored tweets using data-flow graphs. In *Proceedings of the 33rd Annual Meeting of the Society for Political Methodology (POLMETH)*, Houston, TX, 2016.

[17] R. S. Tanash, Z. Chen, T. Thakur, D. S. Wallach, and D. Subramanian. Known unknowns: An analysis of Twitter censorship in Turkey. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society*, pages 11–20, Denver, CO, 2015.

[18] TurkeyPurge. *US Human Rights Report: Tens of Thousands Jailed in Turkey with Little Clarity on Charges*, 2017. https://turkeypurge.com/us-human-rights-report-tens-of-thousands-jailed-in-turkey-with-little-clarity-on-charges.

[19] Twitter. *Intentional Slowing of our Traffic in Country*, 2016. https://twitter.com/policy/status/754072148289789952.

[20] Twitter. *Transparency Report July-December 2016*, 2016. https://transparency.twitter.com/en/removal-requests.html#removal-requests-jul-dec-2016.

[21] U.S. Department of State, Bureau of Democracy, Human Rights and Labor. *Turkey 2016 Human Rights Report*, 2016. https://www.state.gov/documents/organization/265694.pdf.

[22] Wikipedia. Modularity (Networks). https://en.wikipedia.org/w/index.php?title=Modularity_(networks)&oldid=751888052, 2016.

[23] J. G. Wong. Social media may have been blocked during Turkey coup attempt. *The Guardian*, 2016. https://www.theguardian.com/world/2016/jul/15/turkey-blocking-social-facebook-twitter-youtube.

[24] W. Worley and H. Cockburn. Prime Minister says 265 people killed in attempted military coup, including at least 100 'plotters'. *Independent UK*, 2016. http://www.independent.co.uk/news/world/europe/turkey-coup-dead-erdogan-military-chief-ankara-istanbul-death-toll-plotters-how-many-killed-wounded-a7140376.html.

Notes

[1] Tor Metrics: https://metrics.torproject.org

[2] Twitter includes a special *limit* field in its streamed JSON data that indicates the number of skipped tweets.

[3] If requesting a tweet by id from the REST API returns an empty string or error, it means that the tweet is no longer reachable.

[4] Twitter deletes accounts after 6 months of inactivity (https://support.twitter.com/articles/15362). Our initial and follow-on samples were at most five months apart, avoiding this issue.