Known Unknowns: An Analysis of Twitter Censorship in Turkey

Rima S. Tanash
Rice University
6100 Main St.
Houston, TX 77005
rima.s.tanash@rice.edu

Zhouhan Chen
Rice University
6100 Main St.
Houston, TX, 77005
zc12@rice.edu

Tanmay Thakur
University of Houston
4800 Calhoun Rd.
Houston, TX, 77004
tathakur@uh.edu

Dan S. Wallach
Rice University
6100 Main St.
Houston, TX, 77005
dwallach@rice.edu

Devika Subramanian
Rice University
6100 Main St.
Houston, TX, 77005
devika@rice.edu

ABSTRACT

Twitter, widely used around the world, has a standard interface for government agencies to request that individual tweets or even whole accounts be censored. Twitter, in turn, discloses country-by-country statistics about this censorship in its transparency reports as well as reporting specific incidents of censorship to the Chilling Effects web site. Twitter identifies Turkey as the country issuing the largest number of censorship requests, so we focused our attention there. Collecting over 20 million Turkish tweets from late 2014 to early 2015, we discovered over a quarter million censored tweets—two orders of magnitude larger than what Twitter itself reports. We applied standard machine learning / clustering techniques, and found the vast bulk of censored tweets contained political content, often critical of the Turkish government. Our work establishes that Twitter radically under-reports censored tweets in Turkey, raising the possibility that similar trends hold for censored tweets from other countries as well. We also discuss the relative ease of working around Twitter's censorship mechanisms, although we can not easily measure how many users take such steps.

1. INTRODUCTION

Twitter, the popular microblogging platform, plays an essential role for communicating news and current events among professional reporters, human-rights activists, and oppressed citizens across the globe. More notably, Twitter was widely used to disseminate news and opinion during a series of uprisings in the Middle East, also known as the Arab Spring, resulting in the overthrow of dictatorships in Egypt, Tunisia, Libya [31][34], and the Taksim Gezi Park protests in Turkey on May 2013 [30].

On January 27, 2012, Twitter announced a new censorship policy known as *"Country-Withheld Content"* by which Twitter enabled governments and their representatives to formally request that Twitter withhold tweets and/or whole accounts within the boundaries of a specific country [28]. Once Twitter receives such a request, it somehow determines whether the request is lawful and then withholds the tweet in question in that specific country, while still allowing it to be visible elsewhere. Twitter claimed that this policy was a business decision that allowed Twitter to exist in parts of the world that have different ideas of freedom of expression, and that this policy will prevent locally offensive content [29]. The policy received criticism that it was nothing more than a form of government censorship and a threat to freedom of speech [15]. Unsurprisingly, trending hashtags like *#TwitterCensored* and *#twitterblackout* followed this announcement, where many users expressed their outrage. Notably, Twitter also announced a partnership with Chilling Effects to publish withheld content "unless they are legally prohibited from doing so" [28]. Twitter also notes, in their "Withholding Transparency Reports" that the reported data is neither 100% comprehensive nor complete. For example, the Politwoops website, dedicated to collect deleted tweets from politicians—often a source of raw and embarrassing statements—announced that Twitter disabled their API feed on June 4, 2015. Twitter claimed that the company violated their developer agreement [14]. Politicians deleting their own tweets is obviously not the same thing as a country censoring its citizens, but all the same, Twitter has demonstrated that it is opposed to external organizations displaying content that is not visible on Twitter itself.

Given this, the research question is easy to pose: how much censorship is really happening on Twitter? How much is not being reported to Chilling Effects? How much is not being reported on Twitter's withholding transparency disclosures? And can we determine anything about the machinery behind the censorship? Are tweets being withheld one by one, based on individual requests to Twitter by foreign governments, or are they being withheld in large groups, perhaps based on hashtags or other keywords? We suspect that there is undisclosed censorship on Twitter. We just do not know the depth of the unknowns.

2. RELATED WORK

Chen et al. described that in recent years, social media has risen in prominence in many countries. In China, social media such as Weibo and Renren plays an important role as a platform for breaking news and political commentary outside of the confines of state-controlled news media. However, like all websites in China, Chinese social media is subject to censorship. The magnitude of censorship varies dramatically across topics, with 82% of posts in

some topics being censored. The paper also finds that censorship of a topic correlates with high user engagement, suggesting that censorship does not stifle discussion of sensitive topics. Furthermore, the authors find that users create variants of words (known as morphs) to avoid keyword censorship [5].

Florio et al. confirmed that in 2014 the Turkish government hijacked DNS traffic to censor users traffic. Users of traditional computers were able to circumvent censorship using TOR and VPNs. Unlike traditional users, mobile users used a specialized Android application designed to allow DNS configurations of mobile devices to circumvent G/4G service-provider censorship. The application was installed by 130,000 devices by August 2014. The researchers collected data that showed that some censorship activities began after the lift of the Twitter ban on Turkey in March of 2014. However, they did not examine the type of data being censored [10].

Aase et al. argued that motivation, resources, and time are three major elements in the application and potential uses of Internet censorship. The challenge for the online censorship research community is to develop tools for measuring these three elements explicitly when conducting measurement studies [1].

Winter et al. proposed techniques to measure and circumvent Internet censorship that they deployed in three countries. The techniques successfully bypassed the Great Firewall of China [35].

Morrison examined the feasibility of automating the detection of censorship in microblogs without using sensitive keywords but using social network graphs properties and communication flow, and found his automated detection methods feasible when studying Sina Weibo [18].

In recent years, a large number of papers have focused on censorship resistance schemes (CRSs). Khattak et al. proposed an attack model to comprehensively explore censorship capabilities and and developed an evaluation framework to test each CRS's flexibility [13].

No previous work has been attempted to quantify Twitter's internal censorship mechanisms.

3. TWITTER'S CENSORSHIP MECHANISMS

To better understand the steps Twitter uses to implement its *"Country-Withheld Content"* policy, it is important to examine both its Transparency Reports and the Chilling Effects website that Twitter uses to publish government withholding notices. The Chilling Effects is an independent third party archiving service that publishes cease and desist requests and other related legal demands [32]. Many companies such as Google and Facebook use Chilling Effects as a method for being more transparent about how they handle censorship-related requests (and, perhaps, as a way of disincentivizing those who might wish to issue them various forms of legal demands).

3.1 Transparency reporting

When Twitter announced its censorship program, they noted: *"...we have expanded our partnership with Chilling Effects to publish not only DMCA notifications but also requests to withhold content—unless, similar to our practice of notifying users, we are legally prohibited from doing so."*[28] Twitter started publishing Transparency Reports in January 2012, with data grouped into 6-month bins on a per-country basis. We summarize 18 months of this data, from January 2012 to June 2013 in Table 1. The bulk of the censorship appears to occur in three countries: Brazil, France, and Russia. A request may specify several user accounts and/or tweets to be withheld, so the number of actual tweets withheld does not necessarily map one-to-one with the number of requests or the number of accounts specified. For example, in the case of Brazil,

Country	Number of Requests	Users Account Specified	Number of Accounts Withheld	Number of Tweets Withheld
Australia	3	3	0	0
Brazil	26	33	1	39
Canada	3	3	0	0
Ecuador	1	1	0	0
France	5	56	0	56
Germany	6	11	4	0
Greece	2	2	0	0
India	4	18	0	3
Indonesia	2	2	0	0
Japan	3	7	0	6
Korea	1	1	0	0
Netherlands	2	2	0	5
Pakistan	1	1	0	0
Russia	17	17	4	8
Spain	2	4	0	0
Turkey	14	46	0	0
UK	9	33	0	0
US	6	23	0	0
Venezuela	1	1	0	0
Total	108	264	9	117

Table 1: Aggregate Transparency Reports disclosed by Twitter for 18 months (Jan. 2012 - Jun. 2013) from all countries.

Twitter withheld 39 tweets and one account, but the report does not specify if the tweets were generated by the withheld account.

From our data collection, it appears that a request to withhold an account causes all tweets from that account to be withheld. In such cases, Twitter appears not to include these tweets from withheld accounts in the overall count of withheld tweets for a given country in its transparency reporting. We verified this manually for Brazil and for Germany, but have not systematically looked at each and every county with withheld accounts.

Our initial effort was to collect as many Twitter-related reports from the Chilling Effects database as we could find. This would enable us, for example, to learn Twitter handles, hashtags, and other sensitive keywords around which we might later build automated searches. Upon initial examination, we quickly discovered how incomplete the Chilling Effects database appears to be. We found a grand total of 33 notices posted to Chilling Effects, across all countries, which is far fewer than just the 108 account withholding requests disclosed by Twitter's own transparency reports. Clearly, the Chilling Effects database is nowhere near a comprehensive disclosure of Twitter's stream of withholding requests.

Despite these shortcomings, we found that the Chilling Effects postings disclosed a fair bit of information, including the stated reason for the request as well as the user and identifier of the tweet being censored. We show a sample censorship request in Figure 1. In 2014, Twitter changed their procedures and began hiding some information about the government agencies requesting the censorship (see Figure 5 for a sample Turkish notice). To build our database of Chilling Effects notices, we converted PDF to RTF and manually extracted tweets, user-names, and other fields. This was sufficiently robust, across multiple languages, that it served our needs.

3.2 The "Country_Withheld" process

As described before, Twitter allows several mechanisms for countries to request that tweets be censored, including email, a web
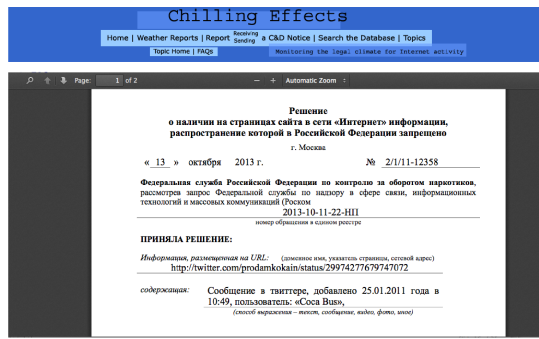
Figure 1: Sample withholding request from Russia, retrieved from Chilling Effects in 2013.
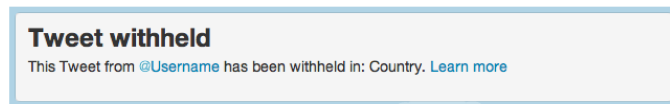

Figure 2: Withheld tweet as it appeared from inside the censoring country.

Figure 3: Withheld Account as it appeared from inside the censoring country.

(Russia)

(USA)

Figure 4: Comparison view of a withheld Russian tweet from Russia (above) and USA (below).

form, and hardcopy notices [29]. If Twitter agrees with the request, the withheld tweet will then be greyed-out with a message: "This tweet from @username has been withheld in Country Name" (see Figure 2). The user may or may not receive a notice that their tweet was being censored, but the censored tweet is generally visible from outside of the censoring country. Twitter offers a similar mechanism to censor an entire user, rather than just an individual tweet (see Figure 3).

3.3 How Twitter appears from different countries

To view how Twitter looks from different countries, we used Tor as a proxy service, allowing us to choose the country of origin for our traffic by selecting different Tor exit nodes. As an example, Figure 4 shows how a Russian-withheld tweet appears respectively from Russia and the US.

3.4 Twitter APIs

Twitter offers a large collection of public REST and Streaming APIs [1]. All API responses return JSON structures; we present an abbreviated response below:

```
...
"id": NumberLong("51707...
"favorite_count" : 703,
"retweeted" : false,
"coordinates":null,
"source":"<a href=\http://twitter.com\...
"in_reply_to_screen_name":null,
```
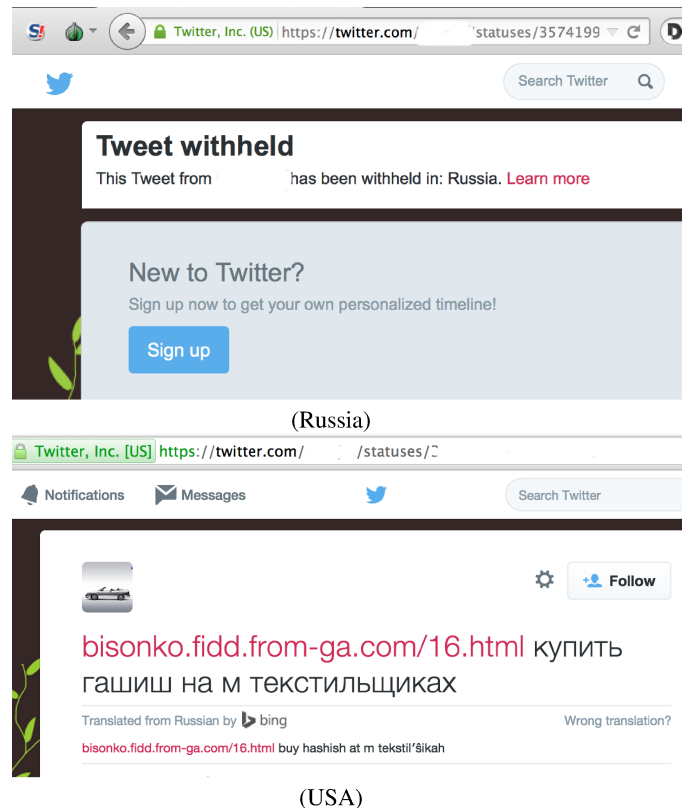
---

[1] **Twitter API** https://dev.twitter.com/docs/api/1.1

```
"id_str" : "51707756700291...
"retweet_count":1828,
"in_reply_to_user_id":null,
"favorited":false,
"withheld_in_countries":["TR"]
"user" : {
...
```

When a tweet is withheld, the corresponding REST API JSON response will include a field called "withheld_in_countries" with a list of two-letter ISO country codes where the tweet is to be censored; otherwise the field is absent. We inspected *all* the collected tweets from Chilling Effects by using Tor with a variety of exit nodes, and validated that the "withheld_in_countries" field is present in the JSON structure if the tweet is withheld in the target countries. We note that *we get the same response, regardless of the country from which we make the request*. This essential observation means that we can collect tweets from our home servers without needing to use any sort of in-country proxy servers. When a user's home country happens to have withheld a tweet, their users won't see it. Figure 2 shows how withheld tweets are rendered. (See also, Section 7 for ramifications on this observation toward bypassing Twitter's censorship mechanisms.)

3.5 Withheld accounts

The Twitter documentation states that when an account is withheld, the "text" field will be assigned a specific value with a specific structure. For example: "text": "@chatty's account is withheld in: Greece, Hong Kong, Malaysia", and that the field "scope" will be assigned the value "user". We found this to be rather inconsistent

13

and in some cases invalid. Throughout our experiments, we found that 72% of withheld accounts that we identified did not have the "user":"scope" field format, but still generated tweets containing a "withheld_in_countries" field, rendering them indistinguishable from those generated by non-withheld accounts. In addition, the "text" field had the original tweet string being withheld, contrary to what the documentation suggested.

Ultimately, the only reliable method we discovered to determine whether an entire account is withheld, rather than just individual tweets, was to simply enumerate all of that user's tweets. If they were all withheld, that was a reliable signal.

4. CASE STUDY: TURKEY

While it would be desirable to collect all withheld content, worldwide, the sheer volume of this would be impractical. Alexa reported that by January 2013 Twitter had 500 million registered users and that the service generates 500 million tweets daily [19]. Twitter enforces rate limits that make it impractical to collect this much data from their service, even when using the obvious bag of tricks (multiple Twitter accounts, multiple IP addresses, etc.—See Section 5.1 for additional details).

Instead, we decided to focus our study on one country with a reputation for censorship. Which one? We selected Turkey, due to its apparently vigorous use of Twitter's censorship mechanism and its generally hostile behavior toward journalists and dissenting political speech [6]. Also, we have access to native speakers to assist us when automated translation falls short.

Notably, in May 2013, the Taksim Square riots [24] led to an "Occupy Gezi" movement (Gezi is a park next to Taksim Square, in Istanbul). Following this movement, Twitter achieved significant popularity in Turkey, gaining over a million new accounts [21].

Subsequently, in March 2014, then-Prime Minister (now President) Recep Tayyip Erdogan ordered Twitter blocked in Turkey for Twitter's failure to implement Turkish court orders seeking removal of some links posted on Twitter. He also demanded that Twitter establish an office in Turkey to ease take down requests and improve Twitter's accountability under Turkish laws. Perhaps unsurprisingly, Twitter declined [23]. During this event, Twitter instructed its users (via tweet) to continue tweeting via SMS. The hashtag #twitterisblockedinTurkey was trending amongst protester and their followers. Later that month, Turkey's highest court rejected the ban as a violation of freedom of expression [36].

In August 2014, a news report claimed that Turkey and Twitter were scheduled to meet for the third time to discuss the establishment of a new office for Twitter representatives in Istanbul [2]. Clearly, negotiations were afoot so Turkey could keep Twitter around, and Twitter could accommodate Turkey's censorship needs.

It would be useful to estimate the volume of tweets per year in Turkey. According to Fox [11], 3.0% percent of Twitter's active users are in Turkey. Sysomos [27] doesn't list Turkey anywhere in its list of top countries although Baronchelli et al. [17] found Turkish to be the 10th most popular language on Twitter in 2013. Twitter does not disclose any such data itself. Ultimately, this means that any measurements we make of Twitter censorship can only be treated as a *lower bound* on the total volume of Twitter censorship. We unfortunately have no way to assure that our data collection will be a *representative sample* of the total traffic.

4.1 Disclosed censorship

Starting in 2014, we noticed an increase in the number of Turkish requests posted to Chilling Effects. In addition, the Twitter Transparency Reports published in 2014 showed an increase in withheld

Transparency Report Date	Number of Withheld Tweets in Turkey
2012: Jan 1 - Jun 30	0
2012: Jul 1 - Dec 31	0
2013: Jan 1 - Jun 30	0
2013: Jul 1 - Dec 31	0
2014: Jan 1 - Jun 30	183
2014: Jul 1 - Dec 31	1820

Table 2: Distribution of withheld tweets in Turkey, as reported by Twitter.

Turkish notices following the unblocking of the Twitter service; Twitter itself reports 183 withheld tweets and 17 withheld accounts in the (Jan 1,2014-June 30, 2014) report, and 1820 withheld tweets and 62 withheld accounts in the (July 1, 2014-December 31, 2014) report. Table 2 shows the distribution of number of withheld tweets by reporting period. Clearly, once Twitter was no longer blocked to Turkish citizens, the Turkish government availed itself of Twitter's censorship mechanisms.

4.2 Collecting censored tweets

As we described previously, Twitter posts withholding notices to the Chilling Effects website. A sample Turkish notice is shown in Figure 5. This scanned document, and many more like it, is blurry and partially redacted by Twitter. We extracted all the Turkish Chilling Effects documents posted through March 30, 2015, extracted the tweet IDs from the notices, fetched the tweets with Twitter's REST API, and stored them in a local database. Overall, we identified 2896 tweets from Turkey with this method. Some tweet IDs appeared in multiple notices, so we removed duplicates as well as some apparently malformed responses, ultimately ending up with 2,473 unique tweets, of which 1,340 were still present on Twitter. The remaining tweets were either removed or were perhaps associated with "protected" users, whose tweets are normally only visible to permitted users rather than the whole world; for these tweets, each user would need to grant us permission to see their tweets in order for us to confirm their withholding status. (We decided not to pursue such permissions.) Of these remaining tweets, we confirmed 1,155 withheld Turkish tweets by examining the "withheld_in_countries" JSON field; this also shows that *at least* 86% of the Turkish government's withholding requests for non-protected tweets were approved by Twitter. We cannot estimate the approval rate for "protected" tweets, but assuming their actions on "protected" accounts are consistent with their actions on "public" accounts, we can confirm that Twitter seems to approve most of the withholding requests that it receives from Turkey.

As an interesting aside, it's worth posing a question we cannot answer: how is the Turkish government managing to censor anything from "protected" users? These tweets are not visible to the world yet they appear in withholding requests. This implies several possibilities. Perhaps the censors are requiring keyword or hashtag-based censorship. Perhaps they're demanding access to protected accounts. Or, perhaps the simplest answer is that these user accounts were once public but are now protected. We have inadequate information to determine what happened.

4.3 Collecting censored accounts

As above, we wish to identify censored user accounts, not just censored tweets. We found a total of 80 Chilling Effects requests for user accounts to be censored. Of these, 40 accounts appear to
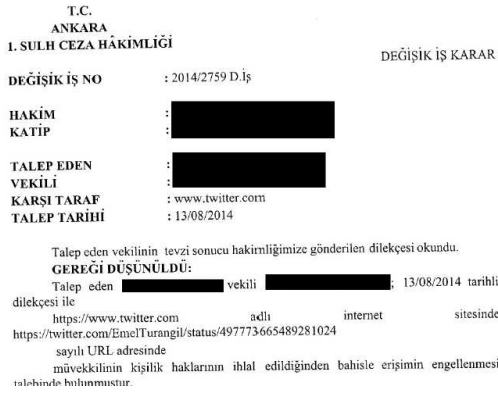
Figure 5: Sample of Withholding Notice from Turkey- Chilling Effects extracted in 2015.

be withheld in Turkey. Of the remaining 40 accounts, 23 accounts appear to be suspended or deleted.

5. BROADER TURKISH DATA COLLECTION

At this point, we hypothesized that the visible Turkish censorship of Twitter was just the tip of the iceberg. To quantify the degree of censorship, we would need to collect much more data.

5.1 Data sources

There are many ways of acquiring or purchasing large-scale datasets of tweets:

The Twitter/Gnip Firehose.

The "firehose" is a special API, only offered to paid customers, that returns the *entire* live stream of tweets, as they occur. In addition, it allows historical retrieval of old tweets from an archive including those that are deleted. We contacted Gnip, a third party tweet-reseller that was recently acquired by Twitter. We received a costly quote of $12,000 per 1 million tweets. They explained that we could purchase Turkish tweets by filtering on the language. However, they could not guarantee to find all tweets containing the "withheld_in_countries" field as censorship may occur after the tweet was captured. They requested that we complete special forms disclosing our research goals to obtain an internal approval from Twitter. We declined to pursue this relationship.

Twitter Free Public APIs.

Many researchers, like us, are ultimately forced to use Twitter's free public Streaming and REST APIs. These only return 1% of the total public stream, with no particular explanation of how the 1% are sampled from the overall population of tweets [19]. Furthermore, these APIs allow only 180 calls per 15 minute interval and requires that we register with a set of OAuth credentials. Ultimately, we procured many sets of these credentials, allowing us to at least partially overcome these rate limits, but we would certainly be unable to fetch every single tweet, much less revisit selected users every few minutes as Zhu et al. [37] did while looking for censorship on Weibo.

Morstatterer et al. [19] found that the volume of tweets obtained using Twitter Streaming API, versus Firehose, depended on the coverage of the streaming API data. The more complete the filtering criteria is, the less coverage we might get. More usefully, they found that when the geo property was used, the coverage was

almost complete. Consequently, this was the route for data collection that we pursued:

- Sampling random Turkish tweets using geo coordinates

- Follow Turkish "sensitive" users

5.2 Data collection

Our general methodology was to collect large volumes of tweets, using these free mechanisms, then revisit them occasionally to discern if any had been censored. In contrast with Zhu et al. [37]'s goal of identifying how fast censorship occurs, we were more interested in collecting as many censored posts as possible. Given the manually intensive review process that Twitter appears to enforce on governments, there does not seem to be any useful information in more precise timing, while volume measurements are still quite valuable.

Phase I: Streaming API using Turkish geo coordinates.

The Twitter *POST statuses/filter Streaming API* includes an optional parameter called "*location*" that takes a set of geo bounding boxes (latitude/longitude) to stream tweets geographically. These appear to be set by Twitter's smartphone clients, although we did not make a detailed examination. Instead, we queried with geobounds corresponding to three major Turkish cities: Izmir, Ankara and Istanbul.

We ran our streamer from October 2014 through January 2015, and collected 17 million tweets.

Phase II: Revisiting Tweets, looking for censorship.

Since there is a time gap between sending the government-request and actually withholding the requested tweets by Twitter, we decided to wait before re-inspecting the collected tweets to see if they became withheld. In February and March 2015, we revisited the previously collected 17 million tweets using the REST APIs, and found 3,258 withheld tweets. Our data clearly shows that there are far more censored tweets than 1,155 we found on Chilling Effects. We also observed that we managed to capture some censorship events of tweets prior to our phase I data collection. These corresponded to users we were following (more on that below), so they are not a representative sample of censorship events, but they are an existence proof of censorship events reaching well into the past.

Phase III: Friends of sensitive users.

In some cases, Twitter's APIs will not only return a stream of tweets from a set of users being queried, but will also return replies, mentions, and retweets. Since followers and friends of censored users are perhaps more likely to be censored themselves, this means that, like Zhu et al. [37], we can spider outward from a small set of known-censored users and derive a larger set of "interesting" users, then collect all of their tweets. Following this process, with our original set of censored tweets as a baseline, we ultimately collected 689 unique user IDs that have been subject to withholding. Ultimately, we collected almost 1.7 million tweets (i.e., an average of 82 tweets per day per user; these are very active tweeters) from these users in March 2015. Of these, 46,769 were withheld.

Repeating the process again, we expanded to the followers of this larger set of censored users, ultimately yielding nearly 85,000 user IDs. We then scanned for each user's tweets, ultimately yielding a total of 171,652 withheld tweets. Curiously, 386 of those tweets were not from Turkey, with the bulk coming from a Brazilian user,
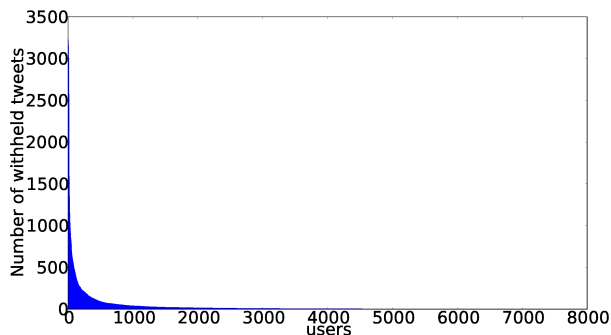
Figure 6: Distribution of number of tweets per user.

Figure 7: Distribution of number of tweets per user (log-log scale).

unhappy at having been fired from his job, he claimed, due to his sexual orientation. We also found 3 Russian tweets and 13 whose country was listed as "XY", which we understand to refer to tweets removed for DMCA violations, along with a corresponding "withheld_copyright" field. Removing all of the non-Turkish tweets, we ultimately found 171,266 withheld tweets in Turkey.

We also looked at censored retweets, leading us to almost 3000 new user IDs and yielding an additional 45,114 withheld tweets beyond the 171,266 above.

Altogether, we discovered 266,407 censored tweets (3258 + 46769 + 171266 + 45114).

5.3 Analysis

Using these methods, as discussed above, we discovered a grand total of 266,407 censored tweets generated by 7,642 distinct users. Of these, we estimate that 46 of the user accounts are, themselves, withheld in their entirety. This represents two orders of magnitude greater censorship of tweets in Turkey than disclosed by Twitter through its Transparency Report and the Chilling Effects website.

This difference is no statistical anomaly. Twitter is deliberately underreporting the volume of withholding in Turkey, perhaps as a result of legal requirements made of it by the Turkish government as part of allowing them to continue operating in Turkey. No such agreements have been publicly disclosed.

We next consider two simple analyses over our data: tweet censorship volume per user and deduplication. We apply machine learning techniques toward topic analysis in Section 6.

5.3.1 Censorship volume per user

We extracted the user accounts with the largest number of withheld tweets, and manually examined their profiles. We found that nine of the ten users tend to tweet political anti government content, and appeared to be influential given the large number of their followers, and one account tweeted hacking information and appeared non influential. Figure 6 shows the probability distribution function (PDF) of withheld tweets by user, where the users are sorted from the most withheld to the least withheld. The distribution is heavily skewed to the left, indicating that a small number of users generated a large number of withheld tweets.This shows that the top 100 censored users generate 44% of the censored tweets, and the top 500 censored users generate 73% of the censored tweets. This observation also corresponds to our withheld account vs. non-withheld account observation, in which withheld accounts are influential accounts tweeting frequently with their whole accounts being withheld. The non-withheld accounts are normal persons who sometimes tweet withheld content.
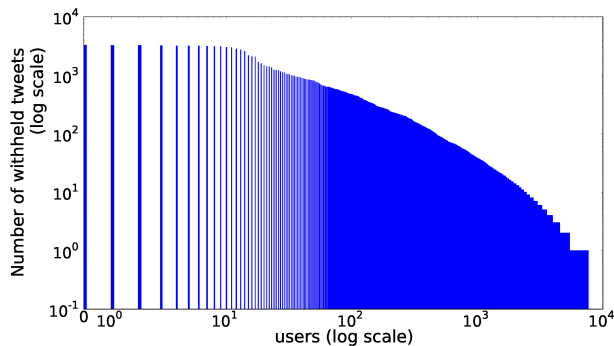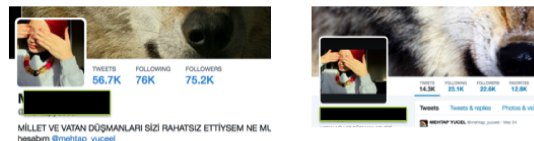

Figure 8: Same user with two profiles withheld in Turkey.

The long-tailed aspect of the censorship distribution (see Figure 7) has important ramifications for understanding Turkey's Twitter censorship. If we saw a short tail, then this would imply that Turkey was focusing its attention on the biggest rabblerousers, but instead we see a broad focus over a large population of users. This also implies that Turkey's censors are using search tools to discover these long-tail tweets that would not otherwise rise to their attention via organic means such as having been posted by celebrity users or having gone viral via retweeting.

The long-tail also implies that popular users cannot simply discard their identities and start over to fly below the radar. As an example of this, we manually dug deeper into our data to identify censored users. Users clearly become aware that they are being censored, then create new accounts which again suffer censorship. Figure 8 shows one example of a withheld user with two accounts and many followers. The older account seems to be the one with over 75k followers while the newer one has 12k followers.

5.3.2 Duplicate text analysis

An interesting question, now that we have established such a large number of withheld tweets, is the extent to which they duplicate one another. Our prior analysis considered tweets with unique IDs, but users can "retweet" one another's tweets, either through official mechanisms (which generate metadata headers) or by simply cutting-and-pasting the text from an old tweet into a new tweet, perhaps prepending the text with in-line metadata, typically "RT @" and the username of the original author.

Stripping out the "RT" prefixes, ignoring retweet metadata (both inline and in the rest of the JSON structure), and focusing strictly on the message body text for the 266,793 withheld tweets, we ultimately yielded 88,276 unique strings. In other words, we found 178,517 retweets of various styles. Digging a little deeper, it appears that all but 15 of them originated with the standard "retweet" button, with those 15 being the result of cutting-and-pasting raw text.

We extracted the top 10 most duplicate tweets and found that all 10 tweets had bad things to say about Recep Tayyip Erdogan, the

current Turkish president, and appeared to be generated by supporters of *Fethullah Gulen*[2].

Additionally, we looked if there are users that reposted or retweeted the same thing more than once; we found 295 users posted a total of 1,503 tweets in this category. We manually reviewed the top-ten most prolific users in this set. Two users regularly post anti-government topics and have large number of followers, suggesting that they are influential. The third account appeared to be a marketing bot for a software product; it has a low number of followers, but generates a large number of tweets.

We finally ask the question of whether censorship of a retweet has any bearing on censorship of the original. We note that the JSON structure for a retweet contains the ID of the original, so we collected the withholding status of each original tweet for every withheld retweet. We ultimately discovered that 92% of these original tweets are also withheld. What about the remaining 8%? Half of them survived, uncensored, while the other half belong to withheld accounts (see Section 3.5). This suggests that, through whatever mechanism Turkey is directing its remarkable volume of Twitter censorship, there is either some amount of human discretion involved, or the mechanism has some degree of inaccuracy in its targeting.

6. CENSORSHIP TOPICS

Now that we have established a remarkable volume of Turkish censorship of Twitter, the next question is to understand what topics the Turkish censors are interested in. This sort of analysis is valuable for understanding the political aims of the Turkish censors. It is also pragmatically valuable to determine hashtags and topics that would help in discovering additional censored tweets that our earlier methods may have missed.

Topic extraction and clustering is a standard feature of natural language processing systems. The key concept behind automatic topic extraction is to assign weights to terms and sentences based on their frequency of appearance. We applied non-negative matrix factorization (NMF) combined with term frequency–inverse document frequency (*tf-idf*) to extract hot topics. *tf-idf* is a weighting schema widely used in document classfication. The tf-idf value of a word increases proportionally to the number of times it appears in the document, but is also offset by the frequency of the word in the corpus. The advantage of *tf-idf* over a simple word frequency is that it can effectively adjust weights of words that are very frequent but not informative [20].

NMF is a technique to factor document-term matrix into a term-topic and a topic-document matrix. The topics are derived from the contents of all tweets. NMF has been widely used in text mining related applications including clustering on email message, scientific journals and Wikipedia articles [9]. The method has been shown to be effective in monitoring underlying semantic features (topics) in a general way [3].

6.1 Hot topic clustering

We first tokenize each censored tweet into a list of words, eliminating Turkish stop words[3]. After this, we applied *tf-idf*, built the document-term matrix, and used non-negative matrix factorization on the document-term matrix to extract the top 5 topics with 10 words for each topic. The result of this process is a series of Turkish words that may be best interpreted by a Turkish speaker, although

---

[2]**Fethullah Gülen**: a Turkish preacher and founder of the Gülen movement [33].

[3]Google has a nice library for stop-word removal. https://code.google.com/p/stop-words/

Google's Translation service is very helpful. See Table 3 for the hottest censored topics.

(This method could potentially be improved in a variety of ways, e..g, exploring the quality and stability of topic clusters as a function of the number of topics, using n-grams rather than unigrams, or varying the number of words in the *tf-idf* document representation. Nonetheless, our method still yields interesting results.)

Topic 0 is about cursing media group owned by Aydin Dogan, largest media corporation in Turkey, and calling them dogs and having no morals and dignity. This media group is known to favor the leading Justice and Development Party (AKP). Since 2011, AKP has increased its restrictions on freedom of speech and television content. It has used legal and administrative measures against opponent media groups and journalists [4]. The issue became prominent during 2013 Taksim Gezi Park Protest, when most news outlets were loath to irritate the government because their owners' business interests relied on government support [16].

Topic 1 is the only topic for which we did not find a clear interpretation. The topic seems to be related to Şekerbank, a financial institution in Turkey, and a person name Ibrahim Karaca.

Topic 2 discusses the Koç family in Turkey. Vehbi Koç, who founded Koç Holding A.S. in 1963, is a Turkish entrepreneur and philanthropist. His son Rahmi Koc took his father's position as the chairman of the company in 1984 and retired in 2003. Following the crackdown of the Taksim Gezi Park protests, Erdogan went after the Koç family as they had criticized his actions. He used the Turkish equivalent of the Internal Revenue Service to go through the Koç Holdings financial books, imposing less than $100,000 in fines [26].

We see Judaism referenced as well in Topics 2 and 4. Both the president Erdogan and the prime minister Davutoglu make a habit of criticising Jews. For example, Erdogan has openly claimed that Muslims lived in American before the Jewish Christopher Columbus sailed to liberate Jerusalem from Muslims and accidentally spotted the new continent; he also suggested that the protest in Taksim Gezi Park was part of an effort by the American "Jewish lobby" to undermine Turkey's government [22].

Topic 3 discusses the controversy ignited by Lütfi Elvan, head of Minister of Transport, Maritime and Communication in Turkey. Elvan proposed to establish Turkish own web protocol using "ttt" as a prefix instead of "www". This suggestion has led to broad criticism inside Turkey.

Topic 4 is about people's disappointment and condemnations of current Turkish Prime Minister Ahmet Davutoglu and the topical words found by our analysis appear relate to his anti-Jewish rhetoric and appear to summarize the use of vulgar language to describe him.

Evidently, strongly worded and vulgar political discourse are top on the minds of Turkey's censorship authorities.

6.2 Hot topics from withheld accounts

We applied a similar topic analysis methodology toward tweets from withheld accounts (i.e., accounts where every single tweet has been withheld versus accounts where only some tweets are withheld). There are 46 such accounts in our dataset. However, the algorithm failed to extract meaningful topics due to the apparently diverse contents of the withheld tweets. The resulting keywords were mostly usernames and hashtags. Consequently, we performed a manual inspection of the withheld accounts by looking at their Twitter profiles and their timelines tweets. Table 4 summarizes our impressions of these accounts.

Users from the "Politics" category constantly post anti-government comments. These user also openly criticize current Turkish presi-

	Turkish "topic"	English translation
#0	doğan aydın medya şerefsiz grubu medyası vatan hurriyet değil köpeği	Aydin Dogan (a person who owns the biggest media in Turkey) media dishonest group home freedom not dog
#1	co rt şekerbank şapşik in hikayesi dikkat chp ibrahim karaca	Şekerbank (a Turkish financial institution) attention in story Ibrahim Karaca (a Turkish name)
#2	koç un vehbi oğlu ın aydın rahmi doğan yahudi nahum	Vehbi Koç (a Turkish entrepreneur / philanthropist) son enlightened womb rising Jewish
#3	elvan lütfi bakan bakanı ulaştırma ttt aptal bi www adam	Lütfi Elvan (a Turkish government minister) stupid man minister transportation
#4	davutoğlu ahmet başbakan lan pic sikeyim yahudi gavat göt vatan	Ahmet Davutoglu (current Turkish prime minister) prime minister man bastard fuck Jewish pimp ass

Table 3: Hot topics from withheld tweets from non-withheld accounts.

Category	Number of accounts
Politics	36
Pornography	2
Advertising Bots	1
Unidentified	3
Not Found	4

Table 4: Topics discussed in withheld accounts.

dent Erdogan by attacking his personality or posting political caricatures. They also tend to have a large number of followers.

Users from the "Pornography" group, constantly posts explicit images and video clips with links to other pornography-related Twitter accounts or websites. Users from this group also have a reasonably large number of followers.

Users from the "Advertising Bot" category repeatedly send tweets with similar or identical contents. Each tweet has a link at the end that will redirect to a company's website.

We classified three users in an "Unidentified" category. They did not appear to be engaged in political speech or anything else which would appear to be worthy of censorship.

Finally, users from the "Not Found" category were originally in our dataset but when we later looked at their profiles, we could not find their accounts anymore. These accounts could have been deleted by the users' own actions or through the actions of Twitter (perhaps on the behest of the Turkish government, perhaps not).

7. BYPASSING CENSORSHIP

In April 2015, we followed a group of withheld accounts in Turkey and noticed that at least 7 users were still tweeting from inside the country despite having their entire accounts withheld. We manually inspected each profile and found that these users tweeted topics political in nature, specifically criticizing Erdogan's leadership. We also noticed that all of these users had large number of followers, and were thus presumably quite influential.

We note that these users were originally found using the Streaming API with geo-bounded boxes set to Turkish cities. Presumably, they all live physically in Turkey and wish to be read by a Turkish audience. So how are they being seen?

We found that Twitter manages location information by a cookie set in the browser. We tested this by viewing a known-censored tweet using a Turkish proxy server and a regular Internet browser from our home institution. In both cases the known-censored tweets in Turkey were not presented as having been withheld. Conversely, using a proxy did not help us bypass censorship. However, when we changed the location setting in the Twitter application and set it to "Turkey", surprisingly the tweet appeared withheld despite our

physical location outside of the country. It is possible that this trivial hack is used inside Turkey. All a Turkish user needs to is simply set their "country" to be elsewhere and they can see everything. We searched the web both in English and Turkish languages for instructions on bypassing censorship in Turkey, to see if this seemingly obvious advice was described anywhere on the Web. We found many Turkish descriptions on how to use Tor or paid VPN services to evade censorship (see, e.g., [7, 12]). However, we found no results on evading Twitter censorship using our method (i.e., changing the location setting). It is likely that this method will gain popularity after the publication of this paper.

Perhaps Twitter is being deliberately simplistic with its censorship mechanism, doing what it interprets to be the bare minimum necessary in order to operate its service in countries that might otherwise ban it altogether. Certainly, Twitter is attempting to walk a fine line to allow its users to have their tweets as widely visible as possible, despite different nation-states having different legal policies for what must be censored in their respective jurisdictions.

7.1 Censorship escalation

Consider what might occur if Twitter gets more serious about censorship. With basic IP geolocation techniques, which Twitter appears to already employ, Twitter could base censorship decisions on these geolocations rather than on its easily-changed cookie. In response, users could use proxy or VPN services (Tor, etc.) to obfuscate their locations. If this continues, the logical conclusion is that Twitter will have no reliable signals that indicate a user's location.

What then? Countries may attempt to strong-arm Twitter into replacing its "withholding" mechanism with a more draconian deletion policy, or else ban them from the country. In such a circumstance, what would it mean for a user in one country, criticizing another country, to find their tweet censored worldwide through no action of their home government? (Example: if Twitter cannot reliably distinguish an American writing a tweet about Turkey from a native writing the same tweet, and the Turks press their case, they may demand the ability to censor the American's tweet.) This draconian progression is the seemingly inevitable path that Twitter will be forced to follow. Google is fighting against requirements that it follow a similar path with respect to European "right to be forgotten" restrictions on its search results [25]. Both Twitter and Google's mechanisms of withholding content in one country while preserving it in others are fundamentally fragile and would seem unlikely to survive in the face of insistent government-sponsored censorship.

7.2 How many users use Tor in Turkey

As discussed above, Turkish users are already aware of Tor as a mechanism for overcoming censorship. How popular, then, is Tor

in Turkey? The Tor project has a public tool "Tor Metrics-Direct users by country"[4] that estimates the daily number of directly-connecting clients to the Tor network from any country, and also displays "indications of censorship events" [8]. Figure 9 shows a Tor metrics graph for Turkey. Red dots indicate possible censorship events and blue dots indicate a "release" of censorship events. We also plotted the distribution of all withheld tweets we detected and those we collected from Chilling Effects, ordered by the "created_at" month as shown in Figure 10.
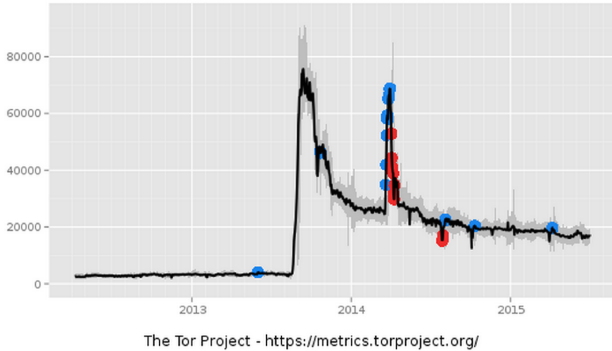
Figure 9: Tor Metrics: users connecting from Turkey.
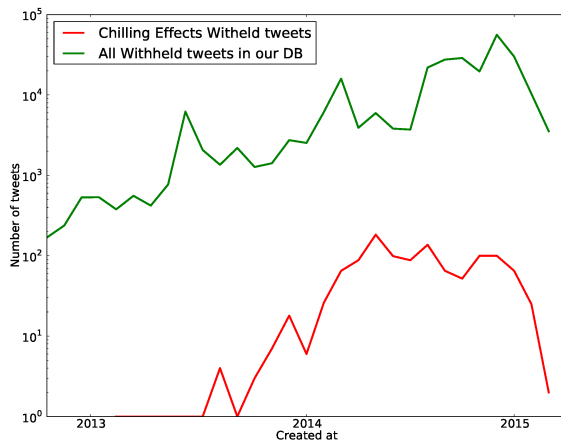
Figure 10: Distribution of withheld tweets by month (log scale).

Both graphs follow roughly the same time period, from 2013 through 2015. There are two noticeable peaks on the Tor graph, in May 2013 and March 2014, which correspond to the end of the Taksim Gazi Park protests (2013) and a brief ban on Twitter across all of Turkey (2014). The withholding graphs, based on our own data collection, have peaks as well, which seem to correlate with these same two events, but the overall trajectory of the withholding is very much upward, versus the apparently otherwise steady-state usage of Tor. This suggests that, to the mind of the Turkish Twitter censors, the withholding mechanism is "good enough" for their needs. The Tor volume suggests that the Turkish population agrees. The y-axis of the Tor graph is the number of daily unique users of Tor, so the Turkish government may be willing to allow 20,000 people to have unfettered Twitter access so long as

---

the larger population's ability to see and spread information is sufficiently squelched.

8. CONCLUSIONS AND FUTURE WORK

In this work, we collected data on the frequency of Turkey's use of Twitter's censorship mechanisms. Through Twitter's own transparency reporting, Turkey is one of the countries with the most censorship so we focused our attentions there and found two orders of magnitude more censorship than Twitter officially reports. The actual numbers might be even higher, but it is difficult to be certain due to Twitter's rate limits and other restrictions on external groups who want to crawl their content. Certainly, our results make it clear that Twitter's transparency reports are entirely unrepresentative of the actual scale and scope of censorship that Twitter enables.

There are a variety of directions for future work related to this study. Expanding this work to study Twitter censorship in other countries is an obvious direction, particularly in countries with no *reported* censorship requests. A full country-by-country study would require significantly greater resources than we could muster from our home institution. While we certainly could scale up with modern cloud services, engineering a full-scale web crawler dedicated to downloading each and every tweet as it is posted, with periodic followups to check every tweet's withholding status, would represent a non-trivial load on Twitter's service; Twitter would presumably then take technical and/or legal steps to block such a research crawler. Alternatively, Twitter could allow researchers access to its "firehose" without restrictive agreements as to how they may or may not use the data, although this seems politically unlikely.

Twitter ultimately faces a stark challenge. By doing business in countries that require censorship of tweets, including the requirement that Twitter wildly distort its reporting of *statistics* over those tweets, while simultaneously allowing users outside of those censorship regimes to read and respond to those same tweets, Twitter is trying to placate its censors while still maximizing what free speech might be leftover. This is not a long-term stable business strategy, and we fear freedom of speech may be the loser in this battle.

Acknowledgements

9. REFERENCES

[1] N. Aase, J. R. Crandall, Á. Díaz, J. Knockel, J. O. Molinero, J. Saia, D. Wallach, and T. Zhu. Whiskey, Weed, and Wukan on the world wide web: On measuring censors' resources and motivations. In *USENIX FOCI*, Bellevue, WA, USA, 2012. https://www.usenix.org/conference/foci12/workshop-program/presentation/Aase.

[2] Avrupa Times. *Turkey, Twitter meet for the third time in Istanbul*, 2014. https://www.avrupatimes.com/technology/7293-turkey-twitter-meet-for-the-third-time-in-istanbul.html.

[3] M. W. Berry and M. Browne. Email surveillance using nonnegative matrix factorization. *Computational and Mathematical Organization Theory*, 2005. http://link.springer.com/article/10.1007/s10588-005-5380-5.

[4] D. Bilefsky and S. Arsu. Charges Against Journalists Dim the Democratic Glow in Turkey. *NY Times*, 2012. http://www.nytimes.com/2012/01/05/world/europe/turkeys-glow-dims-as-government-limits-free-speech.html?_r=0.

[5] L. Chen, C. Zhang, and C. Wilson. Tweeting under pressure: Analyzing trending topics and evolving word choice on Sina Weibo. In *ACM Conference on Social Networking*, Boston, MA, USA, 2013. http://www.ccs.neu.edu/home/cbw/pdf/weibo-cosn13.pdf.

[6] S. Cornell. Erdoğan's looming downfall. *Middle East Quarterly*, 2014. http://www.meforum.org/3767/erdogan-downfall.

[7] Daily Turkiye. *Engelli sitelere girme kılavuzu*, 2015. http://www.dailyturkiye.com/Haber/engelli-sitelere-girme-kilavuzu.html.

[8] G. Danezis. An anomaly-based censorship detection system for Tor. *The Tor Project*, 2011. https://research.torproject.org/techreports/detector-2011-09-09.pdf.

[9] A. N. Finn. Clustering of scientific citations in Wikipedia. *Wikimania 2008*, 2008. http://www.researchgate.net/publication/220489653_Clustering_of_scientific_citations_in_Wikipedia.

[10] A. D. Florio, N. V. Verde, A. Villani, D. Vitali, and L. V. Mancini. Bypassing Censorship: a proven tool against the recent Internet censorship in Turkey. In *Software Reliability Engineering Workshops (ISSREW)*, Naples, Italy, 2014. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=6983872&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D6983872.

[11] Z. Fox. Half of all active Twitter users live in five countries. *Mashable*, 2013. http://mashable.com/2013/11/20/twitter-users-countries/.

[12] Habber. *Twitter gizli hesapları görme: "This account has been withheld in Turkey"*, 2014. http://www.habber.net/gundem/twitter-gizli-hesaplari-gorme-this-account-has-been-withheld-in-turkey-h719.html.

[13] S. Khattak, L. Simon, and S. J. Murdoch. Systemization of pluggable transports for censorship resistance. *arXiv CoRR*, abs/1412.7448, 2014.

[14] S. Machkovech. Political deleted-tweet archive shuttered by Twitter over privacy expectation. *Middle East Quarterly*, 2015. http://arstechnica.com/tech-policy/2015/06/political-deleted-tweet-archive-shuttered-by-twitter-over-privacy-expectation/.

[15] N. Mandell. Twitter users protest new twitter policy-#twittercensored #twitterblackout. *NY Daily News*, 2012. www.nydailynews.com/news/money/twitter-users-protest-new-twitter-policy-twittercensored-twitterblackout-article-1.1013499.

[16] W. R. Marczak, J. Scott-Railton, and M. Marquis-Boire. When governments hack opponents: A look at actors and technology. In *USENIX Security*, San Diego, CA, USA, 2014. https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-marczak.pdf.

[17] D. Mocanu, A. Baronchelli, N. Perra, B. Gonçalves, Q. Zhang, and A. Vespignani. The Twitter of Babel: Mapping world languages through microblogging platforms. *PLoS ONE*, 8(4), 2013. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061981.

[18] D. Morrison. *Trends and Applications in Knowledge Discovery and Data Mining*, chapter Toward Automatic Censorship Detection in Microblogs. Springer International Publishing, 2014. http://link.springer.com/chapter/10.1007/978-3-319-13186-3_51#.

[19] F. Morstatter, J. Pfeffer, H. Liu, and K. Carley. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's firehose. In *International AAAI Conference on Weblogs and Social Media*, Boston, MA, USA, 2013. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6071/6379.

[20] A. Rajaraman and J. D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2011. http://www.langtoninfo.com/web_content/9781107015357_frontmatter.pdf.

[21] E. Reddy. Occupy Gezi: How Twitter facilitated a social movement in Turkey. *gnip*, 2014. https://blog.gnip.com/occupy-gezi-twitter/.

[22] Rehmat's World (blog). *Turkey's $ 350m Presidential Palace irks Jewish Lobby*, 2014. http://rehmat1.com/2014/12/06/turkeys-350m-presidential-palace-irks-jewish-lobby/.

[23] Reuters News. *UPDATE 2-Widespread Twitter outages in Turkey after PM threatens ban*, 2013. http://www.reuters.com/article/2014/03/20/turkey-twitter-idUSL6N0MH5KQ20140320.

[24] J. Roos. The Turkish protests and the genie of revolution. *Roarmag*, 2013. http://roarmag.org/2013/06/tahrir-taksim-egypt-turkey-protests-revolution/.

[25] M. Scott. Google fights effort to apply 'right to be forgotten' ruling worldwide. *New York Times (Bits Blog)*, July 2015. http://bits.blogs.nytimes.com/2015/07/30/google-fights-effort-to-apply-right-to-be-forgotten-ruling-worldwide/.

[26] R. Swier. The rise of Erdogan's Islamist tyranny, 2014. http://drrichswier.com/2014/02/01/the-rise-of-erdogans-islamist-tyranny/.

[27] Sysomos. *Inside Twitter: An In-Depth Look Inside the Twitter World*, Apr. 2014. http://sysomos.com/uploads/Inside-Twitter-BySysomos.pdf.

[28] Twitter. Country withheld content policy, 2012. https://support.twitter.com/articles/20169222#.

[29] Twitter. Tweets must flow, 2012.

[30] Wikipedia. Gezi Park protest in Turkey, 2013. http://en.wikipedia.org/wiki/Gezi_Park_protests.

[31] Wikipedia. Arab spring, 2015. http://en.wikipedia.org/wiki/Arab_Spring.

[32] Wikipedia. Chilling Effects, 2015. http://en.wikipedia.org/wiki/Chilling_Effects.

[33] Wikipedia. Fethullah Gülen, 2015. https://en.wikipedia.org/w/index.php?title=Fethullah_G%C3%BClen&oldid=671559583.

[34] Wikipedia. Libyan civil war, 2015. http://en.wikipedia.org/wiki/Libyan_Civil_War.

[35] P. Winter. *Measuring and circumventing Internet censorship*. PhD thesis, Karlstad University, 2014. http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A758124&dswid=4913.

[36] E. Yeginsu. Turkey lifts Twitter ban after court calls it illegal. *NY Times News*, 2014. http://www.nytimes.com/2014/04/04/world/middleeast/turkey-lifts-ban-on-twitter.html.

[37] T. Zhu, D. Phipps, A. Pridgen, J. R. Crandall, and D. S. Wallach. The velocity of censorship: High-fidelity detection of microblog post deletions. In *USENIX Security*, Washington, D.C, USA, 2013. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/paper/zhu.