**Censorship Versus Legal Protections in Takedown Requests Sent to Google, Facebook, and Twitter by Governments**

A Thesis in Quantitative Methods in the Social Sciences

Graduate School of Arts and Sciences

Columbia University

**By Cara Feldkamp**

May 20, 2020

Thank you to Greg Eirich, PhD, and Elena Krumova, PhD, for their continued guidance and support on this project and throughout this process.

Cara Feldkamp

QMSS Thesis

May 20, 2020

**Censorship Versus Legal Protections in Takedown Requests Sent to Google, Facebook, and Twitter by Governments**

1. **Abstract**

As evidence is mounting that certain governments are taking advantage of social media algorithms to amplify desirable messages and drown out undesirable messages, I aim to use this knowledge as a basis for investigating whether these same governments may also be using the content flagging and removal tools provided by major web platforms to effectively censor information that they do not want to be seen. This research paper will focus on takedown requests sent to Google, Twitter, and Facebook by government agencies, courts, and law enforcement organizations and seeks to better understand how and why countries are using these removal request features. If content removal requests are at least partially motivated by a desire to suppress information in an easy and legally-mandated way, then we would expect to see countries that the international community considers to have a poor record on free expression sending these requests more often than countries that allow free or freer speech.

1. **Introduction**

*1.1 Background*

The US election in 2016 revealed some of the strongest and most widely publicized evidence of government actors taking steps to interfere with the way the public consumes information. An investigation by the US government found that on social media, state-sponsored bot and troll accounts coming from Russia were part of a coordinated attack meant to misinform American voters, incite violence, and disrupt the electoral process. Officials in France, the UK, Canada, several other countries claimed to find information indicating that there were similar attacks aimed at disrupting their elections. The conversation that followed in the public discourse about “fake news,” political advertising, and how and why we see the content that we see online quickly distilled into a collective paranoia that, in many ways, has become a hallmark of the modern age.

As evidence is mounting that certain governments are taking advantage of social media algorithms to amplify desirable messages and drown out undesirable messages, I aim to use this knowledge as a basis for investigating whether these same governments may also be using the content flagging and removal tools provided by major web platforms to effectively censor information that they do not want to be seen. This research paper will focus on takedown requests sent to Google, Twitter, and Facebook by government agencies, courts, and law enforcement organizations and seeks to better understand how and why countries are using these removal request features. If content removal requests are at least partially motivated by a desire to suppress information in an easy and legally mandated way, then we would expect to see countries that the international community considers to have a poor record on free expression sending these requests more often than countries that allow free or freer speech.

Obviously, there are many legitimate reasons why government agencies might file a request to remove content from the internet. Some of the most common reasons are that the content violates the law of that country, endangers children, reveals sensitive information that could threaten national security, is baseless and harmful slander, is part of an ongoing legal investigation, or reveals the identity and/or location of someone who is in witness protection or otherwise at risk of violence. Google, Facebook, and Twitter claim that they only remove content if they are required to do so by law or court order, and leave the rest intact unless it violates their Terms of Service (TOS) agreements. All three companies have reported that they regularly receive and deny requests from government entities to remove content that does not violate the law. These companies regularly release Transparency Reports that include descriptions of certain requests that they deem to be “of public interest,” many of which describe attempts to remove unflattering photos of politicians, writing criticizing heads of state, and unpatriotic sentiments.

*1.2 Literature Review*

**Internet Filtering and Censorship in Practice**

According to the OpenNet Initiative a partnership between Harvard University and the University of Toronto aimed at monitoring and studying barriers to free and open internet usage - there are four main approaches to internet filtering that countries typically employ: [[1]](#footnote-1)

1) *Technical blocking:* Technical blocking involves IP blocking, URL blocking via proxy, DNS tampering, and occasionally keyword blocking; these techniques are used to prevent access to certain webpages, domains, or IP addresses.[[2]](#footnote-2)

2) *Search result removals:* Search result removals require cooperation from internet search engines (e.g. Google) to delist certain websites from search results, or remove them from the search results in a country’s web domain, which makes those pages much more difficult to find.

3) *Takedowns:* Takedowns are requests to a website or platform to remove content or make it unavailable within a certain country’s web domain via withholding; often, these requests come in the form of a court order claiming that the content is illegal in the requesting country.

4) *Inducing self-censorship among citizens:* Self-censorship is often imposed on citizens of countries without free speech by creating fear of retribution for posting or viewing certain content online, either through formal legal measures, or through informal threats and intimidation.

Many countries use all four of these techniques simultaneously, however, this paper will focus on search result removals and content takedowns, which are often more difficult to identify as censorship. Complicating things further, there is also a lack of consensus within the international community on what constitutes censorship. Technical blocking tactics and imposing self-censorship via threats is considered unethical among most of the countries that participate in international governance organizations. The United Nations (UN) Human Rights Council, for example, now considers blocking citizens’’ access to the internet or content on the internet to be a human rights violation.[[3]](#footnote-3) However, the UN takes a position in favor of restricting freedom of expression in the context of policing hate speech; the United Nations Office of the High Commissioner for Human Rights created a legally-binding covenant mandating that “any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law.”[[4]](#footnote-4)

Many of the UN’s member states, particularly European countries within the European Union (EU), have similar legal measures against hate speech, which they actively enforce by either directly or indirectly restricting access to online content that they consider to be hate speech or promoting hate. American law tends to have fewer restrictions on speech and expression, except against expression that violates copyright or intellectual property laws. [[5]](#footnote-5) The Digital Millennium Copyright Act (DCMA) is the law that requires internet companies to filter allegedly copyright-infringing content and have a process in place for rapidly handling takedown and/or search delisting requests.[[6]](#footnote-6)

In his Duke Law Review paper on “cybersieves,”[[7]](#footnote-7) Derek E. Baumbauer writes that “most countries use cybersieves to try to filter undesirable content and make it disappear from the Web. Whether it is copyrighted songs in America or political dissent in Iran, the goal is the same; only the targeted material varies.”[[8]](#footnote-8) He goes on to say that taking an ethical or values-based approach to censorship is unhelpful because it happens nearly everywhere; rather, he provides a framework for evaluating the degree to which governments are open and honest about what content they are filtering and why, and how successful their methods are at filtering out only the intended material.[[9]](#footnote-9)

**Platform Policies and Legal Compliance**

Google, Twitter, and Facebook claim in their Transparency Reports that they only remove content if they are required to do so by law or court order, or if the content violates their usage policies.[[10]](#footnote-10) [[11]](#footnote-11) [[12]](#footnote-12) Otherwise, they typically comply with the requesting country’s laws by “withholding” the content from internet users within that country’s domain and allowing it to remain viewable elsewhere.[[13]](#footnote-13) European “right to be forgotten” laws pose additional compliance challenges because EU citizens are able to make requests for web content containing personal information about themselves to be removed for any reason, and there is little guidance or precedent for how to handle requests by political actors and other influential people. However, Google won a case in the European Court of Justice in September 2019 where the court ruled that “right to be forgotten” laws only apply within EU web domains.[[14]](#footnote-14)

Google, Twitter, Facebook, and other internet companies have largely attempted to position themselves as being politically neutral service providers, but some of their usage policies and processes can facilitate content filtering that benefits governments with an interest in suppressing dissenting voices. Marjorie Heins’ 2014 paper[[15]](#footnote-15) in the Harvard Law review points out that Google, Twitter, and Facebook are headquartered in the United States, and under American law, search engines and content-hosting web platforms are not legally liable for content generated by their users; according to Heins, none of these companies or their subsidiaries are legally obliged to censor anything, as long as they respond to takedown requests for copyright and intellectual property violations.[[16]](#footnote-16) Nonetheless, all of the major social media platforms’ usage guidelines prohibit posts and videos containing nudity, hate speech, and excessive violence, and each platform has its own confidential, internal algorithm for detecting and banning these kinds of content. Such policies could easily censor important information posted by eyewitnesses, whistleblowers, and “citizen journalists.”[[17]](#footnote-17) In response to the 2019 livestreamed mosque shooting in Christchurch, New Zealand, the Australian parliament passed the Unlawful Showing of Abhorrent Violent Material Bill, which would issue hefty fines to social media companies who do not remove violent content “expeditiously;” members of the Australian Law Council have raised concerns that the law could "silence and criminalise whistleblowers trying to bring attention to violent atrocities occurring overseas."[[18]](#footnote-18)

While no such laws exist in the United States, YouTube, which is owned by Google, uses an upload filter to check for content that violates the platform’s usage policies and instantly deletes videos in violation of those policies.[[19]](#footnote-19) YouTube also has a three-strike policy for automatically shutting down a user’s account if it has been flagged for Terms of Service or Community Guidelines violations, and it can take up to 90 days to get an account reinstated if the affected user files a successful appeal.[[20]](#footnote-20) The quick shutdown process facilitated by an algorithm and subsequent lag time in the platform’s account reinstatement procedures could be exploited by state actors who are trying to prevent certain information from being seen.

The internet companies themselves are already aware of government entities trying to use their content removal features to quietly suppress various material, and these types of requests were the impetus for creating the Transparency Reports. In an effort to provide guidelines and legal support for companies dealing with these issues, a group of stakeholders from the tech industry, civil liberties law, and academia came together to create the Global Network Initiative (GNI) in 2008.[[21]](#footnote-21) Google and Facebook are members of GNI, and Google, Facebook, and Twitter have all expressed desire to uphold the Santa Clara Principles, which are a shortlist of guidelines for ensuring transparency in online content moderation.[[22]](#footnote-22)

Nevertheless, when faced with stern requests and court orders, Google, Facebook, and Twitter are often pressured to cave to the demands of regimes with questionable records on human rights and free speech, in an effort to avoid expensive and potentially damaging lawsuits. Twitter has been criticized by a number of journalists and activists for withholding or suspending entire accounts in cases where they could have withheld individual tweets.[[23]](#footnote-23) Buzzfeed maintains a public, running list of highly followed Twitter accounts that have been withheld in at least one country, many of which are accounts run by journalists and news outlets.[[24]](#footnote-24)

**Social Media and Search Engines as Tools for Burying Dissident Views**

The topic of filtering content on the internet via indirect means is somewhat new and, since 2016, has primarily been discussed in the context of state-sponsored online content manipulation as a political tool. Some recent studies have been published where the researchers found evidence of coordinated attempts by government actors to control a political narrative by using bot and troll accounts to dominate and bury the organic conversation happening on trending hashtags. These studies do not involve content removal requests, but are still relevant because they provide insight into which countries seem to be engaging in malicious indirect filtering tactics online. Sending a takedown request to Google, Twitter, or Facebook requires less time and effort than any of the activities described in the research papers below, so it seems logical to assume that government actors who are likely utilizing networks of fraudulent social media accounts may also be submitting a large volume of takedown requests as a first step.

A September 2018 analysis by Gallagher, Suarez-Serrato, and Velazquez Richards[[25]](#footnote-25) examined a series of tweets about the Honduran president after he was elected in 2017; they found over 100 clusters of coordinated social bots that were responding to tweets about the president with positive messaging in an attempt to drown out commentary about the violence and discord in Honduras following the election. The researchers concluded that the bot accounts effectively created a “positive social media fog” in the wake of the elections where the goal was to obscure information about protests and widespread discord in Honduras.

Similarly, Thomas, Grier, and Paxson found in their 2012 study[[26]](#footnote-26) that the same tactics had been used on Twitter to bury undesirable political messages in Russia as early as 2011. Using private data shared with permission from Twitter and using accounts suspended by Twitter’s internal bot detection algorithm, they found that in the days following the Russian parliamentary election in December of 2011, over 25,000 interconnected bot accounts generated more than 440,000 pro-Kremlin tweets that were used to dominate the hashtags that protesters and activists were using to organize. The first spike in activity from these accounts corresponded with the launch of a smear campaign against a prominent Russian anti-corruption blogger. Incredibly, the researchers were able to link the bot accounts that had tweeted to a pool of 975,283 total accounts created en masse in a similar way, 80% of which were dormant accounts that were ready to be used for future spam campaigns.

Although they took a slightly more qualitative approach, the International Press Institute was able to identify a network of troll accounts that were likely used by Turkish security forces to find, capture, and detain Kurdish journalist Nedim Turfent in 2016.[[27]](#footnote-27) Turfent had travelled to rural, southeastern Turkey on an assignment when he started getting a barrage of comments, replies, and direct messages on social media from accounts with anti-Kurdish names, issuing death threats and demanding to know his whereabouts. These accounts also began to threaten and intimidate local people living in the area, berating them to reveal Turfent’s location. The next day, villagers tweeted that Turkish special forces had raided and imposed a curfew in their region, and one day later, Turfent was captured, arrested, and imprisoned on charges of terrorism - a common accusation leveled against journalists in Turkey.[[28]](#footnote-28) This paper comes the closest to concretely proving that the troll accounts were being run by the military.

**Key Takeaways**

An obvious problem is that it is nearly impossible to prove whether government entities’ requests to remove or withhold content are aimed at censoring information in an effort to manipulate a larger message. It is easy to identify countries where the government blocks access to the internet or to certain websites as perpetrators of censorship; however, when a search engine or web platform does the blocking on behalf of a domestic or foreign government, the line is less clear. In this paper, I will focus on trying to discover whether there are any trends in the way countries use platform removal request tools and whether those trends connect back to activities and behaviors that many experts whose fields intersect with this topic consider to be censorship with malicious intent.

*1.3 Research Questions*

**Q1:** *How similar are countries with a large number of content removal requests?*

**Q2:** *Do countries known to engage in censorship make more content removal requests than countries where free or freer speech is the norm?*

*1.4 Hypotheses*

**H1:** *Countries that sent a similar volume of removal requests will cluster into groupings with countries whose legal or political systems have a similar approach to free speech.*

**H2:** *Countries that do not have free speech but do have high internet usage make more content removal requests than countries with freer speech and/or lower internet usage.*

*1.5 Methodology*

I used government removal request data from the Twitter, Google, and Facebook Transparency Reports to investigate governments’ activities efforts to remove content from the web. These three internet companies are not the only platforms that produce Transparency Reports, but these three started collecting and reporting on this data several years before other web platforms (Apple, Reddit, Vimeo, GitHub, etc.) ; the data from Google, Twitter, and Facebook are downloadable, fairly comprehensive, and represent the bulk of government removal requests on the internet, whereas other companies’ Transparency Reports had only a small amount of data, presented in the aggregate, in a non-downloadable format.

I also included Facebook’s data on reported internet outages and the World Bank’s World Development Indicators dataset to provide descriptive features to help identify any external, explanatory variables and to help assess whether countries that submit a large volume of removal requests have any key characteristics in common. The internet outage dataset is small and was started in 2016, but it contains a detailed record of how long each outage lasted and in what specific regions the outages were observed. The World Development Indicators (WDI) are a large dataset, containing over 1,500 features.

As a measure of which countries are generally considered to have a poor record on freedom of expression and other civil liberties, I used three commonly-cited index scores: The World Bank’s World Governance Indicators, Reporters Sans Frontières’ Press Freedom Index, and Freedom House’s World Freedom Index. Because there is no generally agreed-upon means of evaluating a metric as subjective as a country’s level of freedom, these index scores represent different efforts to apply quantitative standardization to qualitative information. As such, these scores were created using methodology that is inexact and prone to bias.[[29]](#footnote-29) However, all three scores are used by countless government, academic, and civil society organizations to evaluate and compare countries; because they are so widely used, it may be possible to conduct or reference a meta-analysis of other studies using these same scores and attempt to control for bias.

After running some initial regressions and trend analyses to gain a better understanding of the relationships between the removal request variables and the free speech/free press index scores, I combined the removal requests, index scores, internet disruptions, and development indicators by country and year. Next, I ran a Principal Component Analysis (PCA), both for dimensionality reduction and to get a better sense of feature importances. I decided to use PCA because the algorithm can handle high-dimensional datasets with multicollinearity. I then used the principal components to do hierarchical clustering and look for distinguishing characteristics of the clusters of countries.

1. **Description of Datasets**

*2.1 Content Removal Requests*

**Google Transparency Report: Government Removal Requests[[30]](#footnote-30)**

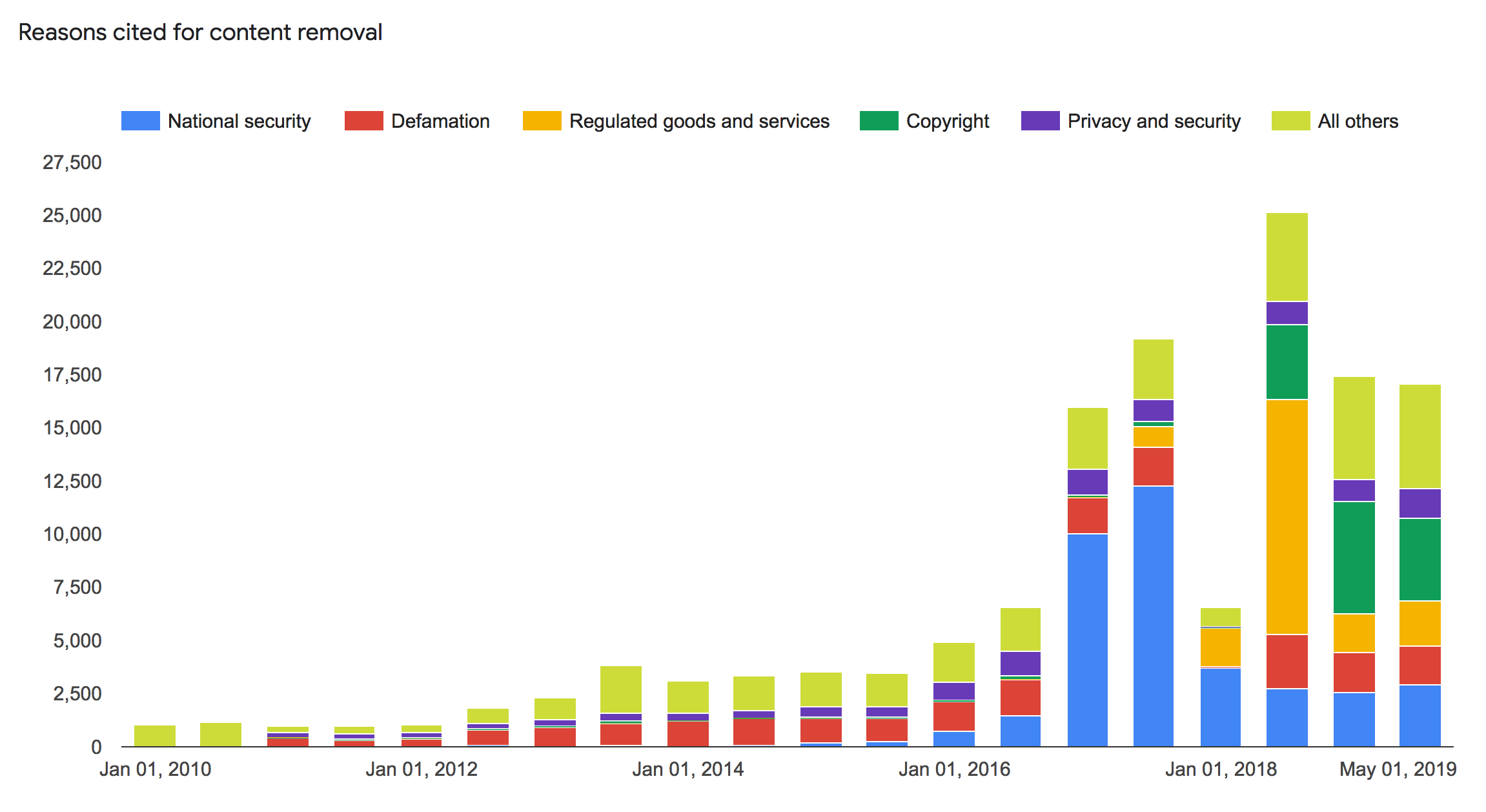
Starting in 2009, Google has released a biannual Transparency Report that includes data about requests by government entities, domestic and foreign, to remove content from Google search results and from the websites of other Google-owned platforms, including YouTube, Blogger, etc. When someone fills out a request to remove content from a Google product or from web searches, Google uses an automated process to check whether the content has been changed in the location where it is hosted and no longer matches what appears in Google search results. If it has changed, Google removes the out-of-date content from searches and other places where it appears via an automated process. Otherwise, the removal request is reviewed by employees at Google and, if necessary, passed along to legal counsel and/or the public policy team.

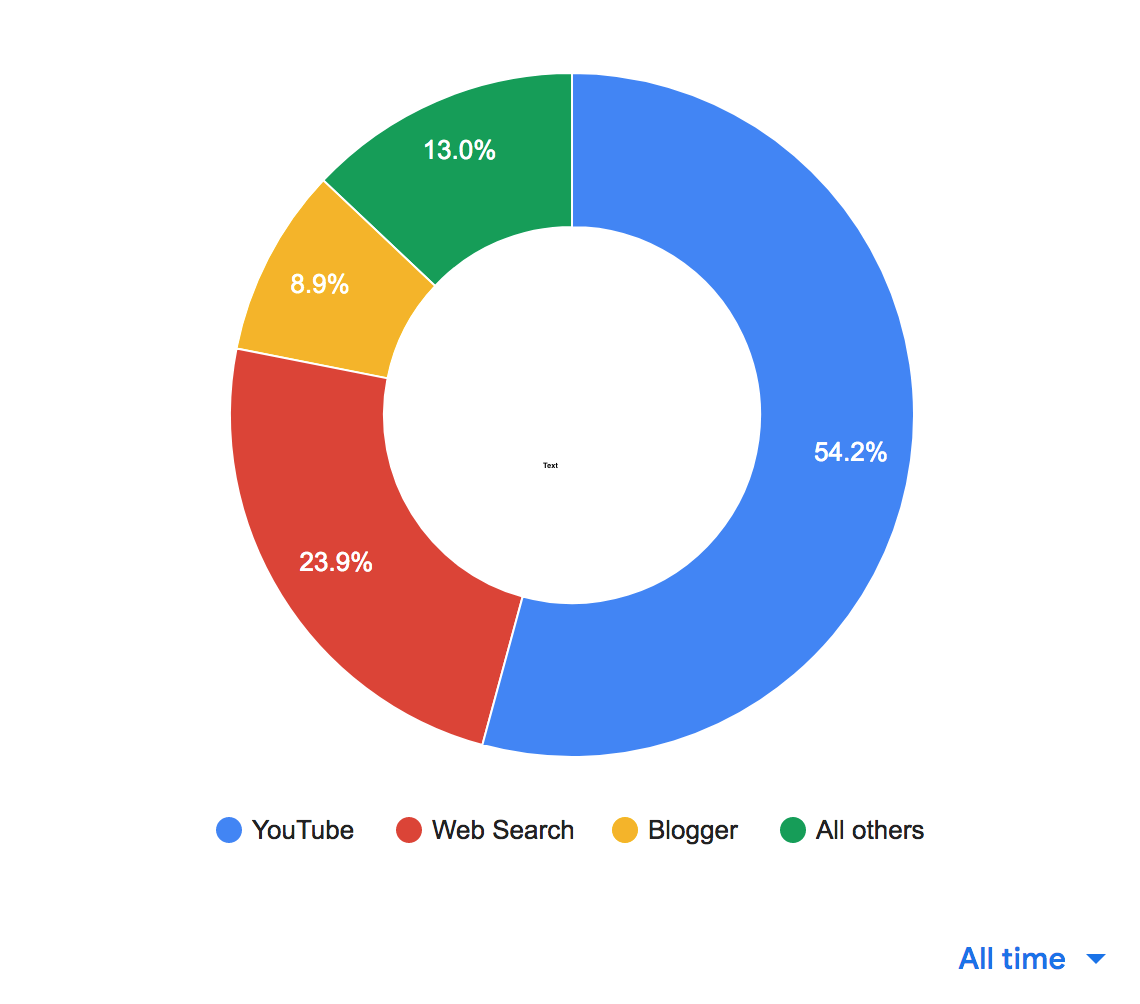
The government requests section of the report offers two datasets available for download: one that has aggregated quarterly information by country, and a more detailed dataset that has requests by content type and reason for the request. Unfortunately, neither dataset includes fields that would allow someone to identify the specific content being requested for removal. However, Google has published some of the more contentious requests that seem to be aimed at censorship as “items of public interest.”

The two data files contain the following fields:

* Period Ending: the end date of the quarter
* Country
* CLDR Territory Code
* Product: Google product where content is hosted
* Reason
* Court Orders: Number of Requests
* Court Orders: Number of Items Requested to be Removed
* Court Orders: % Fully or Partially Complied With
* Other Requests (Executive, Police, etc.): Number of Requests
* Other Requests (Executive, Police, etc.): Number of Items Requested to be Removed
* Other Requests (Executive, Police, etc.): % Fully or Partially Complied With
* All Requests: Number of Requests
* All Requests: Number of Items Requested to be Removed
* All Requests: % Fully or Partially Complied With

In their most recent report, Google provided the following breakdown of reasons for requesting that content be removed, withheld, or delisted:[[31]](#footnote-31)



Similarly, Google’s report showed that out of their suite of products, the majority of requests were regarding content hosted on YouTube.

**Twitter Transparency Report: Government Removal Requests[[32]](#footnote-32)**

Similar to the report by Google, Twitter publishes a biannual Transparency Report that includes information about requests to remove content from Twitter, either by withholding specific tweets from public view or by suspending accounts entirely. The data from the Twitter report is divided into 6 month sections going back to 2012, each of which contain aggregated data with the following fields:

* Country
* Court Orders
* Other Legal Demands: Other legal compliance reasons
* Percentage of Legal Demands Where Some Content Withheld
* Accounts Specified
* Accounts Withheld
* Tweets Withheld
* Accounts (TOS): Accounts violating Twitter’s Terms of Service

**Facebook Transparency Report: Content Restrictions[[33]](#footnote-33)**

Since in 2013, Facebook has released a biannual Transparency Report, much like Google Twitter’s reports, providing aggregated data at the country-level about removal requests and content withholding. However, unlike Google and Twitter’s reports, Facebook does not report the number of requests, only the number of content items withheld. An important distinction is that Facebook’s dataset reports on the number of removal requests from government entities that *did not violate Facebook or Instagram’s Community Guidelines*, but did violate local laws, and were ultimately withheld in the requesting country. This dataset is simple and straightforward, containing only the country name and a count of the items withheld.

*2.2 Datasets with Related Descriptive Features*

**Facebook Transparency Report: Internet Disruptions[[34]](#footnote-34)**

In the second half of 2016, Facebook started releasing a biannual report of countries where they had observed internet service disruptions. These internet disruptions were widespread, with internet access turned off for whole regions, and sometimes whole countries, for a period of time. Alongside this data, Facebook links to a report[[35]](#footnote-35) by the Global Network Initiative, of which Facebook is a member, where they provided a third-party investigation into internet service disruptions. This report makes it clear that many of these service disruptions appear to be deliberate, government-mandated attempts to restrict the flow of information and communication. Facebook’s internet disruptions dataset includes country name and a count of service disruptions for H2 2016 – H1 2017. In the second half of 2017, they started reporting the time stamped start and end dates of each of the internet service disruptions and commentary on which regions/provinces they were observing the disruptions.

**The World Bank: World Development Indicators[[36]](#footnote-36)**

The World Development Indicators (WDI) are a popular dataset containing over 1,600 country-level features pertaining to economic, population, social, health, poverty, inequality, environmental, and labor topics. The dataset is updated annually, drawing from information reported by countries’ national statistical agencies, central banks, and by international organizations. Not every country reports every metric every year, but the feature coverage tends to be fairly comprehensive. I downloaded and used the data from 2011 to 2018, since many of the 2019 metrics had not yet been reported at the time of writing.

*2.3 Measures of Freedom of Speech and Freedom of the Press*

**The World Bank: Worldwide Governance Indicators[[37]](#footnote-37)**

The World Bank also releases an annual dataset of Worldwide Governance Indicators (WGI), which contains country-level scores for six dimensions of governance for 214 countries. The full dataset includes the following fields:

* Country
* Voice and Accountability
* Political Stability and Absence of Violence
* Government Effectiveness
* Regulatory Quality
* Rule of Law
* Control of Corruption

The WGI “combine the views of a large number of enterprise, citizen and expert survey respondents in industrial and developing countries”[[38]](#footnote-38) and come from more than 30 data sources produced by “a variety of survey institutes, think tanks, non-governmental organizations, international organizations, and private sector firms.”[[39]](#footnote-39) In order to combine data from different sources together, The World Bank assigns survey questions to one of the six governance categories, rescales the data to range from 0 to 1, with higher scores meaning better outcomes. They then use an “Unobserved Components Model” (UCM) to get a weighted average of the data from each source based on how much the different sources are correlated, then standardize the data for all countries to a normal distribution.

One of the most common criticisms of the World Governance Indicators is that the World Bank does not release enough information about how the underlying datasets were weighted to create the scores. However, an external analysis found the scores to be reproducible and robust for EU countries. [[40]](#footnote-40) The “Voice and Accountability” score is derived from survey data related to freedom of speech, freedom of expression, and freedom of the press; I used data from 2011 to 2018, since the 2019 data has not yet been released as of May 2020.

**Reporters Sans Frontières: World Press Freedom Index[[41]](#footnote-41)**

Reporters Without Borders (RSF) conducts an annual survey that asks journalists, lawyers, and sociologists about freedom of the press and the safety of journalists in their respective countries. RSF selects these experts and invites them to participate in the survey. RSF also maintains a network of in-country correspondents who are tasked with keeping a tally of abuses against journalists. This “abuses indicator” is used to weight the survey results, giving heavier weight to responses from countries with more acts of violence against journalists.

The survey questions cover a range of topics, including Pluralism, Media Independence, Environment and Self-Censorship, Legislative Framework, Transparency, News/Information Infrastructure, and Abuses. The survey results are used to calculate a score from 0 to 100 for each of the above categories, with lower scores corresponding to better conditions for journalists. These intermediary scores are then combined into an overall country-level score, with the Abuses score being weighted based on the amount of time that imprisoned journalists have been imprisoned. The World Press Freedom Index has been calculated using this methodology since 2012. Reporters Without Borders sometimes considered to have a French-centric bias; the organization was founded and is based in France, has an almost entirely French board of directors, and receives substantial funding from the French government’s national development agency.[[42]](#footnote-42)

**Freedom House: “Freedom in the World” survey data[[43]](#footnote-43)**

Freedom House is a US-based NGO, policy research, and advocacy organization that publishes an annual “Freedom in the World” report on the status of freedom, democracy, political rights, and civil liberties in 195 countries. Freedom House’s index scores come from a survey that they conduct each year, using categories from the United Nations’ Universal Declaration on Human Rights to evaluate freedom and personal rights. Freedom House also publishes an annual “Freedom on the Net” report, but they have only reported scores for a hand-picked group of 15-65 countries depending on the year, and the methodology for that report relies on a report from, in some cases, a single expert in-country.[[44]](#footnote-44) Because the “Freedom on the Net” scores contain fewer data points and the methodology is even more qualitative than for the “Freedom in the World” scores, I opted to use the “Freedom in the World” scores. The “Freedom in the World” scores ask about many forms of political and civil liberties that could help build a more complete picture of the countries I will be clustering, and the survey still includes questions about freedom of expression on the internet. One common criticism of Freedom House and its index scores are that the organization has American-centric biases; Freedom House was founded in the United States during the start of the Cold War and is funded in large part by the US Agency for International Development (USAID) and US State Department.[[45]](#footnote-45) [[46]](#footnote-46)

1. **Analysis**

*3.1 Preprocessing*

I cleaned and combined all of the datasets listed above by country and year. Since the removal request data and Facebook’s internet disruptions data are reported twice a year and the index scores and country-level level indicators are reported once a year, I made one dataset of just the removal request data at half year intervals to use for comparisons between the removal request features, then made a second dataset where I rolled the removal request raw data up to the full year, recalculated any rates or percentages, then combined it with the development indicators and index scores.

I rescaled the Press Freedom Index scores so that a higher score corresponds to more freedom, which is the same direction of scale as the other freedom index scores. I used the WDI estimates of population and percentage of the population using the internet to calculate the estimated number of internet users per country per year. I then used those estimates to calculate the number of removal requests and content items withheld per internet user for each platform.

One important caveat to note is that Google, Twitter, and Facebook do not report their removal request data in exactly the same way. Google and Twitter report the number of requests, the number of items within the requests, and the number/percentage of requests with which the company fully or partially complied. Facebook does not report the overall number of requests, only the number of content items withheld, so we can only compare the rate of noncompliance between Google and Twitter. Similarly, Google does not report on the number of content items withheld or listed, only the number of requests they complied with in full or in part. In all three datasets, the amount of items withheld/requests complied with do not include items that violated the platform’s Community Guidelines.

There were also differences in the timeframes reported. Google’s data went back the furthest: to 2010. However, Google mentions in their Transparency Report that the way the data was collected and categories was not so reliable that first year, so I only used the data from 2010 to the first half of 2019, which was the latest reported. Twitter’s data was reported from 2012 to the first half of 2019, and Facebook’s data spanned from 2013 to 2019. Facebook started reporting its internet outage data in late 2016, however, the reports from H2 2016 and H1 2017 only contained counts of the number of outages; they started reporting the length of outages in H2 2017. I downloaded both sets of World Bank data from 2011 to 2019, but did not end up using the 2019 data because not enough values had been reported yet. The Freedom House and Reporters Without Borders data spanned from 2012 to 2019.

I handled missing data by filling in null values in the WDI data with the mean by country. I did not impute any values for countries that were missing one or more of the index scores (there were not many of these); I limited my analysis to the subset of countries that had all of the index scores and estimate of internet users in the country, because I used that to calculate “per internet user” versions of the removal request metrics. I only filled in missing values for the removal requests with zeroes, and only when it was required for the PCA and clustering, taking care not to fill in zeros for any rates or percentages.

*3.2 Data Analysis Process and Tactics*

**Trend Analysis and Initial Regressions**

To get an initial sense of the relationships between the removal request variables and the index scores, I ran a series of linear regressions where I regressed each of the scores and sub-scores pertaining to free speech on the removal request data. In several of my linear models, I tried to account for affects correlated with time, but because I only had a few years of data to work with, the most I could realistically do to account for time was take the first difference of the data; higher order differencing would likely have left too few data points for any results to be useful.

**Principal Component Analysis**

From the regression results, it seemed likely that the internet disruption data had a strongly negative correlation with the freedom index scores, which intuitively makes sense, so I ran my PCAs and cluster analysis on the data from 2016 to 2018 where the internet disruption data was included. This subset of data also had the most countries available to cluster, and are inherently more relevant because those years are the most recent.

Because the combined dataset contained a fairly large number of features with multicollinearity, I used Principal Component Analysis (PCA) to reduce the number of dimensions and gain a better understanding of how the development indicators, freedom and governance index scores, and removal request metrics relate to each other. To avoid comparing different time periods, I ran three PCAs: one for 2016, one for 2017, and one for 2018. For each analysis, I scaled the data, started out by inputting the same number of principal components as features, calculated the amount of variance explained by each principal component from the eigenvalues, then used a scree plot to decide on the number of principal components to use going forward. Then I looked at the loadings for each principal component to better understand which features were most explanatory and most differentiating between countries.

**Hierarchical Clustering**

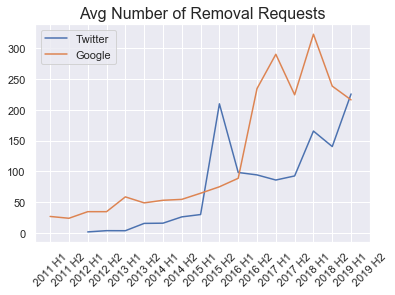
To address my first research question, I used hierarchical clustering with my principal components as input variables to see whether countries with similar removal request activity and similar approaches to free speech clustered together. I produced three sets of clusters: one for 2016, one for 2017, and one for 2018. For each year, I determined the number of clusters to use by calculating and plotting a range of silhouette scores, which are a measure of cluster cohesiveness; a higher silhouette score indicates more similarity within clusters and more dissimilarity between clusters. Finally, I tried plotting the countries using different combinations of the principal components to see what information could be gleaned from the way the countries were clustered.

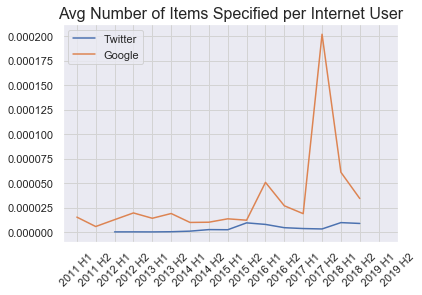
1. **Results and Discussion of Findings**

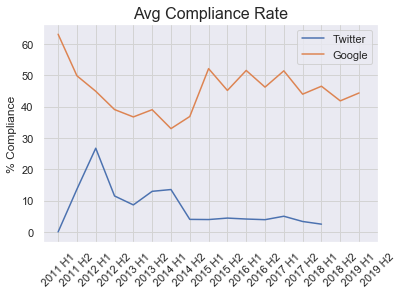
*4.1 Results*

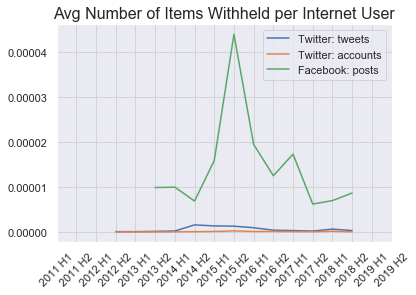
**Trend Analysis**

Upon visualizing the removal request data by 6-month interval, several trends became apparent. We can see that there is an upward trend in removal requests over time for both Google and Twitter, presumably as more people became aware of the removal request features. Facebook does not report on the number of removal requests, only the number of content items withheld that did not violate their Community Guidelines. However, Facebook saw an upward trend in withheld items, which was most likely accompanied by an increase in requests.

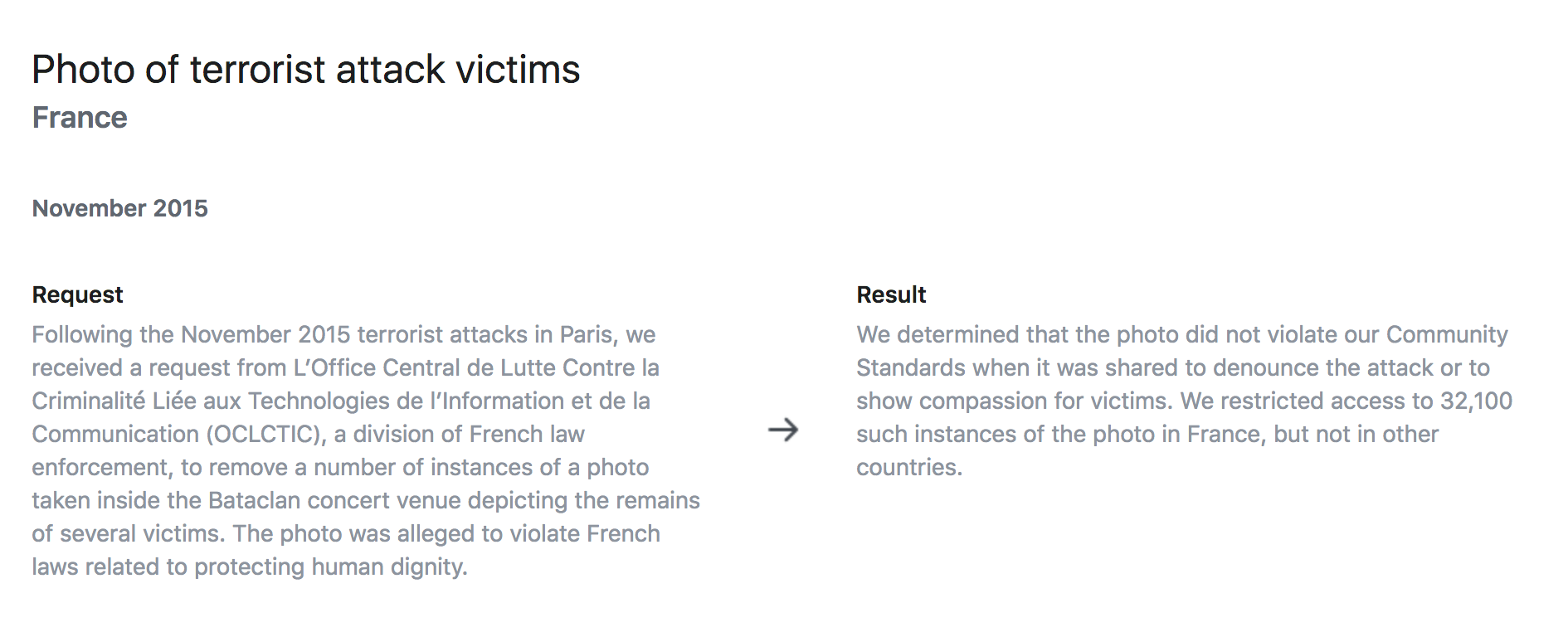




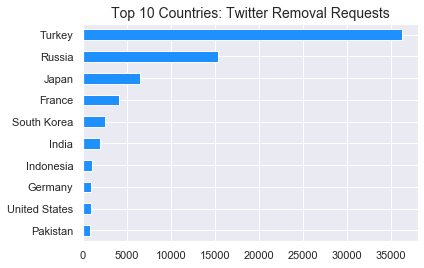


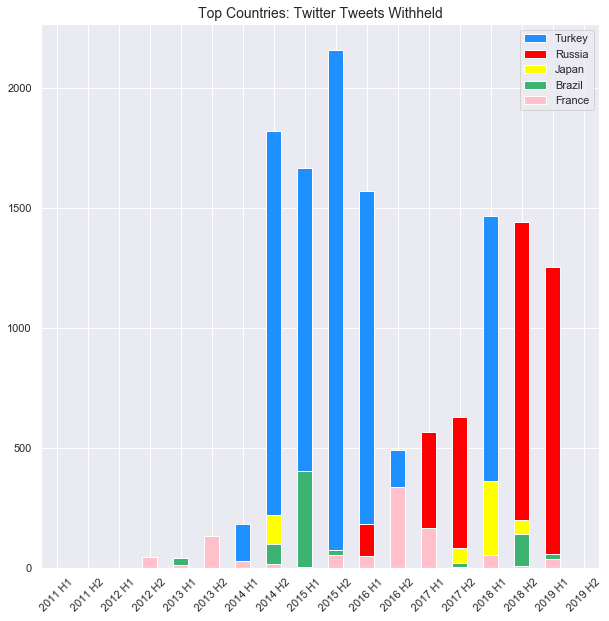


Along with the growth in the number of requests, there was also un uptick in the number of items being requested. All of these averages show a steep increase during 2015 and 2016. Given what we know about what was happening in the world during that time, this could be due to any number of factors. In the case of Facebook, their Transparency Report explains the sharp spike in withheld content in H2 of 2015. Referring back to the selection of text descriptions of what some of the removal requests were about, Facebook mentioned that they removed thousands of copies of a photo related to the terrorist attack at the Bataclan in Paris that year at the French government’s request.[[47]](#footnote-47) Google and Twitter may have received removal requests from France for the same reason around the same time; both had an increase in requests from France, but not as dramatic of an increase as observed on Facebook, and neither Google nor Twitter included a text description related to the terrorist attack.

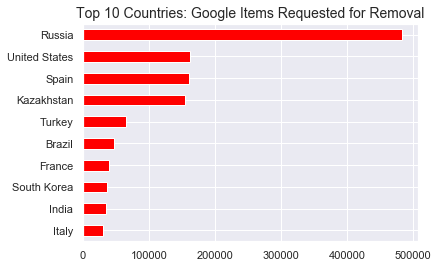


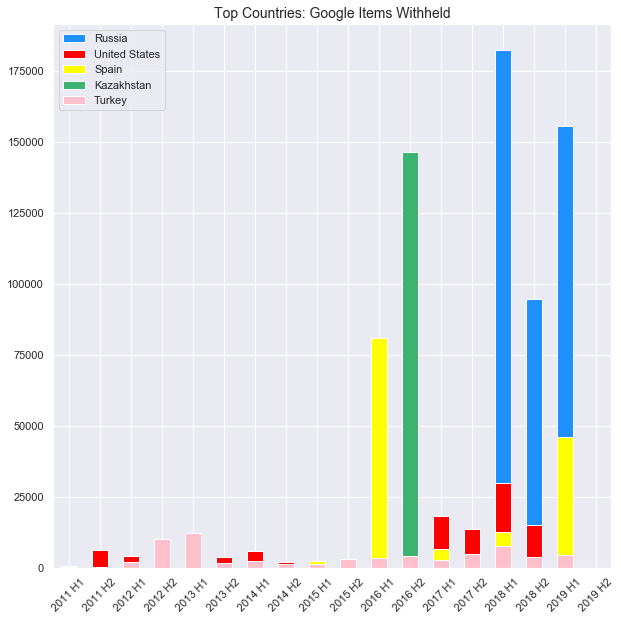
Next, I wanted to see which countries were the heaviest users of the removal request features.

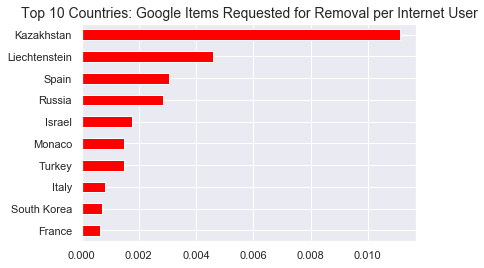




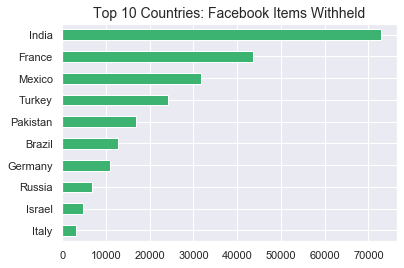
By far, Turkey was the most frequent removal requester on Twitter, followed by Russia. Turkey and Russia were also the top two countries to have tweets they requested withheld to comply with local laws. Expressed in terms of tweets withheld per internet user, Turkey, Israel, and Russia were the top three. Interestingly, the countries that requested the highest percentage of items that violated Twitter’s Terms of Service (which would have been removed anyway) were the United Arab Emirates, South Korea, the United States, and Japan.

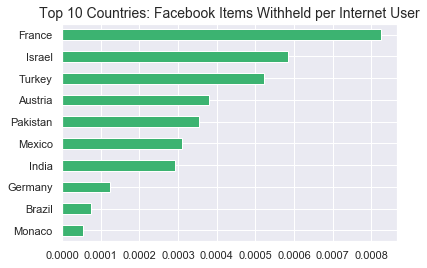


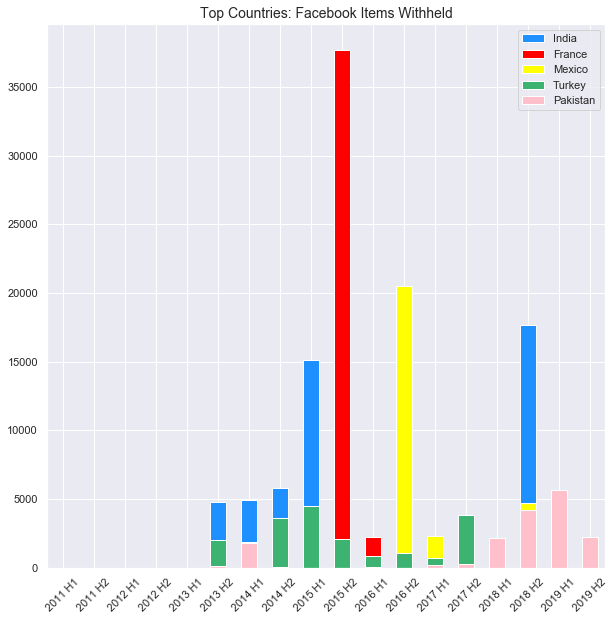




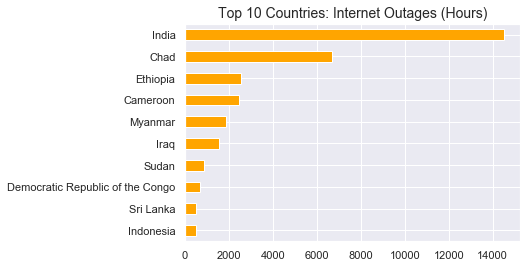
On Google, Russia, the United States Spain, and Spain were the top three removal requesters, but on a per internet user basis, Kazakhstan was the top country by far, followed by Liechtenstein and Spain. Plotted over time, we can see that the United States and Turkey had a fairly consistent number of items withheld during each reporting period, while Russia and Spain started having a large number of items withheld in the last few reporting periods. Kazakhstan’s spike in items withheld all came through in one six month period in 2016.







On Facebook, India had the largest number of items withheld, followed by France, Mexico, Turkey, and Pakistan. On a per internet user basis, France had the most items withheld (due to the terrorist attack victim photos mentioned above), followed by Israel, Turkey, Austria, and Pakistan.

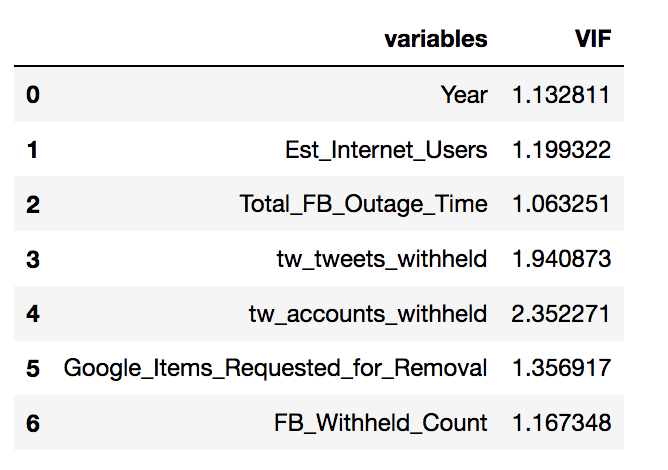


India, and Chad had the longest internet outages by far. Looking back at the comments Facebook provided in the dataset, it is clear that all of the outages in India were in and around the conflict zones in Kashmir. Interestingly, Pakistan had 105 hours of internet outages between 2016 and 2019, compared to India’s 14,508 hours of outages.

**Regression Results**

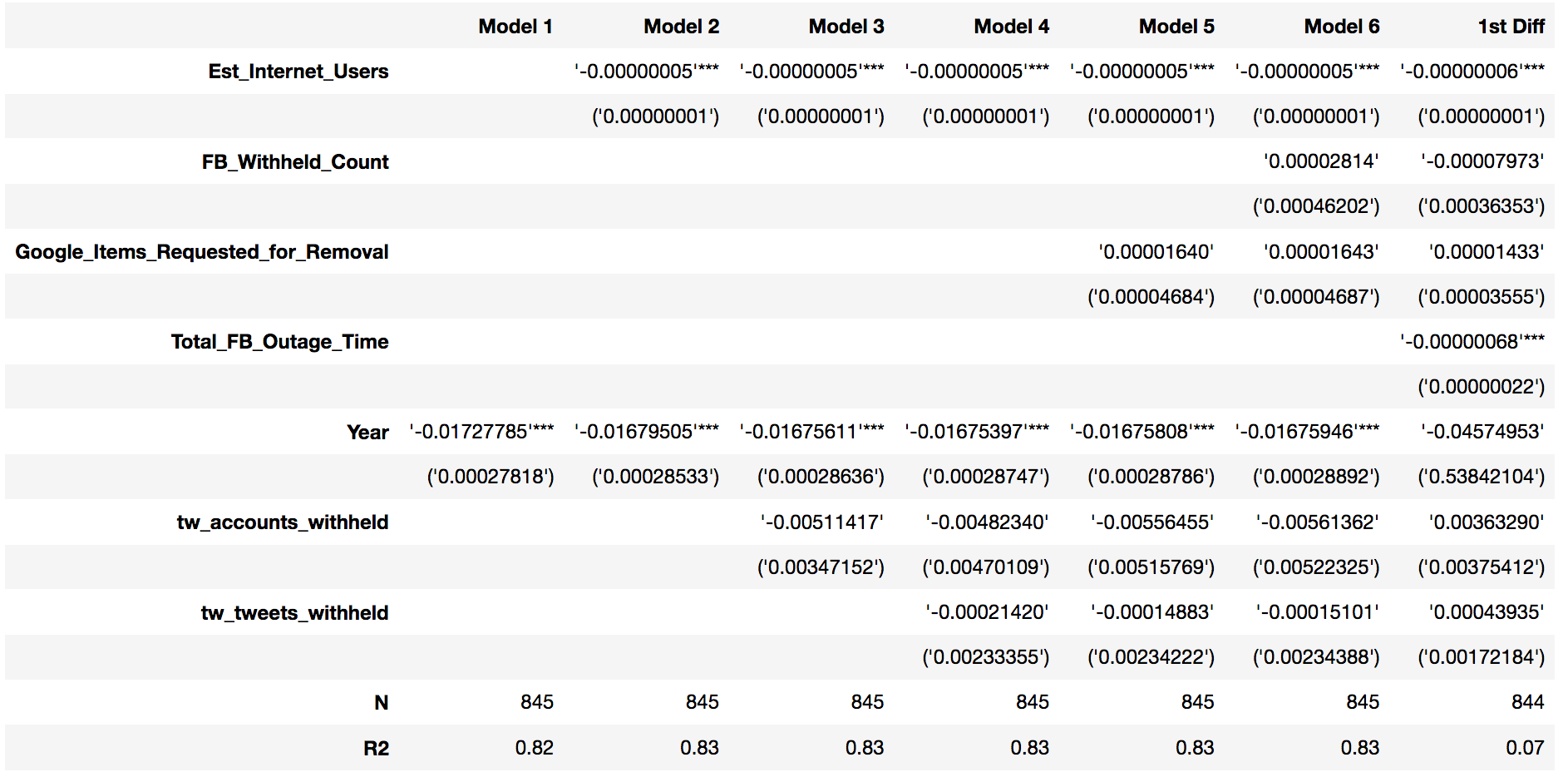
For my initial linear models, I ran a series of naïve OLS regressions predicting each of the different freedom of speech scores using only the removal request and time variables to try to get a sense for how they are related to one another. Since there are only 7-9 years of data for each for each type of content removal request, we are not able to do a proper, in-depth time series analysis. However, I first-differenced the data to try to mitigate some of the effects related to specific points in time.

First, I checked the variance inflation factor (VIF) to see if there was multicollinearity between any of the removal request variables I chose; I knew that requests and items withheld would be correlated, so I chose the withheld variables instead of requests.

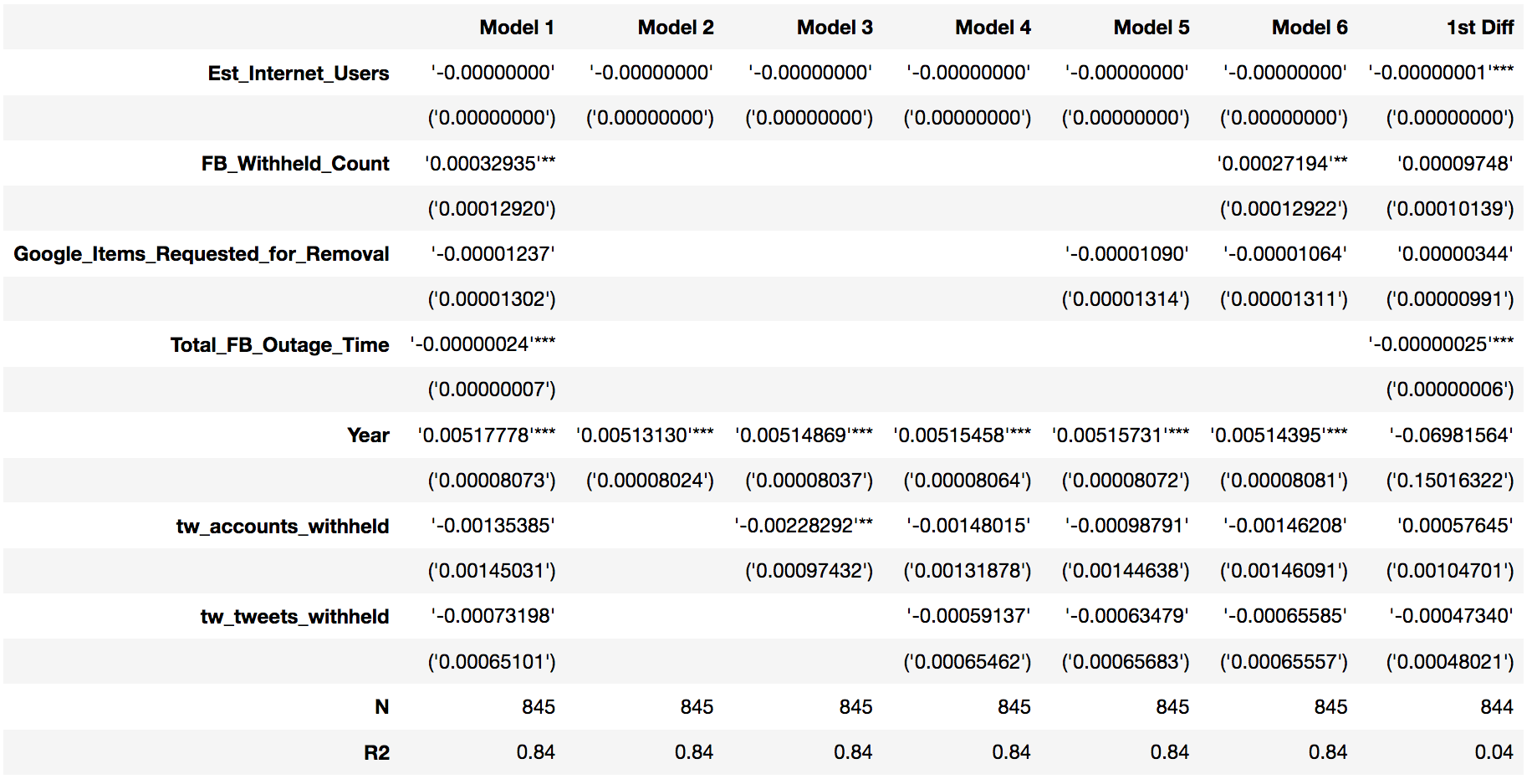


Luckily, there was not enough multicollinearity to be problematic.

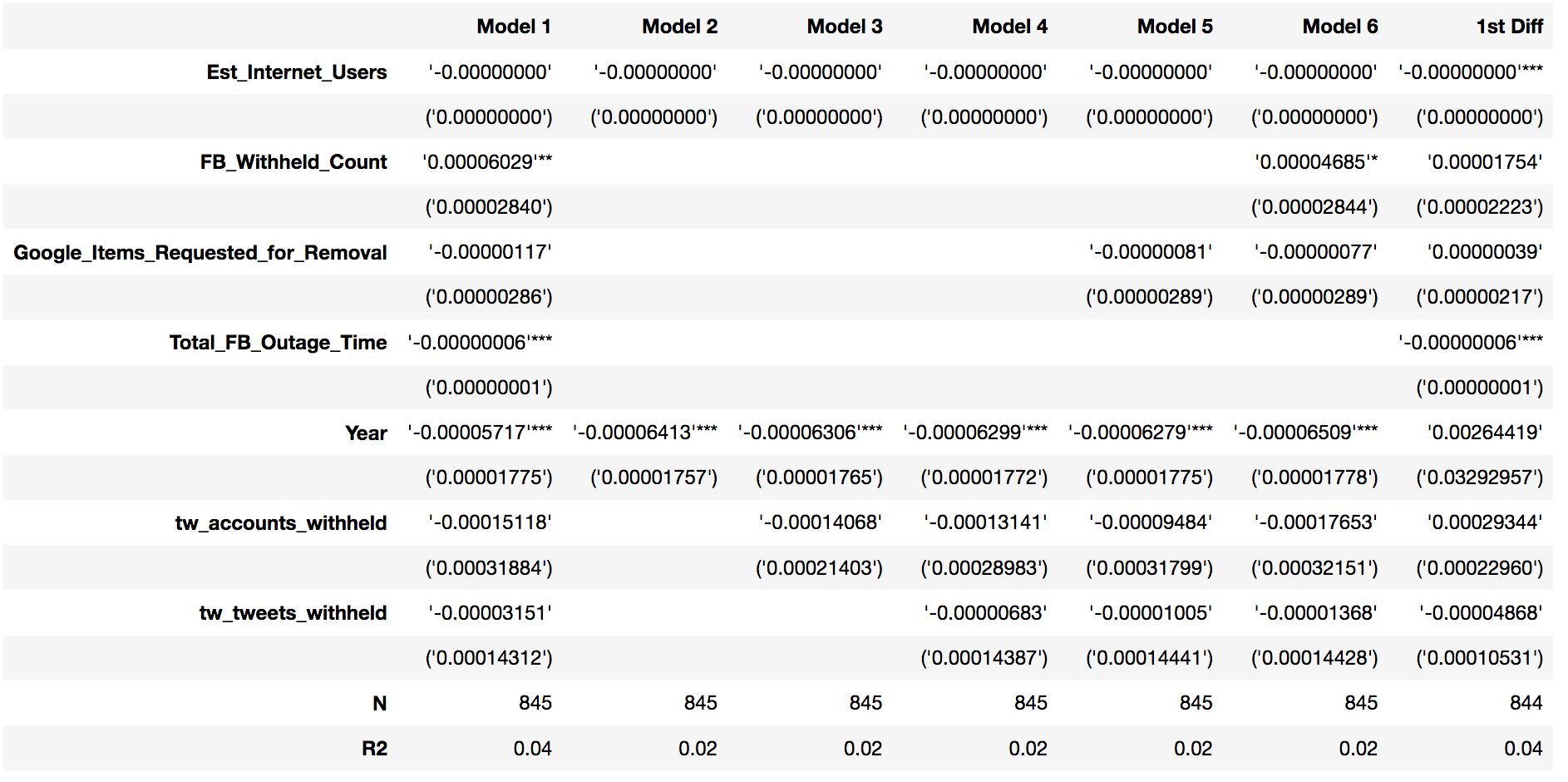
Predicting RSF Press Freedom Index:



Predicting Freedom House “Freedom of Expression and Belief” sub-score:



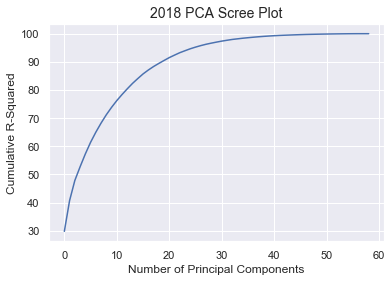
Predicting World Bank “Voice and Accountability” score:

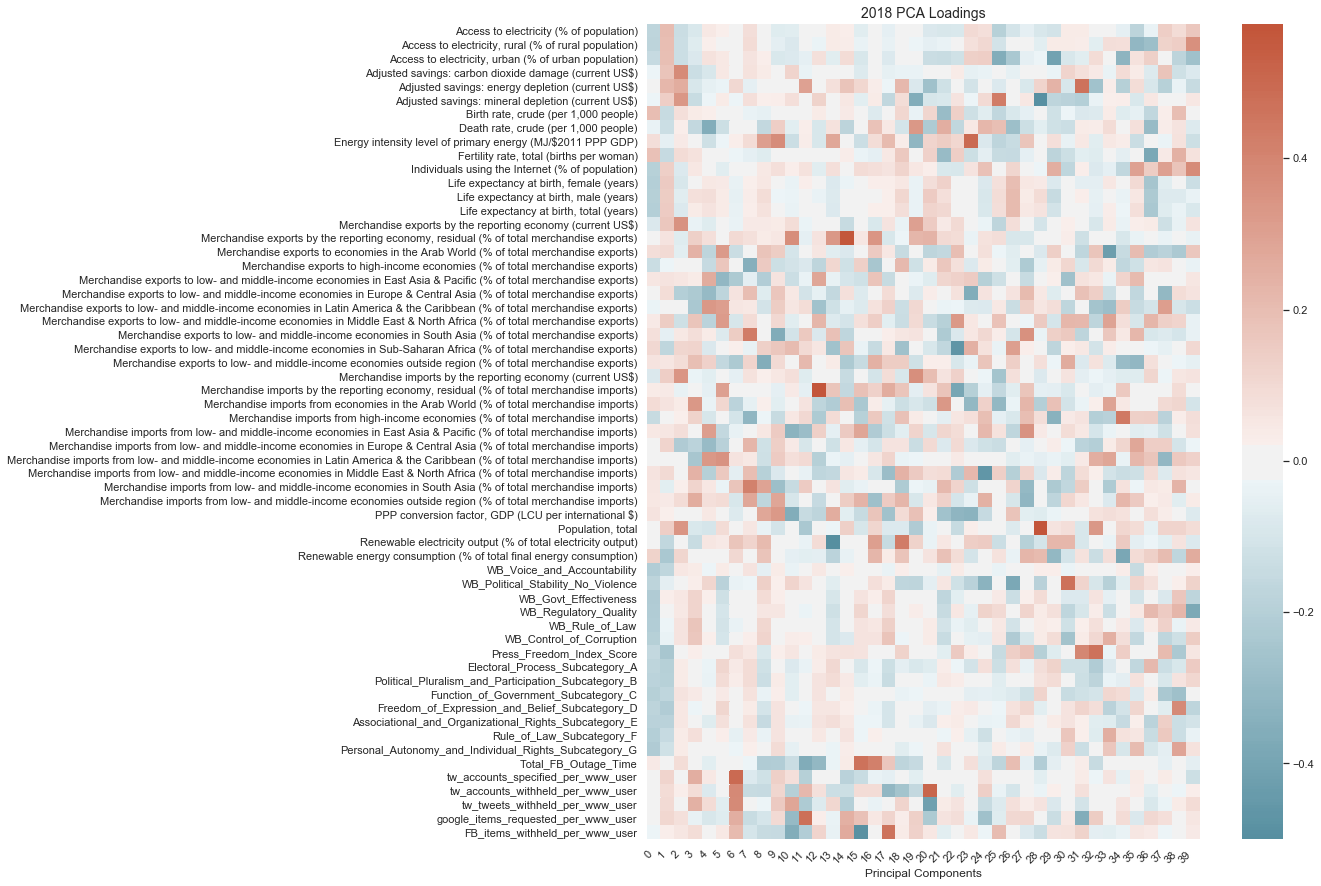


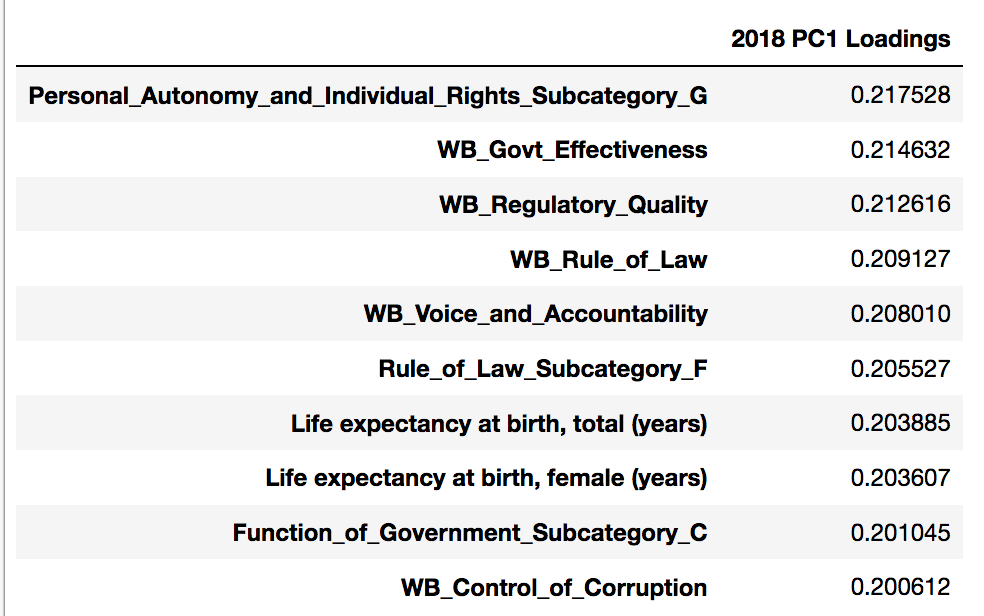
Interestingly, it seems like none of the removal request variables were particularly good predictors of any of the free speech index scores. However, the coefficient on internet outage time was small, but statistically significant, so that variable could be an effective predictor of the level of free speech (which makes sense, intuitively).

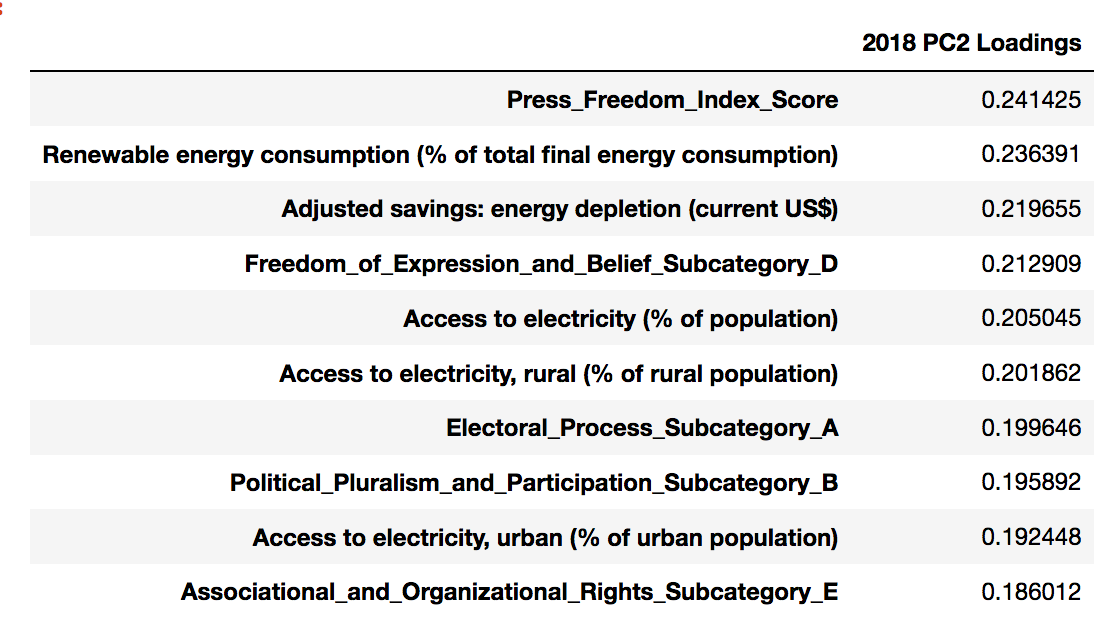
**PCA and Hierarchical Clustering Results**

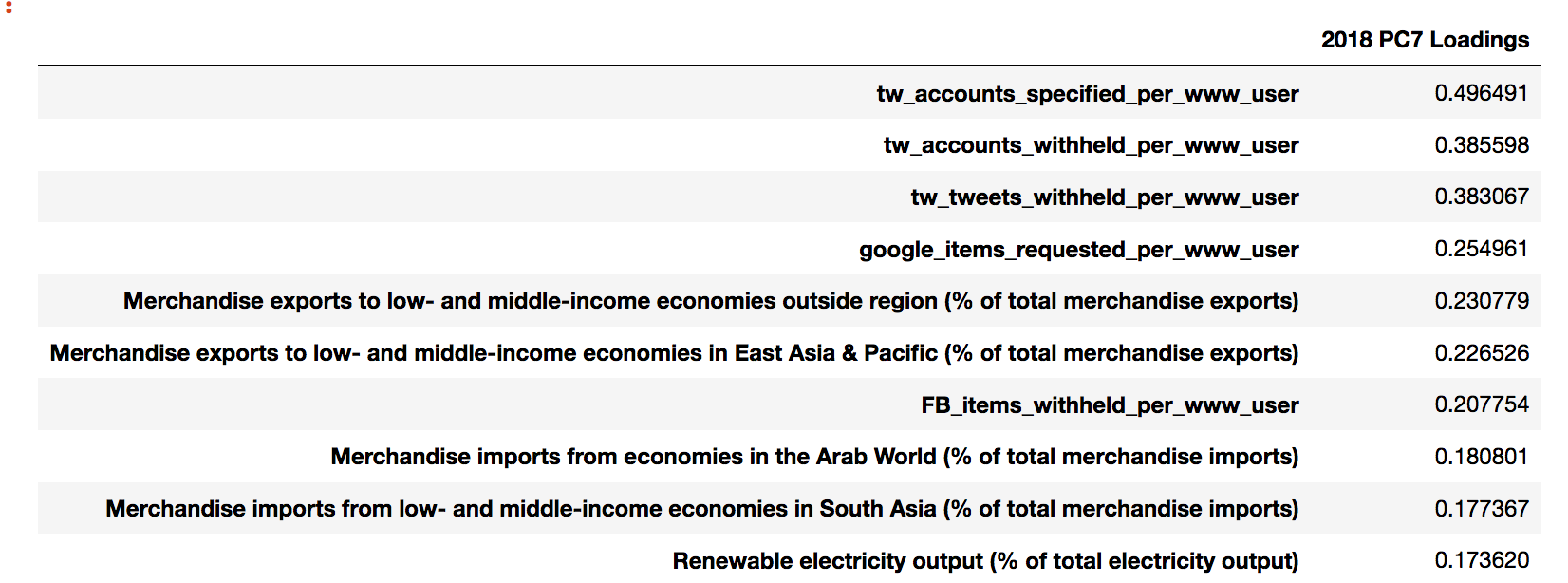
Since the regression results showed that internet outage time could be a better predictor of free speech than the removal requests, I wanted to make sure I included that variable in my PCA and cluster analysis. So I ran PCA and clustered the data for 2016, 2017, and 2018. Overall, the 2016, 2017, and 2018 data reduced down to similarly comprised principal components, and the countries clustered very similarly from year to year. The 2018 plots are shown here because that dataset includes more countries than the previous two years. After dropping null values that could not be filled in using averages by country and the raw values for metrics that were also expressed as percentages, 59 total features remained. The algorithm reduced those features down to 40 principal components, which cumulatively explained nearly all of the variance in the data.

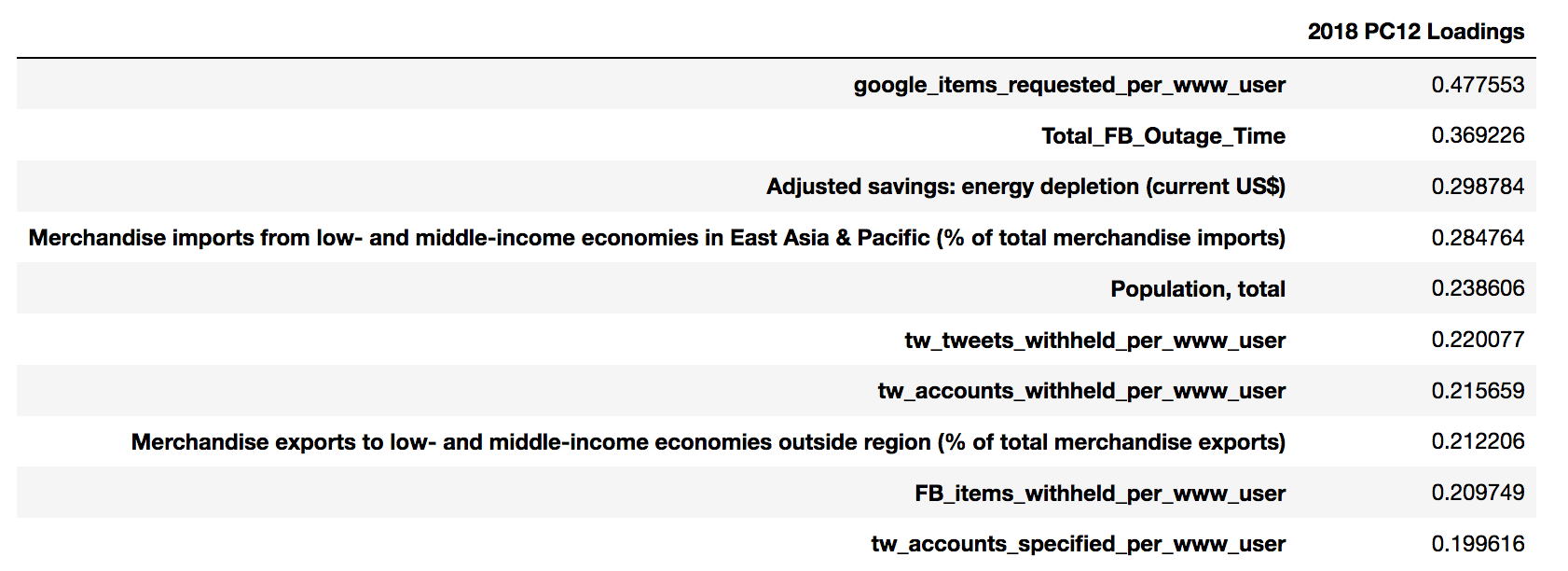




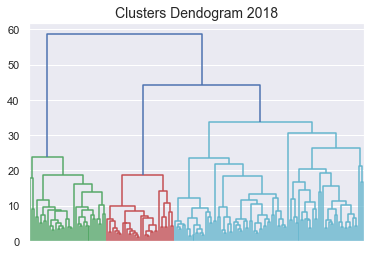


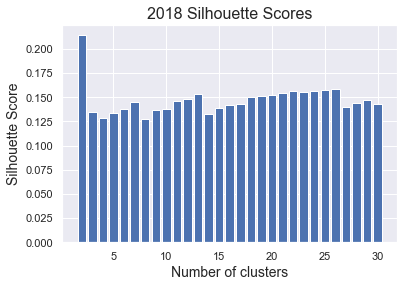






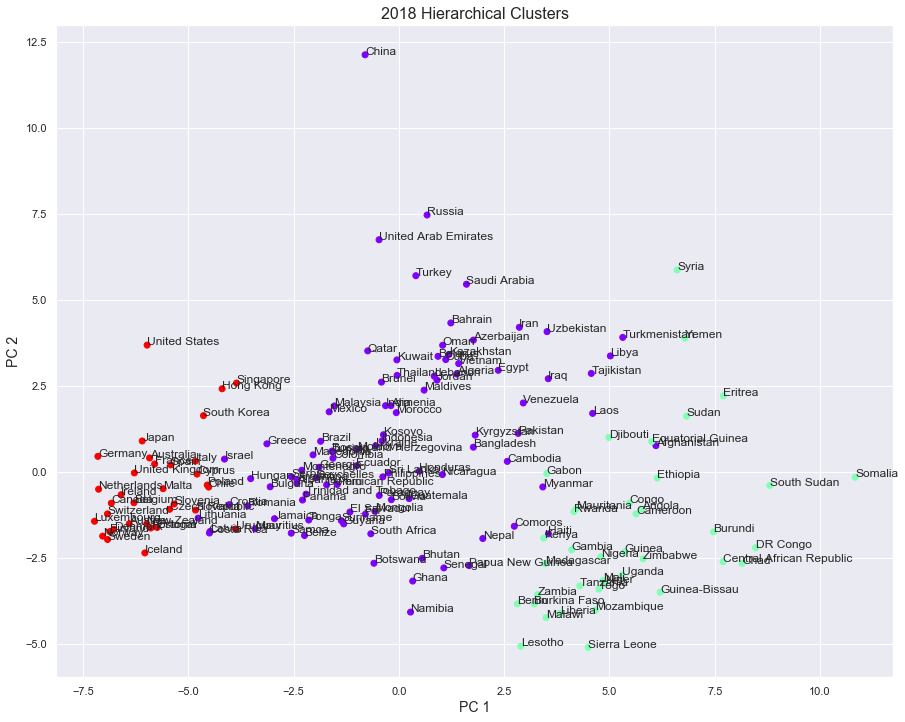
As we can see on the heat map of the PCA factor loadings, the top 10 loadings that were largest in magnitude for the first principal component were on the Freedom House “Personal Autonomy and Individual Rights” sub-score, followed by several of the World Bank Governance Indicators, including the “Voice and Accountability” score measuring free speech. The second principal component was most heavily comprised of the RSF Press Freedom Index Score, several development indicators, and the Freedom House “Freedom of Expression and Belief” sub-score. I also took a closer look at the loadings of the 7th and 12th principal components, where the removal request and internet outage variables had the largest loadings.

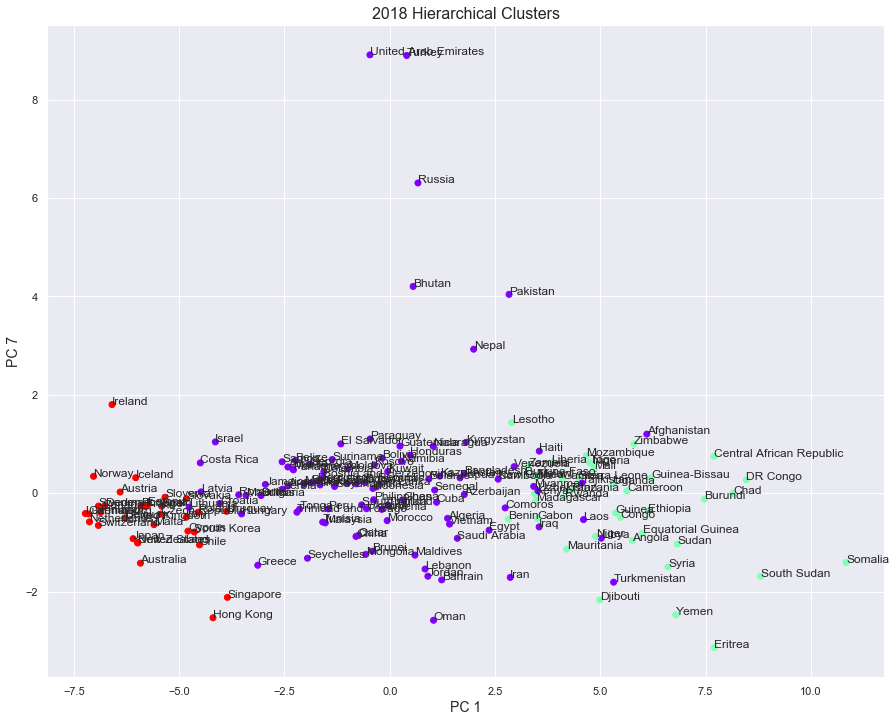


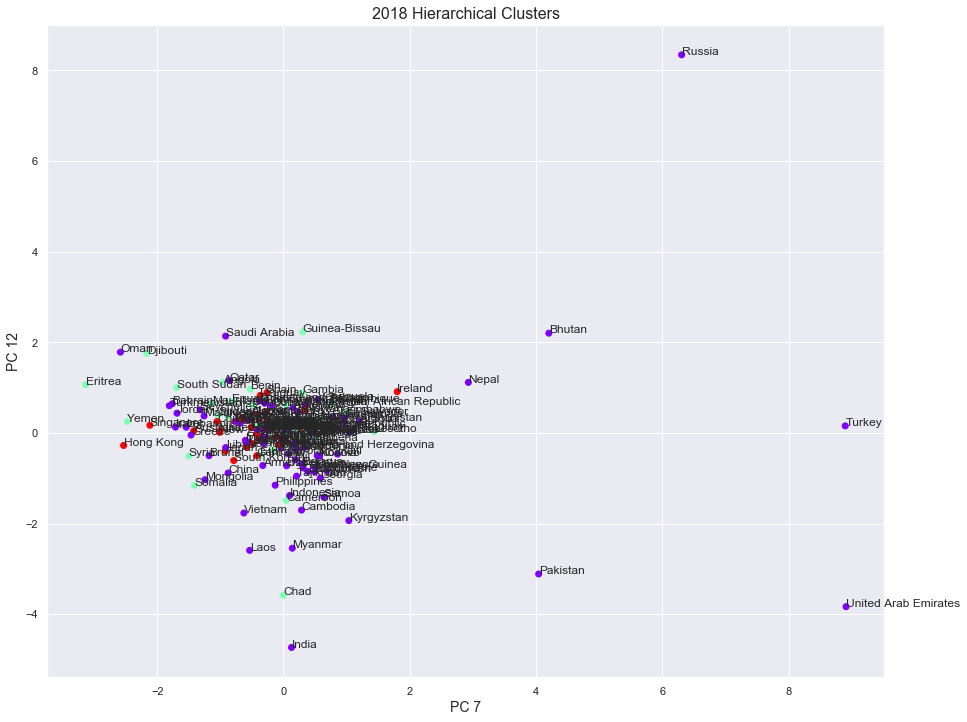


Using Ward linkage and the Euclidean distance measure, I created a dendogram and plot of the silhouette scores to decide on the optimal number of cluster. For the 2016 and 2017 data, both the dendograms and silhouette scores pointed 3 clusters as being most optimal; but, interestingly, the 2018 data did not seem to cluster as neatly. The dendogram suggests that three clusters is optimal, but that the distance between two of the clusters is not very large. The silhouette scores point to two clusters as being optimal compared to three. I tried both options, but found that two clusters was not very helpful, because the countries essentially just clustered into groups of developed and less-developed countries. That does not add anything to our understanding of the data, so I moved forward using three clusters, since that seemed to form distinct enough clusters based on the dendogram anyway. The way that the countries clustered was almost identical to the way they had clustered in the 2016 and 2017 data, but it seems that additional variability in some of the underlying data led to those same clusters having less intra-cluster similarity and less inter-cluster dissimilarity than in 2016 and 2017. Ultimately, my silhouette scores for three clusters were 0.214, 0.215, and .0135 for 2016, 2017, and 2018, respectively.

The graphs below show the clearest separation of the clusters, which were plotted using the 1st, 2nd, 7th, and 12th principal components.





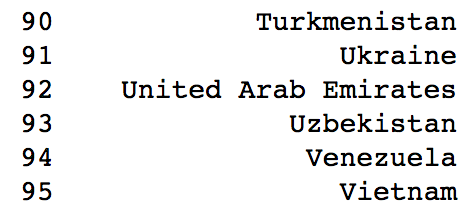
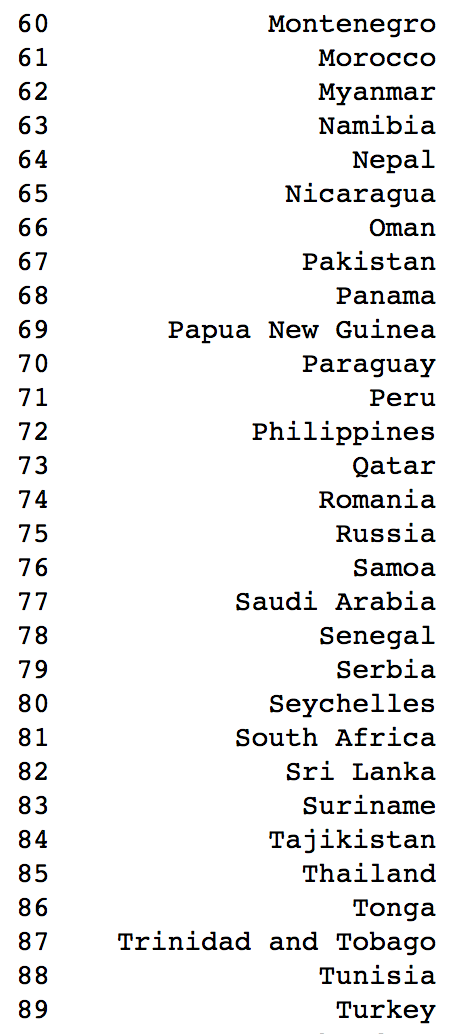
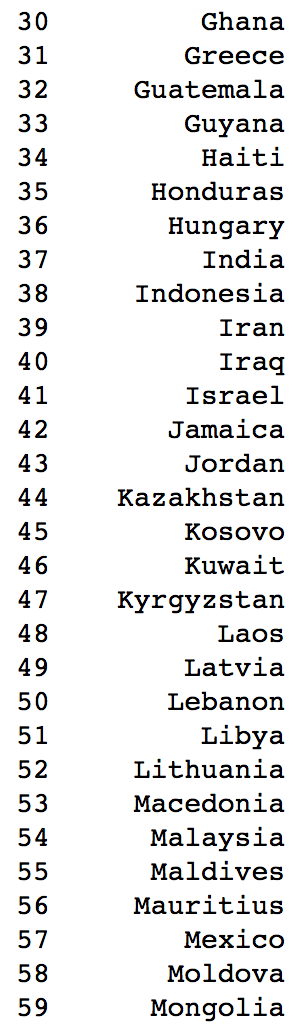


*4.2 Discussion of Findings*

We can see that the three clusters generally split along the lines of developed, semi-developed, and developing countries from left to right, due to the 1st principal component on the x-axis being most heavily comprised of the World Bank Governance scores and population/demographic variables. The vertical split on the first graph is related to free speech and free press, since those scores had the largest loadings in the 2nd principal component. Many of the countries that are known to filter the internet and engage in content manipulation on the internet are shown at the top-center of the first graph, including China, Russia, UAE, Turkey, and Saudi Arabia.

The second plot can be summarily described as a plot of “governance and development levels” versus “removal request activity.” We can see that the countries with the most removal request activity general fall within the middle on the strong/weak governance and developed/developing scale. On the third graph, we can easily identify the extreme outliers as the countries with an unusually high amount of removal requests per internet user, due to the feature loadings that make up the 7th and 12th principal components. However, internet outage time is the second most important feature in the 12th principal component, and we can see that the countries that had both internet outages and lots of removal requests are toward the bottom-right of the third graph.

These countries were grouped into the first (purple) cluster:

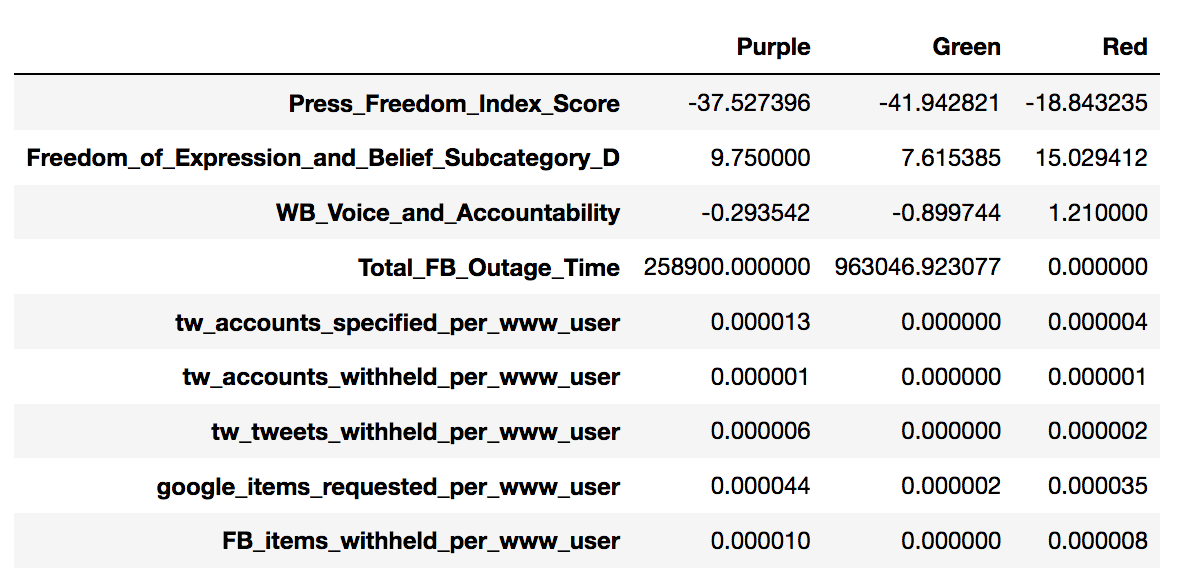


These countries were grouped into the second (green) cluster:



These countries were grouped into the third (red) cluster:





The table above shows the cluster means for each of the three groups of countries. As we can see, the red cluster has the highest average index scores for press freedom, freedom of expression and belief, and voice and accountability. The purple cluster, which ranks 1st for the number of removal requests and content withheld, ranks 2nd for the freedom index scores. The green cluster ranks 3rd for freedom index scores and had almost no removal requests on average, but ranked 1st for internet outage time. The purple cluster also had some internet outages, but the green cluster had outages roughly three times as long, on average. The red cluster had no internet outages.

Essentially, we can see that the heaviest users of the content removal request features tend to cluster in the purple group and so do many of the countries that are known to filter the internet. Some of these countries have shown evidence in other research of using content manipulation tactics on the internet to silence dissidents (e.g. Turkey, Russia, China, Honduras); other countries in this group have their religious and/or moral values codified into law (e.g. UAE, Kuwait, Pakistan). Other countries in the purple group do not have a particularly high number of removal requests and are not necessarily filtering the internet in malicious or culturally-strict ways, but just generally fall into the group of semi-developed countries with mid-level government functioning. Within the red cluster, we can see that most of these countries are liberal democracies, and that some of them have strict anti-hate speech laws (France, Germany, Ireland), copyright laws (United States), and/or have tried very hard to prevent violent content on the internet from reaching their citizens (France, Australia); these laws and behaviors are associated with more removal requests.

*4.3 Study Limitations and Areas for Future Research*

The main limitations of this project had to do with the data, mainly the short timeframe of removal request data available. It would be interesting to rerun these models after several more years have passed to see if any additional trends can be identified. If additional years of data had been available, I would have tried using more in-depth time series techniques, for example, ARIMA modelling and checking for unit roots, to try to isolate any effects that may be significant separate from changes due to time. Another option would be to try using Dynamic Time Warping to calculate the distance between the different time series, then create a custom distance matrix that can be used for hierarchical clustering. I ran my PCAs and clustered using data from one year at a time, but accounting for non-linear temporal trends in this way would allow the for clustering over time.

Other issues with the data were the decisions and estimates I had to make due to missing data. I had to calculate the per person removal request metrics using the estimated number of internet users per country, per year. Surprisingly, I was unable to find consistent usage numbers of each web platform by country; this is proprietary information, and the comparisons made in the base data would be more precise if the number of Twitter, Facebook, and YouTube users were known.

In future iterations of this project, I would also try incorporating other measures of free speech and freedom of expression. As outlined in earlier sections, I chose to use the freedom index scores that I did based on the fact that they are commonly used. However, these scores are problematic for a number of reasons due to the methodologies they use and the biases they convey, and they should certainly not be considered as a source of truth in determining freedom of speech. Another angle would be to directly compare these countries’ legal frameworks in terms of how they interpret and/or allow free speech and the ways in which limits on free speech are enforced on the internet.

## **References**

Bambauer, Derek E. 2009. “Cybersieves.” *Duke Law Journal*, no. Vol. 59, Num. 3 (December). <https://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=1443&context=dlj>.

Bauer, Joseph P. 2010. “Copyright and the First Amendment:  Comrades, Combatants, or Uneasy Allies? .” *Washington and Lee Law Review*, no. Vol. 67, Issue 3: p831-914. <http://law2.wlu.edu/deptimages/Law%20Review/67-3Bauer.pdf>.

Bogle, Ariel. 2019. “Laws Targeting Terror Videos on Facebook and YouTube ‘rushed’ and ‘Knee-Jerk’, Lawyers and Tech Industry Say - ABC News.” ABC (Australian Broadcasting Corporation). April 3, 2019. <https://www.abc.net.au/news/science/2019-04-04/facebook-youtube-social-media-laws-rushed-and-flawed-critics-say/10965812>.

Charron, Nicholas. n.d. “Assessing The Quality of the Quality of Government Data: A Sensitivity Test of the World Bank Government Indicators.” The Quality of Government Institute - The University of Gothenburg. Accessed May 15, 2020. <https://www.qog.pol.gu.se/digitalAssets/1357/1357980_paper-on-sensititivty-tests-of-world-bank-data.pdf>.

Facebook. n.d. “Facebook Transparency Report | Content Restrictions.” Facebook Transparency Report. Accessed May 1, 2020a. <https://transparency.facebook.com/content-restrictions>.

Facebook. n.d. “Facebook Transparency Report | Internet Disruptions.” Facebook Transparency Report. Accessed May 1, 2020b. <https://transparency.facebook.com/internet-disruptions>.

Freedom House. n.d. “Freedom in the World | Freedom House.” Freedom House. Accessed May 1, 2020a. <https://freedomhouse.org/report/freedom-world>.

Freedom House. n.d. “Freedom on the Net Research Methodology | Freedom House.” Freedom House. Accessed May 15, 2020b. <https://freedomhouse.org/reports/freedom-net/freedom-net-research-methodology>.

Gallagher, E., P. Suárez-Serrato, and E. I. Velazquez Richards. 2018. “Socialbots Whitewashing Contested Elections; A Case Study from Honduras.” In *Advances in Intelligent Systems and Computing*, 547–52. Springer Singapore. https://doi.org/10.1007/978-981-13-1165-9\_50.

Global Network Initiative. n.d. “About GNI | Global Network Initiative.” Global Network Initiative. Accessed May 16, 2020a. <https://globalnetworkinitiative.org/about-gni/>.

Global Network Initiative. n.d. “​New Report Reveals the Economic Costs of Internet Shutdowns | Global Network Initiative.” Global Network Initiative. Accessed May 15, 2020b. <https://globalnetworkinitiative.org/%E2%80%8Bnew-report-reveals-the-economic-costs-of-internet-shutdowns/>.

Google. n.d. “Google Transparency Report.” Google Transparency Report. Accessed May 1, 2020. <https://transparencyreport.google.com/government-removals>.

Heins, Marjorie. 2014. “The Brave New World of Social Media Censorship.” *The Harvard Law Review*, no. 127 (June): 325–30. <http://cdn.harvardlawreview.org/wp-content/uploads/2014/06/vol127_Heins.pdf>.

Kelion, Leo. 2019. “Google Wins Landmark Right to Be Forgotten Case - BBC News.” BBC News. BBC News. September 24, 2019. <https://www.bbc.com/news/technology-49808208>.

OpenNet Initiative. n.d. “About Filtering | OpenNet Initiative.” OpenNet Initiative. Accessed May 10, 2020a. <https://opennet.net/about-filtering>.

OpenNet Initiative. n.d. “China | OpenNet Initiative.” OpenNet Initiative. Accessed May 16, 2020b. <https://opennet.net/research/profiles/china-including-hong-kong>.

Raveendran, Nivitha, and Christopher S. Leberknight. n.d. “Internet Censorship and Economic Impacts: A Case Study of Internet Outages in India.” Internet Outage and Economic Impacts . Montclair State University. Accessed May 11, 2020. https://pdfs.semanticscholar.org/9ab7/449a70e4c221dd3d46c7002359e443b3634b.pdf.  
Presented at the twenty-fourth Americas Conference on Information Systems, New Orleans, 2018

Reporters Without Borders. n.d. “2020 World Press Freedom Index | RSF.” RSF. Accessed May 1, 2020. <https://rsf.org/en/ranking>.

Roberts, Margaret E. 2020. “Resilience to Online Censorship.” *Annual Review of Political Science* 23 (1): 401–19. https://doi.org/10.1146/annurev-polisci-050718-032837.

Schneider, Laura. 2014. “Media Freedom Indices: What They Tell Us – And What They Don’t.” BMZ Germany. January 2014. <https://www.dw.com/downloads/28985486/mediafreedomindices.pdf>.

Silverman, Craig, and Jeremy Singer-Vine. 2018. “Here’s Who’s Been Blocked By Twitter’s Country-Specific Censorship Program.” BuzzFeed News. BuzzFeed News. January 24, 2018. <https://www.buzzfeednews.com/article/craigsilverman/country-withheld-twitter-accounts>.

Singer-Vine, Jeremy. n.d. “Country-"Withheld" Twitter Accounts — BuzzFeed News - Google Sheets.” Google Docs. Accessed September 27, 2019. <https://docs.google.com/spreadsheets/d/1ABJKoDKest3ZYny99_GrZ-nFsAPzNBpehfRAAYrjjpY>.

Steiner, Nils D. 2014. “Testing for a Political Bias in Freedom House Democracy Scores: Are U.S. Friendly States Judged To Be More Democratic?” *SSRN Electronic Journal*, January. <https://doi.org/10.2139/ssrn.1919870>.

The International Press Institute. 2018. “Analysis by the Vienna-Based International Press Institute on the Digital Manhunt of Turkish Journalist Nedim Turfent.” *Bloomberg*, April 2018. <https://www.bloomberg.com/features/2018-government-sponsored-cyber-militia-cookbook/data/Turfent.pdf>.

“The Santa Clara Principles on Transparency and Accountability in Content Moderation.” n.d. Santa Clara Principles. Accessed May 16, 2020. <https://santaclaraprinciples.org/>.

The World Bank. n.d. “World Development Indicators | DataBank.” DataBank | The World Bank. Accessed May 1, 2020a. <https://databank.worldbank.org/source/world-development-indicators>.

The World Bank.. n.d. “Worldwide Governance Indicators.” WGI. Accessed May 1, 2020b. <https://info.worldbank.org/governance/wgi/>.

Twitter. n.d. “Removal Requests.” Transparency Report. Accessed May 1, 2020. <https://transparency.twitter.com/en/removal-requests.html>.

UN ONHCR. 1976. “OHCHR |  International Covenant on Civil and Political Rights.” 1976. <https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx>.

United Nations Human Rights Council. 2011. “UN General Assembly Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue.” May 16, 2011. <https://www2.ohchr.org/english/bodies/hrcouncil/docs/17session/A.HRC.17.27_en.pdf>.

YouTube. 2018. “Official YouTube Blog: Faster Removals and Tackling Comments — an Update on What We’re Doing to Enforce YouTube’s Community Guidelines.” Official YouTube Blog. December 2018. <https://youtube.googleblog.com/2018/12/faster-removals-and-tackling-comments.html>.

YouTube Help. n.d. “Community Guidelines Strike Basics - YouTube Help.” Google Help. Accessed September 30, 2019. <https://support.google.com/youtube/answer/2802032?hl=en>.

1. OpenNet Initiative. n.d. “About Filtering | OpenNet Initiative.” OpenNet Initiative. Accessed May 10, 2020a. https://opennet.net/about-filtering. [↑](#footnote-ref-1)
2. Ibid [↑](#footnote-ref-2)
3. United Nations Human Rights Council. 2011. “UN General Assembly Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, Frank La Rue.” May 16, 2011. [↑](#footnote-ref-3)
4. UN ONHCR. 1976. “OHCHR |  International Covenant on Civil and Political Rights.” 1976. [↑](#footnote-ref-4)
5. Bauer, Joseph P. 2010. “Copyright and the First Amendment:  Comrades, Combatants, or Uneasy Allies? .” *Washington and Lee Law Review*, no. Vol. 67, Issue 3: p831-914. [↑](#footnote-ref-5)
6. Bambauer, Derek E. 2009. “Cybersieves.” *Duke Law Journal*, no. Vol. 59, Num. 3 (December). 378. [↑](#footnote-ref-6)
7. Ibid 378 [↑](#footnote-ref-7)
8. Ibid 379 [↑](#footnote-ref-8)
9. Ibid 380 [↑](#footnote-ref-9)
10. Twitter. n.d. “Removal Requests.” Transparency Report. Accessed May 1, 2020. [↑](#footnote-ref-10)
11. Google. n.d. “Google Transparency Report.” Google Transparency Report. Accessed May 1, 2020. [↑](#footnote-ref-11)
12. Facebook. n.d. “Facebook Transparency Report | Content Restrictions.” Facebook Transparency Report. Accessed May 1, 2020. [↑](#footnote-ref-12)
13. Each country has its own top-level web domain, e.g. www.google.co.uk in the UK and www.google.fr in France. [↑](#footnote-ref-13)
14. Leo Kelion. “Google wins landmark right to be forgotten case”. 24 September 2019. BBC. [↑](#footnote-ref-14)
15. Heins, Marjorie. “The brave new world of social media censorship.” June 2014. Harvard Law Review. [↑](#footnote-ref-15)
16. Ibid 328 [↑](#footnote-ref-16)
17. Ibid 330 [↑](#footnote-ref-17)
18. Bogle, Ariel. 2019. “Laws Targeting Terror Videos on Facebook and YouTube ‘rushed’ and ‘Knee-Jerk’, Lawyers and Tech Industry Say - ABC News.” ABC (Australian Broadcasting Corporation). April 3, 2019. [↑](#footnote-ref-18)
19. YouTube. 2018. “Official YouTube Blog: Faster Removals and Tackling Comments — an Update on What We’re Doing to Enforce YouTube’s Community Guidelines.” Official YouTube Blog. December 2018. [↑](#footnote-ref-19)
20. YouTube Help. “Community Guidelines strike basics.” Accessed 30 September 2019. [↑](#footnote-ref-20)
21. Global Network Initiative. n.d. “About GNI | Global Network Initiative.” Global Network Initiative. Accessed May 16, 2020. [↑](#footnote-ref-21)
22. “The Santa Clara Principles on Transparency and Accountability in Content Moderation.” n.d. Santa Clara Principles. Accessed May 16, 2020. [↑](#footnote-ref-22)
23. Silverman, Craig and Jeremy Singer-Vine. “Here’s Who’s Been Blocked by Twitter’s Country-Specific Censorship Program”. 24 January 2018. Buzzfeed News. [↑](#footnote-ref-23)
24. Singer-Vine, Jeremy. “Country-"Withheld" Twitter Accounts — BuzzFeed News.” Accessed 27 September 2019. Buzzfeed News. Google Sheets. [↑](#footnote-ref-24)
25. E. Gallagher, P. Suarez-Serrato, and E.I. Velazquez Richards. “Socialbots Whitewashing Contested Elections; a Case Study from Honduras”. *Third International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing*. 2018; DOI: 10.1007/978-981-13-1165-9\_50 [↑](#footnote-ref-25)
26. Thomas, Kurt, Chris Grier, and Vern Paxson. "Adapting social spam infrastructure for political censorship." In *Presented as part of the 5th USENIX Workshop on Large-Scale Exploits and Emergent Threats*. 2012.

    [↑](#footnote-ref-26)
27. The International Press Institute. “Analysis by the Vienna-based International Press Institute on the Digital Manhunt of Turkish Journalist Nedim Turfent.” April 2018. [↑](#footnote-ref-27)
28. Ibid 13 [↑](#footnote-ref-28)
29. See “Data Description” section [↑](#footnote-ref-29)
30. Google. n.d. “Google Transparency Report.” Google Transparency Report. Accessed May 1, 2020. https://transparencyreport.google.com/government-removals. [↑](#footnote-ref-30)
31. Ibid [↑](#footnote-ref-31)
32. Twitter. n.d. “Removal Requests.” Transparency Report. Accessed May 1, 2020. https://transparency.twitter.com/en/removal-requests.html. [↑](#footnote-ref-32)
33. Facebook. n.d. “Facebook Transparency Report | Content Restrictions.” Facebook Transparency Report. Accessed May 1, 2020. https://transparency.facebook.com/content-restrictions. [↑](#footnote-ref-33)
34. Facebook. n.d. “Facebook Transparency Report | Internet Disruptions.” Facebook Transparency Report. Accessed May 1, 2020. [↑](#footnote-ref-34)
35. Global Network Initiative. n.d. “​New Report Reveals the Economic Costs of Internet Shutdowns | Global Network Initiative.” Global Network Initiative. Accessed May 15, 2020. [↑](#footnote-ref-35)
36. The World Bank. n.d. “World Development Indicators | DataBank.” DataBank | The World Bank. Accessed May 1, 2020. [↑](#footnote-ref-36)
37. The World Bank. n.d. “Worldwide Governance Indicators.” WGI. Accessed May 1, 2020. [↑](#footnote-ref-37)
38. Ibid [↑](#footnote-ref-38)
39. Ibid [↑](#footnote-ref-39)
40. Charron, Nicholas. n.d. “Assessing The Quality of the Quality of Government Data: A Sensitivity Test of the World Bank Government Indicators.” The Quality of Government Institute - The University of Gothenburg. Accessed May 15, 2020. p 23-24 [↑](#footnote-ref-40)
41. Reporters Without Borders. n.d. “2020 World Press Freedom Index | RSF.” RSF. Accessed May 1, 2020. [↑](#footnote-ref-41)
42. Schneider 2014. p 19 [↑](#footnote-ref-42)
43. Freedom House. n.d. “Freedom in the World | Freedom House.” Freedom House. Accessed May 1, 2020. [↑](#footnote-ref-43)
44. Freedom House. n.d. “Freedom on the Net Research Methodology | Freedom House.” Freedom House. Accessed May 15, 2020. [↑](#footnote-ref-44)
45. Schneider, Laura. 2014. “Media Freedom Indices: What They Tell Us – And What They Don’t.” BMZ Germany. January 2014. p 13 [↑](#footnote-ref-45)
46. Steiner, Nils D. 2014. “Testing for a Political Bias in Freedom House Democracy Scores: Are U.S. Friendly States Judged To Be More Democratic?” *SSRN Electronic Journal*, January 2014. [↑](#footnote-ref-46)
47. Facebook Transparency Report. “Case Studies” [↑](#footnote-ref-47)