Measurements for adoption, spread, and prediction of online human behavior

Thesis Proposal

Pamela Bilo Thomas (pthomas4@nd.edu)

March 19, 2020

1 Introduction

With the rapid creation of machine learning systems, scientists are better able to understand the large amounts of data that is being produced on a daily basis around the world. Effective models are needed to distinguish the signal from the noise of large real-world datasets as individuals attempt to make decisions from the data they are collecting.

This so-called "big data" provides researchers with a unique opportunity to study complex problems and human behavior online. However, as a result of social media and other systems, citizens are now more subjected to propaganda campaigns and other types of misinformation. We have seen that around the world, various state agencies and other organizations have used propaganda and misinformation to promote their own candidates or agendas [11]. Since social media and curated news feed have become a prominent component in the news diet of citizens [6], it is imperative to study how information spread through computer mediated social systems is affecting national beliefs and opinions.

Increasingly, as news sources become more fragmented and echo chambers become more prolific, it becomes easier for citizens to become siloed in their thinking and become more susceptible to misinformation [22]. Because of this, it becomes important to understand the process by which individuals are persuaded by what they see on social media - and how we can work to educate people to not be deceived by misinformation.

To study this process, we will summarize an overview of various machine learning algorithms that we used to understand several research questions in health care and social media. Taken together, this analysis gives an overview of several different approaches that were used to help understand these questions. From what we have learned in these models, we will create an additional model to help understand how to educate social media users to be responsible consumers of online information - which, ideally, will result in less spreading of misinformation.

The following chapters will summarize the work that we have done to study this problem. First, we summarize disease prediction algorithms that we have created. Working on these papers were instrumental in building my understanding of network science and began my foray into dealing with large datasets. Additionally, related work has been done to model misinformation spread as a virus in an epidemiological model [73]. Therefore, this work is an important first step on our quest to create new models to show information spread, even if the data comes from a different domain as my eventual proposal.

1.1 Library adoptions on GitHub

We show how people "adopt," or learn, new information by showing how Python libraries spread throughout GitHub projects. Our work summarizes how people on teams work together using new vocabularies (in the form of Python libraries) and eventually become productive after a period of growing pains as individuals compete over what libraries to use. Similarly, we can think of the spread of new Python libraries as like the spread of memes used for propaganda - the study of each problem is rooted in the learning, and adopting, of new terminology and ways of thinking.

1.2 Measuring gatekeeping and censorship during Internet outages in Venezuela

On social media platforms, a gatekeeper is an individual who consumes content which contains a variety of information, but tends to produce content aligning with only one side [29]. Other studies have called social media managers gatekeepers, since they determine what information their followers will see [69]. A user's capacity for functioning as a gatekeeper could be determined by their level of influence, with measures of influence including number of followers, page rank, number of retweets, and number of mentions. However, these measures do not tell a consistent story, as users who might rank highly in one measure of influence may do poorly in another [15] [39]. Gatekeepers with higher levels of trust will be more successful in spreading their messages [34].

Gatekeeping is therefore important as we attempt to understand how information is transmitted between groups of people. As misinformation spreads throughout a platform, some think that censorship is an appropriate response to make sure that individuals do not fall victim to believing lies [41]. Others censor information to maintain control over a situation. As entities resort to censorship, it is important to understand what happens when information is blocked. To that end, we researched what happened during an internet outage in Venezuela during the January 2019 protests. We tracked how information spreads during and after the censorship event, by analyzing how information bursts across languages and locations, and where people go to continue to discuss the censored topic. Additionally, we measured the effects of censorship on Reddit - and shows what happens when toxic communities were banned on the platform.

1.3 Growth and spread of memes

As we hope to understand how memes spread and information goes viral, we used Indonesia and the 2019 election cycle as a case study. Information that is highly popular and spreads quickly is more likely to recur in the future [20]. Additionally, information that is moderately appealing can recur in the future, as information that has already spread widely in the population will be more likely to not be reshared by those individuals who have already seen it [20]. Also, false rumors are more likely to recur in the future, in comparison to true rumors [58]. The social network of a user that is spreading the message is important to its volume, speed, and recurrence [32].

We found that memes can be a vector for propaganda and misinformation [40], so it is important to understand how memes grow and change over time. To do this, we downloaded a large dataset of images and tracked how the memes cluster and spread through apps such as WhatsApp and Instagram, which are important platforms in Indonesia. By detailed observation of these memes, again with the help of machine learning models, we showed how popular ideas evolve over time - and what memes are popular with the public and spread the furthest.

1.4 Reddit models for understanding human behavior

Reddit, or `reddit.com`, is a social network where users can submit content and others can vote on a submission by giving either upvotes or downvotes. When one visits `reddit.com`, submissions that have received more upvotes are listed higher than those that are less popular. Users can also subscribe to different communities, called subreddits, that track interests the users might have.

Many subreddits, which is defined as a community of people on Reddit, are concerned with user retention, and creating an environment where posters and commenters continue to come back to. According to Joyce, et. al, there are at least three common ways in which people will continue to post in a community (and not change their behavior): reinforcement, where a user has positive reactions with others, reciprocity, where a user does something for someone else, and the other will return the favor, and through personal bonds with members [36]. Others have found that stability, cohesiveness, and sociability are important to the future of online communities [43]. As communities grow, participation in group discussion tends to become more concentrated [51]. When it comes to understanding how communities retain members and grow, work has been done to analyze the participation costs (low barrier to entry) and consistency cues (behavior that occurred in the past will continue to happen in the future) [13]. People who join online communities at the same time will have similar experiences, which might happen as norms in communities change over time [7]. Strong connections to parent subreddits, defined as subreddits that spawn new subreddits, lead to stronger growth in communities [61]. When two communities share users, they tend to be excited about the same things, and interest intrinsically springs up in both groups [5]. People tend to leave communities for a variety of reasons, including lack of interesting people, low quality content, or harassment, among others [12]. People stay in communities because of shared interests, their experiences, supportive relationships, strong social feelings of belonging, and a shared identity [12].

1.4.1 Changepoint detection on Reddit

We will then take all that we have learned from this work and attempt to create several models that attempt to explain user behavior online. First, we propose a model that identifies users who have changed, or shifted, their behavior after some event that happened online. To do this, we use Bayesian Changepoint Analysis [4] to identify when users shift their attention permanently. Finding these changepoints - and discovering what communities people are moving to or away from - will be important as we show what community movement looks like when aggregated together.

Changepoint detection has been used in other areas, such as detecting changes in behavior in individuals who are at-risk for suicide attempts by analyzing their Twitter posts [67] and race car driving [70]. It has also been used to examine sociopolitical events based upon historical documents [19].

In addition to the changepoint model, we will also create other frameworks that paint a fully comprehensive picture of user behavior. One of these will be a model which attempts to capture change in user attention at the time of the event. This analysis is different from the previous one, since we are not necessarily identifying only users that change. In fact, user behavior has the potential to stay the same even after a dramatic event happens, such as a ban of the community, or other significant events in news or pop culture. Therefore, this model will attempt to predict where users go, and if their attention shifts after an event. We will also use this model to see if we can learn from past censorship events what will happen in the future.

After creating these models, we will show how censorship impacted users that are like those that were censored, but are not a part of the censored community. Finding control users are important as we show the impact that censorship has on user behavior. We also track what subreddits these users go to after a censorship event. Do they leave the platform, or go invade other subreddits? Do they make new subreddits where they continue their hateful speech? These are important questions, and as we track the frequent posters and compare them to other similar users who have not had their subreddits banned, we will fully understand what happens after a group of people is censored.

Finally, as governments and other political entities are trying to manipulate individuals online through the spread of memes and propaganda, we attempt to try to educate social media users not to be influenced by these campaigns, and to verify and validate what they see online before they share it. Through the results of this campaign launched during the 2019 presidential elections in Indonesia, we will report our analysis of the effect that our social media campaign had on the behaviors of people who viewed the site.

Through these models, we show one perspective of how individuals behave online, and how the influence of censorship or information campaigns can impact user behavior. This multi-pronged approach of using various models and techniques will present a comprehensive study of online human behavior.

Thesis Statement

I propose to create several prototypes that track how people behave on the Internet by modeling their behavior on several platforms. By using various data sources, which consist of Reddit data, Indonesian memes, survey data, and Wikipedia data, we will create several models which can be used to simulate how offline events manifest themselves online, and how these events and other factors, may cause Internet users to change their posting and sharing behavior over time.

Proposal Overview

In the following seconds, we give an overview of work done so far - and how we propose to answer questions raised in the Introduction.

Section 2 will give an overview of current work in various domains, including disease prediction algorithms, GitHub adoptions, censorship events, and meme clustering.

Section 3 will propose future work that will be conducted to further validate (or invalidate) the thesis statement. This future works consists of a changepoint model, an attention model, and an analysis of Indonesian memes and survey data.

Section 4 will summarize the plan, and give a timeline for its completion.

2 Previous Work

Work unrelated to the thesis We had previously authored two studies [46, 64] that apply network science approaches to medical and health science problems. These papers tackle two separate chronic health conditions and use graphs to help predict the onset of disease and other complications. The first paper summarizes disease diagnoses for 1.1 million elderly patients [46], and the second analyzes over 171,000 patients for various conditions, which were drawn from a potential candidate pool of over 805,867 individuals generating over 500 million medical records [64].

By constructing a DAG (directed acyclic graph) of Medicare patients and their visits [46], we found trends in diseases that result in an ultimate diagnosis of heart failure. We conclude that cardiomyopathy is a condition that is commonly associated with heart failure such that screening for cardiomyopathy should be a common part of preventative treatment. Additionally, we know that many patients' first diagnoses on a heart failure path are acute myocardial infarctions, endocardium diseases, and cardiomyopathy. Doctors who see patients for other medical issues, especially mental issues as observed, should know of these complications since they are often the first that show up in diagnoses that do not otherwise lead to heart failure. We also found that rheumatic heart disease, pulmonary congestion and hypostasis, cardiomyopathy, blood poisoning, and valve and aortic diseases are common comorbidities that occur before doctors diagnose patients with heart failure. Because the highest information gains in our DAG are on paths that concern mental disorders such as psychosis, cerebral degeneration, and Parkinson's, the conclusion can be made that patients being seen for these disorders should also be monitored for heart disease. In addition to our heart disease prediction algorithms, we also show that given a patient's disease history and lab results, we can predict their likelihood of developing complications from diabetes [64]. We also show what disease diagnoses or lab results (from our heterogeneous network or graph) are most likely to lead to specific diabetic complications. The diagnoses graphs can help illuminate health problems faced by many patients and what might be the best course of disease management. Not managing complications, especially for fast progressors, can cause rapid development of uncontrolled diabetes, from which it is hard to recover. Moreover, disease diagnoses graphs can also be a useful tool for physicians to understand the effects of co-morbid conditions, and personalize a wellness and disease management plan. This can lead to an improvement in both individual and population health outcomes.

2.1 Dynamics of Team Library Adoptions: An Exploration of GitHub Commit Logs

A manuscript has been published describing the work presented in this section:

Pamela Bilo Thomas, Rachel Krohn and Tim Weninger. Dynamics of Team Library Adoptions: An Exploration of GitHub Commit Logs. *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM), September 2019.

When a group of people strives to understand new information, struggle ensues as teammates argue about implementing various ideas and steep learning curves are surmounted as teams learn together. To understand how these team dynamics play out in software development, we explored Git logs, which provide a complete history of software repositories. In these repositories, we observe code additions, which represent successfully implemented ideas, and code deletions, which represent ideas that have failed or been superseded by others. By examining the patterns between these commit types, we can begin to understand how teams adopt new information. We specifically study what happens after a software library is adopted by a project, or, when a library is used for the first time in the project. We found that a variety of factors, including team size, library popularity, and prevalence on Stack Overflow are associated with how quickly teams learn, implement, and successfully adopt new software libraries.

This work introduces a new approach to study this problem. By investigating how software developers adopt and use software libraries in this specific context, we can better understand how humans learn new technical information and incorporate concepts previously unknown to the user or group. The findings from this study can be generalized to understand how humans work, learn, and fight together, since GitHub provides a rich dataset which approximates the collaborative process.

Additionally, we asked if there exists any competition among team members over the inclusion of a library. When teammates have disagreements about what should be included in a GitHub project, there will ultimately be a winner. Uncovering which user eventually wins these code fights is an interesting research question. We explored competition by examining *code fights* where two users revert each other's code over the course of several commits. Like edit-wars on Wikipedia [66], by looking at the users and libraries that participate in code fights we can learn a great deal about the adoption and diffusion of information. Although the present work focuses specifically on code contributed to public Python projects hosted on GitHub, we hope that other domains can use the methodology of the present work in other explorations of information adoption generally.

To summarize, this work aims to answer the following research questions:

- What are the events that happen when a team adopts a library for the first time?
- Are commits containing new libraries more likely to have deletions than other types of commits?
- Do the answers to these questions vary by library type, team size, or the amount of information available on Stack Overflow?
- Do team members fight over the adoption and usage of a new library?

We found that when teams attempt to learn new information together, it can be challenging to apply these new concepts and there is often a learning curve needed before the new information can fully stabilize within the group project. We further found that that even though learning curves are unavoidable, it helps to have teammates and other online resources that can guide groups towards learning how to adopt new information. When conflict arises, the more experienced team members usually end up winning disagreements when we track whose code ends up in the final version. Through this work, we confirmed that learning new information together can be a difficult process, and that many of the statistics surrounding GitHub projects follow power law distributions (including team size, commits, and adoptions). This work provides a superficial glimpse into an analysis of teamwork on GitHub, though there are still more in-depth questions to be answered to uncover more information about more complex interactions.

2.2 Measuring the effect of Internet outages on social interactions across location and language communities on Twitter

A manuscript describing the work presented in this section is in final stages of preparation:

Pamela Bilo Thomas, Emily Saldanha, and Svitlana Volkova

Many authoritarian regimes have taken to censoring Internet access in order to stop the spread of misinformation, to prevent citizens from discussing certain topics, to prevent mobilization, or other reasons. There are several theories about the effectiveness of censorship. Some suggest that censorship will effectively limit the flow of information, whereas others predict that a backlash will form, resulting in ultimately more discussion about the topic. In this work, we focus on Internet censorship in Venezuela that occurred in January 2019. To gain an understanding of what happened, we study the spread of discussion topics across locations, languages, and time on Twitter. We describe how these external event outages affect social structure before, during, and after an internet censorship event. We also use user centrality to detect key influencers and gatekeepers in the social network by detecting which topics are bursting across different groups and identify sets of individuals that are influencing this spread. Through this analysis, we explore the effect that censorship has on online discussion, how people can spread topics across language and location barriers, and the differences between how information spreads within and outside of a censored environment.

With many regimes around the world resorting to censorship, including China, Indonesia, Iran, Pakistan, Thailand, and others [72], it is important to understand how communication patterns change during a censorship event. In 2017, 38% of Internet users lived in countries where social media or messaging has been blocked in the previous year [24]. Countries block access to Internet sites for political, social, or threats to national security reasons, among others [23]. During a censorship event, it is important to understand how new social groups form, how discussion topics evolve, and which users create and spread viral information. To this end, we examine a collection of tweets from Venezuelan censorship events in January 2019 to address these questions. We focus on analyzing communication patterns across languages and locations, since these can act as natural barriers to information diffusion.

As of 2017, over 12 million Venezuelans use social media, and 3 million of them are active Twitter users [10]. Various groups have attempted to use Twitter to spread their messages in Venezuela. For instance, in 2015, the Venezuelan government attempted to spread pro-government information on Twitter [10]. Most Venezuelans access the internet over one state-run internet provider, CANTV [53]. In 2018, 52 websites were blocked by Venezuelan internet providers [53]. In this work, we focus on the recent outage events occurring in Venezuela on January 21, 2019 [49], January 23, 2019 [48], and January 26-27, 2019 [47] during ongoing protests in the country. The first two outage events primarily included disruptions to social media access, while the third was a total internet outage.

With these events as background, we aim to answer the following research questions:

- How do external outage events affect social structure? What does the social structure look like before, during, and after an internet censorship event?
- How does discussion of prominent entities, such as political figures and social media companies, recur over time during repeated events?
- Which users serve in central roles connecting different groups and communities?
- Which users function as gatekeepers for specific communities, determining whether specific information will spread to users in that community?

Through this work, we have shown the impact of languages and locations on the discussion of three censorship events occurring within a short time frame in Venezuela. We have shown what happens when a censorship event occurs, how information bursts within a social network, the impact of languages and locations on centrality, and a new way to detect gatekeepers. As more countries face censorship from their governments, it becomes crucial to understand how information continues to spread during an event even in the face of potential information spread barriers such as language and location differences.

2.3 An Automated Pipeline for Identifying and Tracking Political Meme Genres with Diverse Appearance

A manuscript describing the work presented in this section has been submitted and is under peer review:

William Theisen, Joel Brogan, Pamela Bilo Thomas, Daniel Moreira, Pascal Phoa, Tim Weninger, Walter Scheirer. Automatic Discovery of Political Meme Genres with Diverse Appearances. *Under peer review*

From silly cat photos to parodies of political candidates, image-based memes represent a vibrant and vital form of communication on social media. While once primarily the domain of the online communities found on 4chan and Reddit, most casual users of social media are now familiar with the concept of a meme and spread such content freely. Moreover, advances in image-editing software have made sophisticated tools accessible to untrained users, which has rapidly increased the production of memes by amateurs and professionals alike. When it comes to politics, the messages contained within memes span a landscape from beneficial "get out the vote" campaigns [68] to anti-social messages stoking violence [8, 55]. In order to monitor social media for emerging trends in memes, researchers have proposed new ways to use computer vision to automatically identify and track specific genres [27, 74].

To study this problem, we performed an expansive study of the 2019 Indonesian election [33], where a diverse set of memes and other image content circulated on social media platforms. During the year both preceding and covering the election, we collected a data set of over two million related images from social media, including 170,000 images from 14 different sources on Twitter, and 1.9 million images from 20 different sources on Instagram. The images were taken from the time period covering May 31, 2018 to May 31, 2019. Our primary goal was to cluster the images in order to study the trends that emerged in the social media space surrounding the election. Through these investigates we found which memes were reaching the Indonesian population during the election and attempted to categorize them as well as understand what messages were being construed.

Indonesia was chosen as a case study for a variety of reasons, including its large number of Internet users, diverse population, and democratic culture. In July of 2016, it was estimated that Indonesia had the 9th most Internet users in the world [14]. As a large, middle-income democracy [62], Indonesia provided a unique opportunity to study the spread of political memes during the 2019 election in a population that is relatively new to the Internet. Results gathered from this case study can then be applied to model what might happen in other middle-income democracies with recent Internet participation. Through tracking opinion polls and election results, we could understand the views that the Indonesian people had about their government, and who would ultimately win the election, as a result of or in spite of the memes that were spread beforehand and the information the public was exposed to. The findings presented in this paper represent the first large-scale visual analysis of political meme content associated with a major world event.

As one of the largest collections of political memes ever gathered, this rich and complex dataset has provided us an extensive look at the clustering of memes. We have gathered millions of Indonesian memes, of which some, like the colored-pinky meme, are unique to countries like Indonesia, where individuals are physically identified after voting. We can assume that spreading memes such as this are a result of civic pride, which provide an uplifting counterpart to the misinformation that can be spread on social media by malicious users. We also see that political and other advertisements are incredibly prolific in this dataset, and that going forward, political memes will undoubtedly have an impact on how people in a democracy discuss and learn about the candidates for which they will be voting.

Previous work from online sources such as GitHub [63] show that groups of people working together can 'adopt' libraries which are eventually incorporated into a group's vocabulary. Similarly, we can think of groups of individuals on social media platforms such as Twitter or Instagram that 'adopt' hashtags or phrases and insert them into memes. Together, these groups of people are working to create a shared online language through hashtags and memes. As a relatively new form of discourse, this image and meme analysis requires further study as researchers uncover how language and images spread across people and culture. Further work can attempt to understand how long it takes for a population to become 'fluent' in a meme and understand its meaning and how to use that particular meme to communicate.

In this work, we show that there is a large quantity of political memes when we sample social media sites such as Twitter and Instagram. We show that at least 1 out of every 3 clusters of memes gathered was fully political in nature. Therefore, we can expect that political meme sharing is having an influence on the segment of the Indonesian population that uses social media. In the future, it is hard to imagine that the role of political memes and hashtags will diminish their importance in our political discourse. Rather, it is more likely that these memes and hashtags are here to stay. While propaganda has always been a part of politics, memes present a relatively new medium for which individuals to express themselves. Additionally, while propaganda and misinformation might have previously been only limited to wealthy people, corporations, or political parties that owned expensive media such as television stations, newspapers, or radio stations, the proliferation of social media and the low cost to entry makes it easy for memes and misinformation to be created cheaply and spread easily. Further study is needed to understand the role of memes in such places such as voter turnout and voter persuasion, and to comprehend how people are influenced by these memes, and potential ways to educate voters such that they can resist misinformation and propaganda.

3 Proposal for future work

Advisory: some of the examples in this section contain explicitly hateful and troubling content.

3.1 Modelling change in user behavior on Reddit

Analyzing how individuals move through online communities and change between them is an interesting problem and important to understand as we try to make sense of how humans adopt and join new groups. Some who have investigated this problem have found that people on Reddit can shift their behavior 'suddenly' between different topics [65]. The internet is a place that allows diverse groups to post their beliefs for anyone to read and consume. As time goes on, people may find themselves changing their beliefs, or joining and participating in different communities. In this work, we propose to present a high-level analysis of what occurs when people change behaviors on the internet on Reddit. We will uncover when changepoints occur in a user's behavior, which marks when an individual stops or starts posting in one subreddit, or begins to post in another. We will use a variety of subreddits to show that it is possible to model change in user behavior and correlate these changepoints to external events.

On Reddit, users can subscribe to different communities, called subreddits, that track interests the users might have. If a user subscribes to a subreddit, they will see posts from that subreddit in their own personalized front page. For users that do not have an account, the front page was populated by "default" subreddits until 2017, when the default subreddit system was replaced by a new algorithm that ranks posts from all popular subreddits [25]. A user does not necessarily have to be subscribed to a subreddit to comment, but subscribing to a particular subreddit means that the user will see popular content from that subreddit when they visit their homepage.

By tracking user engagement across various subreddits, this work will answer the following research questions:

- Can we track changepoints over time and correlate clusters of changepoints with external events?
- Can we model the flow of users towards or away a certain subreddit?
- Can we estimate the likelihood of a Reddit user experiencing a change in behavior?
- When an exogenous event occurs, do users change their behavior? If so, how?

3.1.1 Experiment

To answer these questions, we will gather posting behavior from all people who posted in the following subreddits in the year before and after the events listed (therefore collecting two years of user behavior, unless otherwise noted). An important note is that Reddit is always changing and subreddits can be banned or undergo other changes, which can provide natural experiments for us to study. Each of these events listed below provides an example of one of those natural experiments as we attempt to understand how user behavior changes when an external event happens. These subreddits give us a variety of topics (including sports, politics, hate speech, fandoms, and others) to help us understand how people who have these interests change their behavior over time. Collecting these data over a large period will allow us to understand how the user is behaving prior to the event - and what changes might occur afterward:

- fatpeoplehate - Banned June 10, 2015 due to repeated hate speech violations [26].
- Incels - Banned November 7, 2017 due to repeated hate speech violations [35].
- DarkNetMarkets - Banned March 21, 2018 due to selling and buying prohibited goods [52].
  - The Bitcoin peak happened December 17, 2017 and affected the DarkNetMarkets platform
- ThanosDidNothingWrong - 'Snap' occurred July 9, 2018 which 'banned' half its users to the subreddit InTheSoulStone analogous to fictional events from the theatrical movie series [3].
- StarWars - New movies in the Star Wars series include 2015's The Force Awakens, 2016's Rogue One, and 2017's The Last Jedi [71]. Data collected over a period of five years to track change in behavior when various films were released from December 14, 2014 - December 9, 2018.
- SandersForPresident - The political subreddit for Bernie Sanders supporters [44], we track the behavior of individuals on this subreddit in the year prior to and after the 2016 election on November 8.
  - SandersForPresident was briefly removed from Reddit on July 26, 2016 for toxic behavior [54] until it was restarted after the 2016 election in November.
- the_donald - The political subreddit for Donald Trump supporters [44], we track the behavior of individuals on this subreddit in the year prior to and after the 2016 election on November 8.
- Atheism - removed from default subreddits July 21, 2013 [1].
- ExplainLikeImFive - added to default subreddits July 21, 2013 [1].
- Bitcoin - The Reddit bitcoin subreddit was used as a place to discuss and promote cryptocurrency [30]. After the Bitcoin bubble burst in late December 2017 [59], it would be interesting to see if the people promoting Bitcoin on Reddit continued to post.
- LosAngelesRams/StLouisRams - The NFL team Rams left the city of St. Louis after the 2015 NFL season and moved to Los Angeles. We gather data for all users who posted in both subreddits 1 year before and 1 year after September 1, 2016, to capture user behavior. On Reddit, a new subreddit was created for the new team - the St. Louis Rams subreddit still exists as of January 6, 2020.
- Chargers - Like the Rams, the Chargers left San Diego to head to Los Angeles after the 2016 NFL season. Unlike the Rams, the Chargers kept their subreddit r/Chargers. Like our analysis for the Rams, we track behavior for users on the Chargers subreddit for 1 year before and 1 year after September 1, 2017.

To detect change in user behavior, we will use Bayesian Changepoint Analysis [4], which is a technique used to identify change in the combination of multiple Boolean signals.

3.1.2 Analysis

We plan to collect user behavior from all the users that participate in these communities during the given time frame. Once we gather user behavior, we will then check their posting history to see if their behavior "changed" during the time period. With these changepoints, we will answer these following research questions:

Is it possible to model changes in user behavior? First, we will attempt to see if Bayesian changepoint analysis works as a valid tool for measuring change in user behavior. We expect to see that very few users have a change in their behavior.

Can we track changepoints over time and correlate clusters of changepoints with external events? Our goal is to understand if we can see the effects of external events via change in social media behavior. We want to see if we can detect changes on Reddit policy (for example, change in Reddit default subreddits, banning of certain subreddits) reflected in when our users experience changes. We will also perform a sensitivity analysis - as we expect to see that users who are more active in a subreddit that has been affected by an external event are more likely to experience a change in their behavior.

Can we model the flow of users towards or away from a certain subreddit? Once we identify changepoints, we will then work to explore what is happening to the user's behavior. Once a changepoint is identified, it means that the user is either moving to, or away from, a group of subreddits. As we aggregate the changepoints together, we hope to see patterns emerge and uncover how groups of people change together. We plan to look at these patterns of movement from both a time period perspective (when are people joining or leaving a subreddit at the same time) and from a progression perspective (in what order do people who were active in a particular subreddit move between other subreddits).

Can we estimate the likelihood of a Reddit user experiencing a change in behavior? Ultimately, we want to find out how many users experience changepoints, and what the probability that a user changes their posting behavior as a result of these events. After gathering this information, we can use what we learned to predict what will happen to users during future online events - and ultimately where people will move to and away from. Knowing this information will be helpful for companies that manage social media platforms as they work to better understand effects of their policies (such as banning subreddits or adding or removing subreddits from default lists).

3.2 Models of User Behavior after Community Bans on Reddit and Other External Events

For the most part, Reddit has engaged in a laissez-faire approach to regulating speech on its platform. Each subreddit is regulated by moderators, who control the content that is submitted, and the comments on those posts [9]. Every subreddit has its own set of moderators who have a different set of rules which determine if content is allowed on a subreddit, or if the content should be taken down [9]. Various communities have different norms and rules that their moderators follow [18]. In some communities, hate speech has flourished, even in some subreddits not dedicated to hateful ideas, such as in college subreddits [56]. Reddit has been known for misogynistic content, and it has been shown that female Redditors participate in commenting and submitting posts less than male Redditors do [37]. Additionally, this lack of regulation has allowed for toxic and offensive content to spread through many subreddits [42]. Toxicity has a negative correlation with subreddit growth [45]. Following years of controversy about their approach to regulating speech on their platform, in May 2015, Reddit redefined their approach to harassment to be: "Systematic and/or continued actions to torment or demean someone in a way that would make a reasonable person (1) conclude that reddit is not a safe platform to express their ideas or participate in the conversation, or (2) fear for their safety or the safety of those around them." [1]. This led to the banning of subreddits including 'fatpeoplehate,' [17], which had over 151,000 subscribers before its ban [2]. There was a small user migration from Reddit after the ban, but Reddit's ability to support a large platform offering a variety of content helped it weather the controversy [50].

Since this event, Reddit has gone on to ban other troublesome subreddits from its platform. The goal of this work is to study the effect the ban had on disrupting the r/fatpeoplehate community, along with others, such as r/incels (banned on November 7, 2017) [35] and r/DarkNetMarkets (banned March 21, 2018) [52]. Some found that the banning of the fatpeoplehate subreddit was successful and had a positive impact on the community [57]. We want to fully understand the migration to other subreddits after the ban took hold, and to quantify the effect that removing the subreddit had on disrupting that community. Using each banned community as a model for the others, we want to investigate how transferable these banning events are, and if we can predict what will happen in the future after banning events.

In addition to Reddit bans, we will also attempt to model other behavior patterns. We will analyze how Redditors discuss movie and sports fandoms online by tracking users in the Star Wars and NFL team subreddits. Additionally, we will also track how Reddit was used to spread, propagate, and discuss political ideas, by analyzing two popular subreddits during the 2016 election, SandersForPresident and the_donald. Collectively, this will give us an overview of how Internet users discuss sports, politics, movies, and pop culture through events that can be tracked in real time to understand how Reddit is used in various environments - and when communities turn toxic, what happens when Reddit decides to take action to rid their platform of unsafe behavior.

After a banning event occurs, it is possible for a Redditor who was previously active in that banned community to take one or more of several actions: leave Reddit altogether, post in a newly created community that might exist to take the place of the old space that had been banned, invade a new community (for trolling or other purposes), or continue their same posting behavior. After each banning event, we hope to quantify each of these actions.

Therefore, this work aims to answer the following research questions:

- Can we use various events in pop culture and politics to model how users behave on social media as a response to external events?
- What proportion of Redditors who are active in a toxic subreddit remain on the platform after a banning event?
- Can we model where Redditors migrated to after a subreddit they are active in becomes banned?
- Can we use past censorship events to predict future migration patterns?

3.2.1 Experiment

Similar to the work that we proposed in the changepoint analysis subsection, we will gather user behavior for individuals who posted on those subreddits. We will use the same subreddits to conduct our experiments.

Predicting future behavior Using past events, we will work to construct a model to measure the effect that the event has had on current user behavior. For every individual i who is a member of the community in which the event occurred (for this we say that an individual if they post at least once in a given subreddit), we observe the subreddits s_0 to s_n that the user has posted in, in a given time t before and after the event occurred. Then, we will construct the amount of attention that has been given to that subreddit, by counting the total p_0, p_1, c_0, and c_1 posts and comments that were made before and after the event across all of Reddit. A vector is then constructed which counts the times a user participated by posting or commenting in each subreddit, and then dividing by the total number of comments and posts in this time frame. Each vector will sum to 1, since we are creating a vector which represents the fraction of attention that a user has given to each subreddit, combined to the user's attention. We say that n subreddits are needed to capture p percent of total user interactions (posts, comments, etc.) on Reddit for all the users who are active members of the observed subreddit. It is important that the same set of subreddits is used for analyzing all users, so we find this set of n subreddits across all users.

After we observe this user's behavior before and after the point in time, we will use this vector to create a model which will learn how user behavior changes after these events. For each event and subreddit, we set an activity threshold for users, and various time intervals to predict subreddit activity before and after the event. We propose to use a simple feed-forward neural network, as we have seen that simple models are oftentimes as good as complicated ones but are much more interpretable [60]. Additionally, we will experiment using a variety of input sizes and dimensions to find the best parameters for our model. We will run our model using the Keras neural network package [21], with an Adagrad optimizer [28] and Kullback-Leibler divergence loss function [38].

Gathering aggregate posting data While our previous subsection is focused on modelling change in behavior on an individual level, we are also focused on modelling change in the aggregate. For each of the subreddits listed above, we work to implement the model proposed in [17] to find control users - who are users who post in subreddits that are similar to the banned or example subreddit, but never in the actual subreddit. This methodology is important to help us understand how people that are like those that are actually censored react to a banning event. We find that while many of the users who are censored decrease their activity after an event, we do not see the activity levels exactly mirrored in those users who are identified as matched controls. Rather, we see that the effect of having a community that a user is invested in banned results in decreased behavior for that user across all of Reddit, but does not necessarily decrease behavior for those who are posting in subreddits that are similar to the one that was censored.

To analyze this aggregate behavior change, we will first gather all individuals who have posted in the analyzed subreddit at least 10 times in the year prior to the event. We will then observe how they post before and after an event - if they are posting in a new subreddit (created after the event - this would indicate that the upset members of the banned community are creating new community spaces, potentially to share their frustrations or continue their behavior), invading other subreddits (meaning that the users who have been censored are going into other spaces that they were not currently posting in to express their ideas), or keeping their behavior constant in other subreddits. We will also track absolute number of posts on Reddit - which is useful as we attempt to understand if overall activity levels on Reddit are going down as a result of being censored, and potentially leaving Reddit for another platform.

Tracking this behavior is useful as we attempt to understand how a censored community acts, but it is as important to also track how other individuals, who are like the censored ones, act during a censorship event. To find these controls, we will again follow the methodology of [17] and do the following steps:

- Find all users who post in to studied subreddit s at least 10 times in the year before the event.
- Find all other subreddits these users post to in that year.
- From that list of subreddits, create a list of subreddits that have at least 10 posts by at least 10 authors in the year before the event, to find other subreddits these users are active in.
- Find how many users on all of Reddit post to those subreddits in that period, and then find the top 200 subreddits that have the greatest percentage of users from the subreddit s.
- For those top 200 subreddits, find all users who have posted in any of those subreddits at least 10 times - but have never posted in s, to create a set of control users.
- Using this set of control users, for each user u who posts in s, find a “matching” control from the set of control users that we have identified by using the Mahalanobis Distance Matching algorithm [16] on the dimensions of account age, karma, and number of posts.

3.2.2 Analysis

Through this experiment we return to the following research questions:

Can we use various events in pop culture and politics to model how users behave on social media as a response to external events? We expect that the experiments listed above will give us a comprehensive overview of what happens as a result of these events. We hope to show that our neural network model will give accurate predictions about how a user’s attention will change after the event. We also will attempt to study how we can use past events to predict what will happen when similar incidents occur in the future.

What proportion of Redditors who are active in a toxic subreddit remains on the platform after a banning event? Ultimately, we want to discover through our analysis if censorship works to remove hate speech from a platform. Theoretically, Reddit works to ban communities, not individuals, on a large scale from their platform. Even if Reddit bans an account, it would be easy for that user to simply create a new one. Additionally, if Reddit bans a community, it would be easy to create a new space where the toxic speech is continued. Therefore, if we track the number of people who end up leaving the platform altogether after a community is censored, we can begin to understand how well banning subreddit works - if the goal is to remove harassers or hateful individuals from Reddit.

Can we model where Redditors migrated to after a subreddit they are active in becomes banned? After a subreddit is banned, the users who were previously active in that community have several choices about where to go afterwards: they can form new spaces, invade other subreddits, leave Reddit, or continue their same behavior. After we collect this data, we will model where users go after the event, and summarize what proportion of users, grouped by their activity level in the subreddits, end up doing after their subreddit is banned.

Can we use past censorship events to predict future migration patterns? Ultimately, we want to discover if what we learn about Reddit bans is transferable to other incidents. From our previous research question, we summarize what Redditors do after an event. After this summarization, we look to other instances of censorship and compare the proportion of active community members who engage in various behaviors, such as leaving the platform or invading other subreddits. After this comparison, we see if the numbers after each event are similar. If so, we can predict approximately how many people in the community will leave Reddit, create new communities, etc., after a community is banned. We can therefore use what we learned about these censorship events to predict what will happen in the future if more communities are banned.

3.3 Investigating the social media landscape in Indonesia

The threat posed through online and social media vehicles is particularly compelling in that ordinary citizens will both consume and spread (mis)information through their online activity, affecting thousands of individuals instantly with falsehoods that are then implicitly endorsed by a seemingly qualified source. This issue is especially problematic for new digital arrivals who are least likely to understand the dynamics of these complex social and technical systems. One of the primary consequences of the online and social media environment is that a handful of motivated, malicious individuals can disrupt the information landscape. Tactics vary across regions due to social and design differences in popular online and social media systems. Due the relative ease of social media manipulation and their large impact on behavior and belief, it is widely expected that malicious individuals and groups will continue to spread harmful and false information. This is especially the case in the lead up to national democratic elections.

At a time when communities around the world increasingly turn to digital sources for information, online and social media systems play a critical role in affecting attitudes and behavior. A core problem is that social media channels are being manipulated by malicious groups to spread misinformation in low- and middle-income countries (LMIC) to exacerbate social divides and influence citizen involvement in democratic processes. The spread and adoption of misinformation through digital channels is especially problematic because many users of online and social media systems are not aware of how (mis)information is spread through these channels.

Because of the critical need to address the spread of misinformation, especially among new digital arrivals in LMICs, we propose to research, develop, deploy, and measure the effectiveness of an online and social media literacy campaign through a targeted pilot study.

Therefore, we plan to answer the following research questions:

- In Indonesia, which demographics of people are more likely to believe false, or true, information?
- Can we teach people healthy online behaviors, and what methods work better than others?
- How do people change their beliefs about news stories they see online over time?
- Can we predict when memes or stories will go viral in Indonesia?

3.3.1 Experiment

Our collaborators, IREX and Moonshot, are currently deploying a set of videos on a website that we have created to teach Indonesian citizens to think before they spread messages on social media at `literasimediasocial.id`. This website contains short videos and text which encourage individuals to engage in healthy online habits. Additionally, we are currently conducting a survey before and after the 2019 election which aims to understand how the Indonesian population is using social media and spreading misinformation.

Memes We have gathered over two million memes that have been spread on social media sites Instagram and Twitter, as described in Chapter 2.3. Through the gathering of these memes, we will track which types of memes are most popular in Indonesia. Further, we have downloaded the entire change history of Indonesian Wikipedia to fit a Hawkes process [31] to model when information goes viral in Indonesia. With this information, we will paint a picture of what type of information is popular in Indonesia, and what spreads the furthest the most quickly.

Survey The survey, administered via phone, is given in an attempt to understand baseline beliefs about several true, misleading, and false news stories in Indonesia. Since it is given twice, the survey will help us understand how feelings about these topics change over time and in response to the placement of media literacy content. The survey we administered is given in Appendix A.

The Tracker Our collaborators have launched software (called The Tracker) which is capable of tracking user behavior across a variety of platforms, including Google, Bing, Yahoo!, Twitter, Facebook, Telegram, Instagram, YouTube, Discord and Google Plus. This includes exclusive tools to track individual searches on Google, infiltrate closed Telegram Groups, and access millions of videos which have already been removed from YouTube. Where possible, data is disaggregated by demographics, while preserving the anonymity of users. This database will guide The Tracker to provide highly granular mapping of audiences engaging with these indicators on select digital platforms. Where possible, this will include analyses of demographics of audiences that engage with disinformation online; how they encounter such material; the extent to which they repeatedly engage with disinformation; and what other online content they consume.

The Tracker has gathered information from YouTube videos, Facebook pages, and Twitter. Further, it has directed users who it deems especially susceptible to misinformation to our website which attempts to teach individuals to engage in healthy online behaviors. Through the deployment of the Tracker software, we will track what people are doing online - and what the effect of our media literacy campaign is.

3.3.2 Analysis

In Indonesia, which demographics of people are more likely to believe false, or true, information? The survey results will give us an understanding of which news stories are popular with certain groups of people. Because the results are aggregated by demographics, we will fully comprehend which segments of the population is more likely to believe misinformation and false stories which have been promoted by certain groups in Indonesian society. This will further help us in our targeting approach as we attempt to deploy our media literacy campaign to the individuals that are most susceptible to misinformation.

Can we teach people healthy online behaviors, and what methods work better than others? We have deployed and advertised a web site which uses short videos and slogans to encourage responsible social media behavior. After the data collection period is finished, we will examine the results to see if there is a statistically meaningful difference from before and after the treatment, in the form of a website visit, was applied. We will do this by analyzing pieces of information such as Google searches, Facebook posts, and other indications of online behavior to measure if the treatment has had any measurable effect on how people interact.

How do people change their beliefs about news stories they see online over time? We will be administering the survey twice. When we compare the data from the two times the story has been administered, we hope to answer this question by comparing percentages of people who think the story is true or accurate. We hope that more people believe in the true story and disbelieve the false one. This will give us some ideas about the pervasiveness of information over time: is it more likely that people will have time to research true stories and believe them, or do the lies manage to spread further when they have been circulating for longer?

Can we predict when memes or stories will go viral in Indonesia? We have several tools at our disposal to do this. First, we have the collection of memes that we have gathered from Instagram and Twitter that will help us understand what types of memes are popular in the country. Secondly, we have the downloaded corpus of Indonesian Wikipedia edits which will allow us to track when ideas or subjects are popular in the country, since we can see when spikes in activity occur.

3.4 Evaluation Plan

As we analyze Reddit, we plan on continuously updating and refining the models as needed. While the proposed Bayesian changepoint and subreddit flow models are not predictive and just consist of straightforward analyses, the neural network attention model will need to be tested thoroughly to ensure that its predictions are accurate. We plan on using 80/20 5-fold cross validation to ensure that the model is adequately trained and outputs valid testing results.

Our interview given to Indonesian citizens, along with the data about their internet usage, is covered in IRB Protocol ID 18-11-5009. Data is anonymized, but demographic information is included.

Data that we collect about Wikipedia and Reddit usage is anonymous, though there are screen names or IP addresses attached to their usernames. Further, we do not have user names attached to the Indonesian memes we have gathered from Twitter and Instagram.

4 Summary

In this proposal, we have described past work that we have done to understand Internet behavior from many angles: GitHub library adoptions, Internet censorship, and meme creation and spread. As we attempt to fully comprehend how individuals change their behavior online, and how information about politics and other events go viral, we will follow the experiments laid out in Section 3 to create several models which will show the impact of offline events on online behavior, and quantify change in user behavior over time. Our Reddit data will provide us with a large dataset for complex analysis to answer these questions, and the Indonesian set will allow us to test our hypothesis about the ability for user behavior to change via education. Together, we can model how change in online behavior occurs - and if we can act to influence people to practice healthy online habits.

4.1 Timeline

Activity	Month
Finalize data collection for problems	Now - April 2020
Create neural network model	May - June 2020
Create changepoint model	June - July 2020
Create community movement model	August - September 2020
Finalize Indonesian data analysis	October 2020
Write Indonesian social media analysis paper	November 2020
Write community movement model paper	December 2021
Write changepoint paper	January 2021
Write neural network paper	February 2021
Finish up additional tasks and analysis	March - April 2021
Defend	May 2021

References Cited

[1] Promote ideas, protect people. https://redditblog.com/2015/05/14/promote-ideas-protect-people/. Accessed: 2019-06-23.

[2] Reddit bans 'fat people hate' and other subreddits under new harassment rules. https://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit. Accessed: 2019-06-23.

[3] Thanos subreddit successfully bans half its community. https://www.theverge.com/2018/7/10/17548768/thanos-subreddit-bans-half-community-marvel. Accessed: 2019-12-23.

[4] R. P. Adams and D. J. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007.

[5] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54. ACM, 2006.

[6] E. Bakshy, S. Messing, and L. A. Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[7] S. Barbosa, D. Cosley, A. Sharma, and R. M. Cesar Jr. Averaging gone wrong: Using time-aware analyses to better understand behavior. In *Proceedings of the 25th International Conference on World Wide Web*, pages 829–841. International World Wide Web Conferences Steering Committee, 2016.

[8] BBC. Indonesia post-election protests leave six dead in jakarta, 2019.

[9] I. Birman. Moderation in different communities on reddit–a qualitative analysis study. 2018.

[10] R. Bolgov, O. Filatova, and E. Semenova. Social media in mexico, argentina and venezuela: legal and political framework. In *2017 Conference for E-Democracy and Open Government (CeDEM)*, pages 253–259. IEEE, 2017.

[11] S. Bradshaw and P. Howard. Troops, trolls and troublemakers: A global inventory of organized social media manipulation. 2017.

[12] P. B. Brandtzæg and J. Heim. User loyalty and online communities: why members of online communities are not faithful. In *Proceedings of the 2nd international conference on INtelligent TEchnologies for interactive enterTAINment*, page 11. ICST (Institute for Computer Sciences, Social-Informatics and . . . , 2008.

[13] B. S. Butler, P. J. Bateman, P. H. Gray, and E. I. Diamant. An attraction-selection-attrition theory of online community size and resilience. *Mis Quarterly*, 38(3):699–728, 2014.

[14] Central Intelligence Agency. The world factbook - central intelligence agency, 2019.

[15] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *fourth international AAAI conference on weblogs and social media*, 2010.

[16] M. P. Chandra et al. On the generalised distance in statistics. In *Proceedings of the National Institute of Sciences of India*, volume 2, pages 49–55, 1936.

[17] E. Chandrasekharan, U. Pavalanathan, A. Srinivasan, A. Glynn, J. Eisenstein, and E. Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):31, 2017.

[18] E. Chandrasekharan, M. Samory, S. Jhaver, H. Charvat, A. Bruckman, C. Lampe, J. Eisenstein, and E. Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):32, 2018.

[19] A. Chaney, H. Wallach, M. Connelly, and D. Blei. Detecting and characterizing events. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1142–1152, 2016.

[20] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec. Do cascades recur? In *Proceedings of the 25th International Conference on World Wide Web*, pages 671–681. International World Wide Web Conferences Steering Committee, 2016.

[21] F. Chollet et al. Keras. https://keras.io, 2015.

[22] W.-Y. S. Chou, A. Oh, and W. M. Klein. Addressing health-related misinformation on social media. *Jama*, 320(23):2417–2418, 2018.

[23] J. D. Clark, R. M. Faris, R. J. Morrison-Westphal, H. Noman, C. B. Tilton, and J. L. Zittrain. The shifting landscape of global internet censorship. 2017.

[24] L. Collins. Gatekeepers of our lives [internet censorship]. *Engineering & Technology*, 12(10):28–31, 2017.

[25] S. Datta and E. Adar. Extracting inter-community conflicts in reddit. *arXiv preprint arXiv:1808.04405*, 2018.

[26] C. Dewey. Censorship, fat-shaming and the 'reddit revolt': How reddit became the alamo of the internet's ongoing culture war, 2015.

[27] A. Dubey, E. Moro, M. Cebrian, and I. Rahwan. Memesequencer: Sparse matching for embedding image macros. In *IW3C2 Web Conference*, pages 1225–1235, 2018.

[28] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[29] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 World Wide Web Conference*, pages 913–922. International World Wide Web Conferences Steering Committee, 2018.

[30] M. Glenski, E. Saldanha, and S. Volkova. Characterizing speed and scale of cryptocurrency discussion spread on reddit. In *The World Wide Web Conference*, pages 560–570, 2019.

[31] A. G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971.

[32] D. Henry, E. Stattner, and M. Collard. Social media, diffusion under influence of parameters: survey and perspectives. *Procedia Computer Science*, 109:376–383, 2017.

[33] J. Hollingsworth. Joko widodo secures second term as indonesia's president, 2019.

[34] C. Hui, M. Goldberg, M. Magdon-Ismail, and W. A. Wallace. Simulating the diffusion of information: An agent-based modeling approach. *International Journal of Agent Technologies and Systems (IJATS)*, 2(3):31–46, 2010.

[35] S. Jaki, T. De Smedt, M. Gwóźdź, R. Panchal, A. Rossa, and G. De Pauw. Online hatred of women¡? br?¿ in the incels. me forum: Linguistic analysis and automatic detection. *Journal of Language Aggression and Conflict*, 7(2):240–268, 2019.

[36] E. Joyce and R. E. Kraut. Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3):723–747, 2006.

[37] D. K. Kilgo, Y. M. M. Ng, M. J. Riedl, and I. Lacasa-Mas. Reddit's veil of anonymity: Predictors of engagement and participation in media environments with hostile reputations. *Social Media+ Society*, 4(4):2056305118810216, 2018.

[38] S. Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[39] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. AcM, 2010.

[40] C. Machado, B. Kira, V. Narayanan, B. Kollanyi, and P. Howard. A study of misinformation in whatsapp groups with a focus on the brazilian presidential elections. In *ACM World Wide Web Conference*, pages 1013–1019, 2019.

[41] B. Martin. Censorship and free speech in scientific controversies. *Science and Public Policy*, 42(3):377–386, 2014.

[42] A. Massanari. # gamergate and the fappening: How reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346, 2017.

[43] B. McEwan. Communication of communities: linguistic signals of online groups. *Information, Communication & Society*, 19(9):1233–1249, 2016.

[44] R. A. Mills. Pop-up political advocacy communities on reddit. com: Sandersforpresident and the donald. *AI & SOCIETY*, 33(1):39–54, 2018.

[45] S. Mohan, A. Guha, M. Harris, F. Popowich, A. Schuster, and C. Priebe. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer, 2017.

[46] S. Nagrecha, P. B. Thomas, K. Feldman, and N. V. Chawla. Predicting chronic heart failure using diagnoses graphs. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 295–312. Springer, 2017.

[47] Netblocks. Evidence of regional internet blackouts across venezuela.

[48] Netblocks. Major internet disruptions in venezuela amid protests.

[49] Netblocks. Social media outage and disruptions in venezuela amid incident in caracas.

[50] E. Newell, D. Jurgens, H. M. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[51] E. Panek, C. Hollenbach, J. Yang, and T. Rhodes. The effects of group size and time on the formation of online communities: Evidence from reddit. *Social Media+ Society*, 4(4):2056305118815908, 2018.

[52] K. Porter. Analyzing the darknetmarkets subreddit for evolutions of tools and trends using lda topic modeling. *Digital Investigation*, 26:S87–S97, 2018.

[53] T. A. Press. Venezuelan opposition targeted by internet censors, 2019.

[54] J. Roozenbeek and A. S. Palau. I read it on reddit: Exploring the role of online communities in the 2016 us elections news cycle. In *International Conference on Social Informatics*, pages 192–220. Springer, 2017.

[55] V. M. Rumata and A. S. Sastrosubroto. Net-attack 2.0: Digital post-truth and its regulatory challenges in indonesia. In *International Conference of Communication Science Research (ICCSR 2018)*. Atlantis Press, 2018.

[56] K. Saha, E. Chandrasekharan, and M. De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 11th ACM Conference on Web Science*, 2019.

[57] H. M. Saleem and D. Ruths. The aftermath of disbanding an online hateful community. *arXiv preprint arXiv:1804.07354*, 2018.

[58] J. Shin, L. Jian, K. Driscoll, and F. Bar. The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, 83:278–287, 2018.

[59] N. Smith. Yep, bitcoin was a bubble. and it popped, 2018.

[60] B. Strang, P. van der Putten, J. N. van Rijn, and F. Hutter. Don't rule out simple models prematurely: A large scale benchmark comparing linear and non-linear classifiers in openml. In *International Symposium on Intelligent Data Analysis*, pages 303–315. Springer, 2018.

[61] C. Tan. Tracing community genealogy: how new communities emerge from the old. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[62] The World Bank. The world bank in indonesia, 2019.

[63] P. B. Thomas, R. Krohn, and T. Weninger. Dynamics of team library adoptions: An exploration of github commit logs. *arXiv preprint arXiv:1907.04527*, 2019.

[64] P. B. Thomas, D. H. Robertson, and N. V. Chawla. Predicting onset of complications from diabetes: a graph based approach. *Applied network science*, 3(1):48, 2018.

[65] C. M. Valensise, M. Cinelli, A. Galeazzi, and W. Quattrociocchi. Drifts and shifts: Characterizing the evolution of users interests on reddit. *arXiv preprint arXiv:1912.09210*, 2019.

[66] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM, 2004.

[67] M. J. Vioulès, B. Moulahi, J. Azé, and S. Bringay. Detection of suicide-related posts in twitter data streams. *IBM Journal of Research and Development*, 62(1):7–1, 2018.

[68] Wall Street Journal. Indonesian voters dip their fingers and pose, 2014.

[69] K. Welbers and M. Opgenhaffen. Social media gatekeeping: An analysis of the gatekeeping influence of newspapers public facebook pages. *new media & society*, 20(12):4728–4747, 2018.

[70] C. Widanage, J. Li, S. Tyagi, R. Teja, B. Peng, S. Kamburugamuve, D. Baum, D. Smith, J. Qiu, and J. Koskey. Anomaly detection over streaming data: Indy500 case study. In *2019 IEEE 12th International Conference on Cloud Computing (CLOUD)*, pages 9–16. IEEE, 2019.

[71] M. J. Wolf. Beyond vader. *Disney's Star Wars: Forces of Production, Promotion, and Reception*, page 289, 2019.

[72] J. Wright, A. Darer, and O. Farnan. Automated discovery of internet censorship by web crawling. Association for Computing Machinery, 2018.

[73] S. Wu and B. Mai. Talking about and beyond censorship: Mapping topic clusters in the chinese twitter sphere. *International Journal of Communication*, 13:23, 2019.

[74] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *ACM Internet Measurement Conference*, pages 188–202, 2018.

A Appendix

In addition to demographic information, the following questions were asked via a phone survey by our partners GeoPoll in Indonesia. First the survey asks about general news behavior, and then about specific news stories that we have deemed to be false, questionable, or true. Our goal is to understand which demographics of people believe which false news stores - and how the belief in these stories change over time.

- How much do you follow national political affairs and events in Indonesia?
- How much do you follow international current affairs and events?
- How much do you follow local current affairs and events in your commmunity?
- How frequently do you access Social Media platforms such as Instagram/Facebook/WhatsApp ?
- How frequently do you watch TV?
- How frequently do you listen to radio?
- How frequently do you read a physical newspaper or magazine?
- What is your primary source of news? Reply with the actual name of the platform or outlet you get your news from.
- Other than for personal stories and photos, which of these social media platforms do you regularly read/watch local news and events on?
- Other than for personal stories and photos, which of these social media platforms do you regularly read/watch global news and events on?
- Do you have Facebook account?
- Do you have WhatsApp account?
- Do you have Youtube account?
- Do you have Twitter account?
- Do you have any Other social media account?
- What other social social media account do you use most? Reply with the name of the account you use most not already mentioned.
- Thinking back to the last time you used social media, how often do you read news stories before you share them with others?
- Do you share news stories that you know to be false?

- Why do you share news stories that you know to be false? Reply with the main reason why you share.
- Why don't you share news stories that you know to be false? Reply with the main reason why you do not share.

Now we're going to ask about some recent news events:

- (TRUE) In February of 2019, Fajar reported that the use of herbicides on chilli crops is not meant to increase the price of chilli, but because they are old and new crops need to be planted. Did you hear of this story? How do you rate its accuracy?
- (TRUE) In July of 2018, The Guardian reported that 'Team buzzer' is spreading misinformation in Indonesia. Did you hear of this story? How do you rate its accuracy?
- (TRUE) In July of 2018, Katadata reported that Pertaminia (state-owned oil company in Indonesia) is not bankrupt. Did you hear of this story? How do you rate its accuracy?
- (TRUE) In MONTH of YEAR, SOURCE reported that people are lying to you using the Quran (in the speech that resulted in his imprisonment). Did you hear of this story? How do you rate its accuracy?
- (TRUE) In MONTH of YEAR, SOURCE reported that nobody survived the Lion Air Flight 610 flight. Did you hear of this story? How do you rate its accuracy?
- (MISLEADING) In December of 2018 Turnbackhoax reported that Indosat did not send chain SMS messages during the 212 alumni reunion event. Did you hear of this story? How do you rate its accuracy?
- (MISLEADING) In August of 2018 Eliefweb reported that public assistance being sent to Lombok after the earthquake is not directly from the government or the president. Did you hear of this story? How do you rate its accuracy?
- (MISLEADING) In September of 2018 Detik News reported that the founder of NetTV Wishnutama Kusubandio appeard in a video with current VP candidate Sandiaga Uno in July 2016. He did not mean for it to be political. Did you hear of this story? How do you rate its accuracy?
- (MISLEADING) In February of 2019 Merdeka reported that the man who expressed support for Prabowo at the UN is not a diplomat or delegate. Did you hear of this story? How do you rate its accuracy?
- (MISLEADING) In October of 2018 Merdeka reported that Habieb Rizieq's children were turned around at the Yemen/Oman border, but ALL foreigners are not allowed to pass through that border. Did you hear of this story? How do you rate its accuracy?
- (FALSE) In MONTH of YEAR, SOURCE reported that Chinese citizens were arrested by TNI AD members for making fake ID cards. Did you hear of this story? How do you rate its accuracy?
- (FALSE) In MONTH of YEAR, SOURCE reported that Vaccines are poison. Did you hear of this story? How do you rate its accuracy?
- (FALSE) In MONTH of YEAR, SOURCE reported that religious education will be removed in Indonesia. Did you hear of this story? How do you rate its accuracy?
- (FALSE) In MONTH of YEAR, SOURCE reported that Ahok gets special treatment in prison. Did you hear of this story? How do you rate its accuracy?