Cyprus University of Technology

Faculty of Engineering and Technology

Doctoral Dissertation

Towards Understanding the Information Ecosystem Through the Lens of Multiple Web Communities

Savvas Zannettou

Limassol, October 2019

CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING, COMPUTER ENGINEERING AND INFORMATICS

Doctoral Dissertation

Towards Understanding the Information Ecosystem Through the Lens of Multiple Web Communities

Savvas Zannettou

Limassol, October 2019

Approval Form

Doctoral Dissertation

Towards Understanding the Information Ecosystem Through the Lens of Multiple Web Communities

Presented by

Savvas Zannettou

Supervisor: Dr. Michael Sirivianos, Assistant Professor at Cyprus University of Technology

Signature _____

Member of the committee: Dr. Andreas Andreou, Professor at Cyprus University of Technology

Signature _____

Member of the committee: Dr. Thorsten Strufe, Professor at Karlsruhe Institute of Technology

Signature _____

Cyprus University of Technology

Limassol, October 2019

Copyrights

The approval of the dissertation by the Department of Electrical Engineering, Computer Engineering and Informatics does not imply necessarily the approval by the Department of the views of the writer.

First, I would like to acknowledge my advisor, Michael Sirivianos, and my de facto co-advisor, Jeremy Blackburn, for their continuous support and feedback throughout my PhD journey. Their support and guidance was instrumental in turning me into an independent and competent researcher. Second, I would like to thank Emiliano De Cristofaro and *The Legendary* Gianluca Stringhini, who I consider my "second advisors." Their expertise and feedback complemented the one received by my two advisors, hence helping me in further expanding my research and writing skills. More importantly, I am grateful to these four individuals mainly because they ensured that this journey was fun while undertaking important and cool research.

Also, I want to thank various colleagues from Cyprus University of Technology, Telefonica Research, University College London, Princeton University, and University of Illinois at Urbana-Champaign: their help and feedback was pivotal for undertaking the studies presented in this thesis.

Furthermore, I would like to thank my family for their support, patience, and encouragement, which ensured that I was mentally strong to overcome all the obstacles faced during my PhD journey.

Finally, I would like to thank anonymous users on 4chan for their comments on our papers, as well as for creating some high quality memes about our research: their comments and "meme magic" made this journey much more enjoyable.

ABSTRACT

The Web consists of numerous Web communities, news sources, and services, which are often exploited by various entities for the dissemination of false or otherwise malevolent information. Yet, we lack tools and techniques to effectively track the propagation of information across the multiple diverse communities, and to capture and model the interplay and influence between them. Furthermore, we lack a basic understanding of what the role and impact of some emerging communities and services on the Web information ecosystem are, and how such communities are exploited by bad actors (e.g., state-sponsored trolls) that spread false and weaponized information.

In this thesis, we shed some light on the complexity and diversity of the information ecosystem on the Web by presenting a typology that includes the various types of false information, the involved actors as well as their possible motives. Then, we follow a data-driven cross-platform quantitative approach to analyze billions of posts from Twitter, Reddit, 4chan's Politically Incorrect board (/pol/), and Gab, to shed light on: 1) how news and image-based memes travel from one Web community to another and how we can model and quantify the influence between the various Web communities; 2) characterizing the role of emerging Web communities and services on the information ecosystem, by studying Gab and two popular Web archiving services, namely the Wayback Machine and archive.is; and 3) how popular Web communities are exploited by state-sponsored actors for the purpose of spreading disinformation and sowing public discord.

In a nutshell, our analysis reveal that small fringe Web communities like 4chan's /pol/ and The_Donald subreddit have a disproportionate influence on mainstream communities such as Twitter with regard to the dissemination of news and image-based memes. We find that Gab acts as the new hub for the alt-right community, while for Web archiving services we find that they are popular on fringe Web communities and that they can be misused by Reddit moderators in order to penalize ad revenue from news sources with conflicting ideology. Finally, when studying state-sponsored actors, we find that they exhibit substantial differences compared to random users, that their tactics change and evolve over time, and that they were particularly influential in spreading news on popular mainstream communities like Twitter and Reddit.

Keywords: misinformation, disinformation, Twitter, Reddit, 4chan, Gab, information warfare

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

/pol/:	4chan's Politically Incorrect board
/sp/:	4chan's Sports Board
/int/:	4chan's International Board
/sci/:	4chan's Science Board
API:	Application Programming Interface
AUC:	Area under the ROC Curve
CNN:	Convolutional Neural Networks
CDF:	Cumulative Distribution Function
ICO:	Initial Coin Offering
ISIS:	Islamic State of Iraq and the Levant
KYM:	Know Your Meme
LDA:	Latent Dirichlet Allocation
ML:	Machine Learning
NSFW:	Not-Safe-For-Work
OSN:	Online Social Network
ROC:	Receiver Operating Characteristic
RT:	Russia Today
SOM:	Self-Organizing Map
SVM:	Support Vector Machines
T_D:	The_Donald subreddit
TF-IDF:	Term Frequency-Inverse Document Frequency
VPN:	Virtual Private Network

LIST OF PUBLICATIONS

A large body of work presented in this thesis is already published in peer-reviewed journal, conference, and workshop papers. Specifically:

- Zannettou, S., Sirivianos, M., Blackburn, J. and Kourtellis, N., 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality (JDIQ), 11(3), p.10.

- Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G. and Blackburn, J., 2017, November. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens Of Mainstream and Alternative News Sources. In Proceedings of the 2017 Internet Measurement Conference (pp. 405-417). ACM.

- Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G. and Suarez-Tangil, G., 2018, October. On the Origins of Memes By Means of Fringe Web Communities. In Proceedings of the Internet Measurement Conference 2018 (pp. 188-202). ACM **(Distinguished Paper Award)**.

- Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G. and Blackburn, J., 2018, April. What is Gab: A Bastion of Free Speech or An Alt-Right Echo Chamber. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 1007-1014).

- Zannettou, S., Blackburn, J., De Cristofaro, E., Sirivianos, M. and Stringhini, G., 2018, June. Understanding Web Archiving Services and Their (Mis) Use on Social Media. In Twelfth International AAAI Conference on Web and Social Media.

- Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G. and Blackburn, J., 2019, May. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 218-226). ACM **(Best Paper Award)**.

- Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G. and Blackburn, J., 2018. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. In Proceedings of the 2019 ACM on Web Science Conference. ACM.

Chapter 1

Introduction

Over the past decades, the Web became the predominant medium for the rapid acquisition of information. Unfortunately, the Web has also become a medium where false information, hateful content, and weaponized information is disseminated. Recently, we have seen extensive anecdotal evidence suggesting that users on the Web are exposed to such information. Some examples include the spread of false and weaponized information by state-sponsored actors, namely Russian trolls were disseminating weaponized information related to the vaccine debate [1] and the 2016 US elections [2] to sow public discord and likely change voting preferences of people. Another example is the one of the now infamous Cambridge Analytica: during the 2016 US elections the company targetted millions of people in the US on Facebook by exposing them to political weaponized information with the goal to shift their voting decision [3]. On top of this, regular users on the Web are also involved in the dissemination of false, hateful, or weaponized information, which further compounds these emerging problems on the Web. Therefore, there is a pressing need to understand how information is shared on the Web, who the main entities involved are, and how information shared on the Web can alter real-world behavior.

At the same time, the information ecosystem on the Web has become an enormous and complex medium. It involves various entities that contribute to the dissemination of information: ranging from Web communities like Twitter, where users can share information, to news sources that disseminate articles to users. On top of this, the barrier of entry of new communities and news sources is minimal, hence the Web is becoming more complex as more communities and news sources are added. Due to the increased complexity of the ecosystem a lot of its aspects are relatively unstudied by the research community.

In this thesis, we focus on the following aspects of the information ecosystem: 1) how we can track the propagation of information across multiple Web communities and how to study the interplay and influence between these communities; 2) characterizing the role of emerging Web communities and services; and 3) understanding the exploitation of Web communities from bad actors for the purpose of advancing an agenda or sowing public discord. We focus on these mainly because we argue that providing knowledge and tools to analyze these aspects is a significant step towards understanding and mitigating emerging socio-technological phenomena on the Web. Such phenomena include studying the spread of hateful, fake, and weaponized information that can have great impact on the world (e.g., excessive spread of weaponized and targeted information can lead to shift in voting results of major elections). We elaborate on each of these aspects below.

Spread of information across multiple Web communities. As the number and diversity of communities and news sources grow, so does the opportunity for the production and dissemination of hateful or fake content. Nevertheless, previous work (see Chapter 2) only examined the propagation of information on the Web, to the best of our knowledge, by looking at specific communities in isolation. In reality, however, the various communities on the Web do not exist in vacuum. Users are members of multiple communities and they can share information seen on one community to another, possibly mutating it along the way. Such interactions, indicate that information travels from one Web community to another, hence denoting influence from the *source* to the *destination* Web community. Furthermore, anecdotal evidence emerged suggesting that fake news dissemination might start on fringe Web communities, eventually reaching mainstream communities and likely affecting the opinion of a vast amount of people [4, 5, 6]. Nevertheless, as a research community, we lack tools to effectively track the propagation of information across multiple Web communities, and more importantly, we lack knowledge on understanding the interplay and influence between multiple Web communities. Gaining this understanding will be extremely important for the research community and the public, allowing us to understand and mitigate emerging pressing issues of our era like the spread of hateful, weaponized, and fake information across the Web.

Characterizing the Role of Emerging Web Communities and Services on the Information Ecosystem. As new communities are added on the Web, we have limited knowledge of their role on the ecosystem. One example is the Gab social network [7], which was introduced back in 2016 as an alternative to Twitter. This specific Web community claims to be all about free speech and welcomes users banned from other Web communities. However, anecdotal evidence suggests that this community has become the new hub for the alt-right community and

likely it is used for the dissemination of false or hateful content [8]. Therefore, it is important to gain knowledge on what content is disseminated in these emerging Web communities, what users are attracted, and how such emerging communities affect the Web.

Information can be extremely diverse: the same piece of information can be disseminated via text, images, and URLs, hence constituting the tracking of information on the Web a non-straightforward task. In particular, URLs have several aspects that need to be considered. First, the information provided by the URL can change when the source updates the page. Second, URLs can get inaccessible after some time, a problem known as *link rot* [9], which can affect references across the Web. Third, there are several services that work with URLs that add complexity to studies that use URLs. An example is URL shorteners that generate a shortened URL, which redirects the user to the source page (i.e., different URLs point to the same information). Another, more interesting example, is the one of archiving services. These services, archive the content of the URL at a specific point in time and provide a new URL. However, in contrast with URL shorteners, the page is served by the archiving service itself without redirection to the source. This aspect can have important implications to the Web, as the content will not be able to be changed by the original author and because traffic is taken away from the original source (i.e., content is served by the archiving service). Despite these interesting aspects, Web archiving services are relatively unstudied: we lack a general understanding of how these services are used on the Web and what is their role and impact on the information ecosystem.

Exploitation of Web communities from bad actors. The Web provides an ideal environment for the diffusion of information to a vast amount of people in a short period of time. Clearly, this aspect of the Web can become an extremely powerful and dangerous tool when exploited by bad actors. Recently, anecdotal evidence emerged that highlights how popular mainstream Web communities like Twitter were exploited by state-sponsored actors that disseminated disinformation on a wide-variety of subjects ranging from health issues [1] to politics [2]. These actors are employed by governments and they possess several online personas that disseminate specific information that helps in pushing the agenda of their government. Motivated by the real-world impact that these actors can have, popular mainstream communities like Twitter and Reddit, started working on identifying and removing such actors from their platforms. However, as a research community, we lack an understanding on the behavior of these actors on the Web and how they impact and disrupt the Web's information ecosystem.

Motivated by the above aspects of the information ecosystem on the Web, we focus on

providing answers to the following research questions (RQs):

- **RQ1:** What are the various types of false or otherwise malevolent information (e.g., propaganda) that exist on the Web, what are the main actors that contribute to the dissemination of false information, and what are their possible underlying motives?
- **RQ2:** How information propagates across multiple Web communities and how can we quantify the influence between Web communities?
- **RQ3:** What type of content is disseminated in small fringe Web communities like Gab, what user base they attract, and what is the influence and impact of these communities to the rest of the Web?
- **RQ4:** What is the role of Web archiving services and how are these services exploited by users on various Web communities. Also, how do such services impact the information ecosystem on the Web?
- **RQ5:** How are state-sponsored actors exploiting mainstream Web communities in order to disseminate weaponized and possibly fake content? Do these actors have substantial differences when compared to random users? More importantly, how do these actors evolve over time, and what is their influence on the Web.

To provide answers to these research questions, we follow a large-scale cross-platform data-driven quantitative approach. To do so, we first implement a data collection infrastructure that consists of various crawlers, which allow us to collect large-scale datasets from the Web. Then, we apply various statistical analysis and machine learning techniques to extract meaningful insights from the large-scale datasets. Specifically, we use the main following techniques:

- **Hawkes Processes [10]**: A statistical analysis framework that enable us to investigate possible causalities between events. We use this technique to assess the influence that various Web communities have to each other by modeling and fitting our datasets with Hawkes Processes. More details regarding this technique can be found in Section 2.2.
- **Changepoint Analysis [11]**: A statistical analysis technique that allows us to extract points in a time series where statistically significant changes occur. This is particularly useful as it allows us to isolate significant days in a time series and investigate why these changes occur on the various Web communities we study, and possibly link them to real-world events. More details on the methodology and application of this technique can be found in Section 5.1.

- **Neural Networks:** We apply neural networks for various purposes. For instance, we use word2vec [12], which are shallow neural networks, to understand the use of language in the various communities we study. Also, we use neural networks to build custom classifiers: e.g., we use Convolutional Neural Networks to build a custom screenshot classifier (see Section 4.2.2).

- **Graph Analysis & Visualization:** We leverage several graph analysis and visualization techniques to analyze data that can be modeled with graphs. Among other things, we use community detection techniques (e.g., the Louvain method [13]) to detect meaningful communities from the underlying graph structure, and graph layout techniques (e.g., OpenOrd [14] and ForceAtlas2 [15]), which allow us to lay out graphs in the space where the distance between nodes represents something useful (e.g., nodes that are layed out closer means they are more similar).

- **Clustering Algorithms:** We use traditional clustering algorithms for the purpose of creating groups of similar information. For instance, we use the DBSCAN algorithm [16] to cluster images that are visually similar with the ultimate goal to track the propagation of memes across the Web (more details can be found in Section 4.2.2).

1.1 Contributions

This thesis makes several contributions towards understanding the information ecosystem on the Web. We make contributions in three main lines of work: 1) understanding the spread of information across multiple Web communities; 2) characterizing emerging Web communities like Gab and assessing the role of Web archiving services on the information ecosystem; and 3) understanding the behavior and influence of state-sponsored actors on the Web's information ecosystem. In more detail, we make the following contributions:

- We provide a comprehensive overview of the information ecosystem on Web. To do this, we present a typology that sheds light into the types of false information on the Web, the actors that are involved as well as their possible underlying motives (**RQ1**).

- We introduce a novel methodology, based on Hawkes Processes, to quantify the influence between multiple Web communities. We applied this methodology to several datasets with the goal to quantify the influence that each community has on other communities with respect to the dissemination of news and image-based memes (**RQ2**).

- We present the first study of mainstream and alternative news shared on Twitter, Reddit, and 4chan, measuring how mainstream and alternative news flow between these platforms, and demonstrating how alt-right communities have surprisingly high influence on Twitter (**RQ2**).

- We design and implement a highly scalable processing pipeline that is able to track the propagation of image-based memes across multiple Web communities.[1] By applying the proposed pipeline to 160M images posted on Twitter, Reddit, 4chan, and Gab, we study the memes ecosystem and characterize each community with respect to the memes their users share (**RQ2**).

- We provide some exploratory analyses on some relatively unknown, by the time we started looking at them, communities and services. Specifically, we provide the first study on Gab, finding that it is becoming the new alt-right's hub despite the fact that it started as a social network promoting free speech (**RQ3**). Furthermore, we study two Web archiving services, the Wayback Machine and `archive.is`, and their use on Twitter, Reddit, 4chan, and Gab, finding several "nuggets": 1) these services are used to archive news content and are extensively used by fringe Web communities like 4chan; 2) these services are exploited to a large extent by Reddit bots; 3) these services can be used to deprive ad revenue from the original source and we find evidence that Reddit moderators actually "force" users to share archived content from sources with opposing ideology in order to deprive them of ad revenue (**RQ4**).

- We study the behavior of state-sponsored actors on the Web, finding that they exhibit substantial differences when compared to a set of random users. We find that they change their behavior and that they target different populations over time. Also, we quantified the influence that these actors had to Twitter, Reddit, 4chan, and Gab, finding that these actors had a disproportionate influence to the rest of the platforms, with respect to the dissemination of news URLs (**RQ5**).

1.2 Peer-Reviewed Papers

A large body of work presented in this thesis is already published in peer-reviewed journal, conference, and workshop papers. Specifically, some aspects of our work (in collaboration with other researchers and academics) appear in the following papers:

[1] We make the memes processing pipeline publicly available so it can be used by other researchers [17]

- Zannettou, S., Sirivianos, M., Blackburn, J. and Kourtellis, N., 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. Journal of Data and Information Quality (JDIQ), 11(3), p.10.

- Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G. and Blackburn, J., 2017, November. The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens Of Mainstream and Alternative News Sources. In Proceedings of the 2017 Internet Measurement Conference (pp. 405-417). ACM.

- Zannettou, S., Caulfield, T., Blackburn, J., De Cristofaro, E., Sirivianos, M., Stringhini, G. and Suarez-Tangil, G., 2018, October. On the Origins of Memes By Means of Fringe Web Communities. In Proceedings of the Internet Measurement Conference 2018 (pp. 188-202). ACM **(Distinguished Paper Award)**.

- Zannettou, S., Bradlyn, B., De Cristofaro, E., Kwak, H., Sirivianos, M., Stringini, G. and Blackburn, J., 2018, April. What is Gab: A Bastion of Free Speech or An Alt-Right Echo Chamber. In Companion of the The Web Conference 2018 on The Web Conference 2018 (pp. 1007-1014).

- Zannettou, S., Blackburn, J., De Cristofaro, E., Sirivianos, M. and Stringhini, G., 2018, June. Understanding Web Archiving Services and Their (Mis) Use on Social Media. In Twelfth International AAAI Conference on Web and Social Media.

- Zannettou, S., Caulfield, T., De Cristofaro, E., Sirivianos, M., Stringhini, G. and Blackburn, J., 2019, May. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In Companion Proceedings of The 2019 World Wide Web Conference (pp. 218-226). ACM **(Best Paper Award)**.

- Zannettou, S., Caulfield, T., Setzer, W., Sirivianos, M., Stringhini, G. and Blackburn, J., 2018. Who Let The Trolls Out? Towards Understanding State-Sponsored Trolls. In Proceedings of the 2019 ACM on Web Science Conference. ACM.

1.3 Thesis Organization

The remainder of this thesis is structured as follows: Chapter 3 describes previous work on: 1) user perception and interaction with false information on the Web; 2) propagation

of information on the Web; 3) detection and containment of false information on the Web; and 4) false information on the political stage. Chapter 2 provides the required background. In Chapter 4 we present our work that focuses on understanding how information spreads from one Web community to another and how to measure the influence between multiple communities. In Chapter 5 we present our work related to characterizing various communities and services on the information ecosystem; namely, we study Gab and Web archiving services. Chapter 6 describes our work on understanding the role and impact of state-sponsored actors on the Web. Finally, we conclude in Chapter 7.

Chapter 2

Background

In this chapter, we present useful background information regarding the Web communities we study, as well as some statistical techniques that we use to analyze data from various Web communities. Specifically, for the former, we briefly describe Twitter, Reddit, 4chan, and Gab, while for the latter, we overview Hawkes Processes for estimating influence between Web communities.

2.1 Web Communities

In this section, we briefly describe the Web communities that we use in our work, as well as the methodology for collecting data from each community.

2.1.1 Twitter

General. Twitter is a popular microblogging social network. Users can broadcast 280-character messages, called "tweets", to their followers. By default, tweets are publicly available, however, users are able to restrict tweets to be available only to their followers. Twitter includes several traditional social networking features like sharing other tweets (i.e., retweet), liking tweets, as well as posting tweets in reply to other tweets. On top of this, users can use hashtags (#) in their tweets, which can help other users to find and weight in on tweets with specific content. Also, users can refer to other users by mentioning them in tweets (i.e., by using the @ character).

Moderation. Twitter moderates content on their site with the goal to remove hateful content

and ban users that incite violence or share hateful content. Some examples include the permanent ban of Milo Yiannopoulos after continuing hateful abuse against Leslie Jones[1] and the permanent ban of several accounts linked to the alt-right because of targeted abuse and harassment of others.[2] Also, Twitter employs a demoting system which automatically hides content that is likely to be abusive and the user can only see the possibly hateful tweets by pressing a button.[3]

2.1.2 Reddit

General. Reddit is called the "front page of the Internet" and it is a popular news aggregator. Users can create threads (called "submissions") by posting a URL along with a title. Other users can reply below in a structured manner (e.g., reply to submission or reply to specific reply). Popularity of content within the platform is determined via a voting system: each comment or submission can be up-voted or down-voted, hence a score can be calculated. Submissions and comments with higher score appears on top of submissions and comments with lower score. Also, there is a user-based "score" called *karma* that is basically the sum of scores for all of the user's comments and submissions. Note that on Reddit the community structure is not defined by the friendship/follower relation like Twitter (a user can list another user as a friend but it does not change anything in the structure or use of the platform).

Subreddits. Reddit is divided into millions of communities called "subreddits".[4] Subreddits are created from users of the platform and this has lead to a plethora of communities discussing a wide variety of topics ranging from video games, to politics, pornography, and even meta-communities that summarize interactions of users on other subreddits/social networks. Subreddits are monitored by Reddit's administrators and they are removed when they share "extremely inappropriate" content. For instance, in the past, Reddit removed subreddits related to the promotion of conspiracy theories (e.g., /r/greatawakening, which promoted the Qanon conspiracy theory [18]), subreddits that shared suggestive photos of underage girls (e.g., /r/jailbait), as well as subreddits sharing "deepfakes" (e.g., /r/deepfakes).[5]

[1] `http://fxn.ws/2zshTl8`

[2] `https://wapo.st/2fYdQRG`

[3] `https://cnnmon.ie/2smPCaF`

[4] According to statista as of 2017 there are nearly 1.2M subreddits (`https://www.statista.com/chart/11882/number-of-subreddits-on-reddit/`).

[5] For a list of controversial subreddits that were removed see `https://en.wikipedia.org/wiki/Controversial_Reddit_communities`.

2.1.3 4chan

General. 4chan is a discussion forum known as an imageboard.[6] Users can create a new thread by creating a post that must include an image. Other users can reply to the thread (images are optional in replies) and possibly add references or quotes to previous posts within the thread. 4chan is an anonymous community: users are not required to have an account in order to create a post. At the same time, users can add a pseudonym when posting, however, their pseudonym is bounded to the specific post and they can use a different one for other posts. On top of this, each post is associated with a *flag*. Usually, the flag is determined based on the location of the user, however, it can be tricked by the use of Virtual Private Networks (VPN). Furthermore, there are communities within 4chan that introduce custom flags. For instance, 4chan's Politically Incorrect board (/pol/), allow users to either add the flag based on their location or from a set of 23 pre-defined flags. Examples of such flags include the flag of Kekistan, the Nazi flag, confederate flag, etc.

Boards. 4chan is divided into multiple communities called *boards*, which are defined by 4chan. Each board has its own general theme and topic, ranging from politics, to sports and pornography, and likely attracts different user bases. For instance, 4chan's Politically Incorrect board (/pol/) focuses on discussions of politics and world news, while the Video games board (/v/) focuses on discussions around video games. As of April 2019, 4chan has 70 different boards. In our work, we mainly focus on 4chan /pol/ board, which is mainly used for the discussion of news and real-world events that are happening. Also, we are particularly interested in this community since previous work showed that it exhibits a high degree of hate speech and racism [19], and because of anecdotal evidence that suggests the community's influence and impact both on the online and offline world (e.g., spread of Pizzagate conspiracy theory [20]).

Moderation. 4chan has an extremely lax moderation: each board has a handful of volunteers called *janitors* that are moderating each board. Janitors can remove posts and threads and recommend user bans to 4chan employees. Generally, Janitors pretty much allow everything to be posted as long as it is relevant to the general topic and theme of the board. Due to this lax moderation and anonymous nature of the community, 4chan users can use whatever tone and language to express themselves, hence 4chan's high degree of hateful content.

Ephemerality. 4chan is an ephemeral community. Each board has a finite amount of active

[6] `http://4chan.org/`

threads. Threads are removed after a relatively short period based on a "bumping system" that considers the posting activity within the thread. That is, creating a new thread, results in the *archival* of the thread with the least recent post. A new post within a thread can help "bump" the thread as it keeps the thread alive and makes the thread appear at the top of the board. To avoid having a thread alive forever, 4chan has *bump* and *image* limits, which determine the maximum number of images and bumps that a thread can receive. Once a thread is archived, it remains to the community for 7 days before getting deleted forever.

2.1.4 Gab

General. Gab is a new social network, launched in August 2016, that "champions free speech, individual liberty, and the free flow of information online.[7]" It combines social networking features that exist in popular social platforms like Reddit and Twitter. A user can broadcast 300-character messages, called "gabs," to their followers (akin to Twitter). From Reddit, Gab takes a modified voting system (which we discuss later). Gab allows the posting of pornographic and obscene content, as long as users label it as Not-Safe-For-Work (NSFW).[8] Posts can be reposted, quoted, and used as replies to other gabs. Similar to Twitter, Gab supports hashtags, which allow indexing and querying for gabs, as well as mentions, which allow users to refer to other users in their gabs.

Topics and Categories. Gab posts can be assigned to a specific *topic* or *category*. *Topics* focus on a particular event or timely topic of discussion and can be created by Gab users themselves; all topics are publicly available and other users can post gabs related to topics. *Categories* on the other hand, are defined by Gab itself, with 15 categories defined at the time of this writing. Note that assigning a gab to a category and/or topic is *optional*, and Gab moderates topics, removing any that do not comply with the platform's guidelines.

Voting system. Gab posts can get up- and down-voted; a feature that determines the popularity of the content in the platform (akin to Reddit). Additionally, each user has its own score, which is the sum of up-votes minus the sum of down-votes that it received to all his posts (similar to Reddit's user karma score [21]). This user-level score determines the popularity of the user and is used in a way unique to Gab: a user must have a score of at least 250 points to be able to down-vote other users' content, and every time a user down-votes a post a point from his user-level score is deducted. In other words, a user's score is used as a form of currency

[7] `http://gab.ai`

[8] What constitutes NSFW material is not well defined.

expended to down-vote content.

Moderation. Gab has a lax moderation policy that allows most things to be posted, with a few exceptions. Specifically, it only forbids posts that contain "illegal pornography" (legal pornography is permitted), posts that promote terrorist acts, threats to other users, and doxing other users' personal information [22].[9]

Monetization. Gab is ad-free and relies on direct user support. On October 4, 2016 Gab's CEO Andrew Torba announced that users were able to donate to Gab [23]. Later, Gab added "pro" accounts as well. "Pro" users pay a per-month fee granting additional features like live-stream broadcasts, account verification, extended character count (up to 3K characters per gab), special formatting in posts (e.g., italics, bold, etc.), as well as premium content creation. The latter allows users to create "premium" content that can only be seen by subscribers of the user, which are users that pay a monthly fee to the content creator to be able to view his posts. The premium content model allows for crowdfunding particular Gab users, similar to the way that Twitch and Patreon work. Finally, Gab is in the process of raising money through an Initial Coin Offering (ICO) with the goal to offer a "censorship-proof" peer-to-peer social network that developers can build application on top [24].

2.1.5 Remarks

In this section, we presented the data sources that we use in this thesis. We select these specific data sources for various reasons. First, our data sources comprise of an interesting mix of both mainstream Web communities (i.e., Twitter and Reddit), as well as fringe Web communities (i.e., Gab and 4chan). This enables us to understand how small fringe Web communities influence large mainstream Web communities. Second, we select these specific fringe Web communities mainly because of extensive anecdotal evidence that suggest that 4chan and Gab are involved in the dissemination of false information [20, 25] and hateful content [8, 26]. Third, we avoid using other popular social networks like Facebook mainly due to limits imposed on their APIs by the company itself, hence constituting the task of obtaining data non-straightforward.

Obviously, it is likely that on other Web communities we can find important differences that might affect the results presented in this thesis, however, this thesis sheds light into the information ecosystem through the lens of multiple Web communities highlighting the need to

[9] For more information on Gab's guidelines, see `https://gab.ai/about/guidelines`.

shift focus into understanding the various Web communities on the Web and study the interplay between them. Despite this fact, as we mention in the next section, our influence estimation experiments via Hawkes Processes allow us to also capture the creation of events from external sources, hence we argue that the Influence Estimation results presented throughout this thesis can be treated as general, as they shed light into the influence that each Web community have to the others by also considering external sources (i.e., communities that we do not study like Facebook).

2.2 Hawkes Processes

In this section, we provide necessary background for Hawkes Processes and how we use them in order to assess the interplay of multiple Web communities and, more importantly, quantify the influence that specific small fringe Web communities (e.g., 4chan) have to mainstream ones like Twitter. In a nutshell, Hawkes Processes is a statistical framework that allow us to assess the causality of events that occur on the Web, and find the possible root causes (i.e., Web community that is responsible for the creation of the events) along with their respective probabilities.

General. Hawkes Processes are self-exciting temporal point processes [10] that describe how events (e.g., posting of a URL or an image) occur on a set of processes (i.e., Web communities). Generally, a Hawkes model consists of a number, K, of point processes, each with a "background rate" of events $\lambda_{0,k}$. The background rate is the expected rate at which events will occur on a process *without* influence from the processes modeled or previous events; this captures events created for the first time, or those seen on a process we do not model and then created on a process we do.

An event on one process can cause an *impulse response* on other processes, which increases the probability of an event occurring above the processes' background rates. The shape of the impulse determines how the probability of these events occurring is distributed over time; typically the probability of another event occurring is highest soon after the original event and decreases over time.

Fig. 2.1 illustrates a Hawkes model with three processes. The first event occurs on process B, which causes an increase in the rate of events on all three processes. The second event then occurs on process C, again increasing the rate of events on the processes. The third event occurs soon after, on process A. The fourth event occurs later, again caused by the

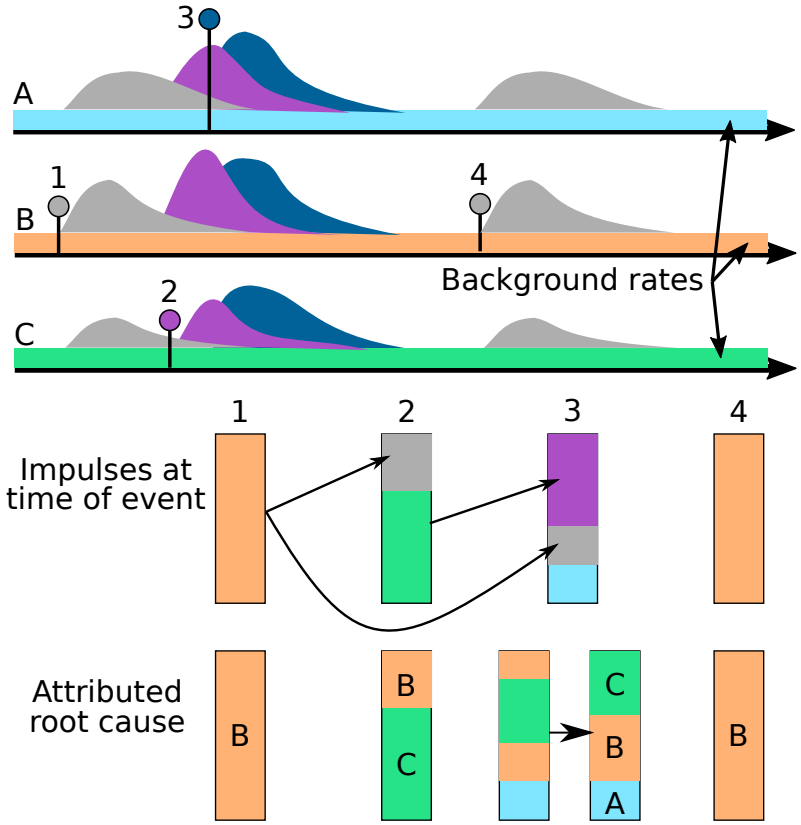


Figure 2.1: Representation of a Hawkes model with three processes. Events cause impulses that increase the rate of subsequent events in the same or other processes. By looking at the impulses present when events occur, the probability of a process being the root cause of an event can be determined. Note that on the second part of the Figure, colors represent events while arrows represent impulses between the events.

background arrival rate on process B, after the increases in arrival rate from the other events have disappeared.

To understand the influence different processes have on the creation of specific events, we want to be able to attribute the cause of an event being posted back to a specific process. For example, if an image is posted on /pol/ and then someone sees it there and posts it on Twitter where it is shared several times, we would like to be able to say that /pol/ was the *root cause* of those events. Obviously, we do not actually know where someone saw something and decided to share it, but we can, using the Hawkes models, determine the *probability* of each community being the root cause of an event.

Looking again at Fig. 2.1, we see that events 1 and 4 are caused directly by the background rate of process B. This is because, in the case of event 1, there are no previous events on other

processes, and in the case of event 4, the impulses from previous events have already stopped. Events 2 and 3, however, occur when there are multiple possible causes: the background rate for the process and the impulses from previous events. In these cases, we assign the probability of being the root cause in proportion to the magnitudes of the impulses (including the background rate) present at the time of the event. For event 2, the impulse from event 1 is smaller than the background rate of process C, so the background rate has a higher probability of being the cause of event 2 than event 1. Thus, most of the cause for event 2 is attributed to process C, with a lesser amount to B (through event 1). Event 3 is more complicated: impulses from both previous events are present, thus the probability of being the cause is split three ways, between the background rate and the two previous events. The impulse from event 2 is the largest, with the background rate and event 1 impulse smaller. Because event 2 is attributed both to processes B and C, event 3 is partly attributed to process B through both event 1 and event 2.

For our purposes, fitting a Hawkes model to a series of events on the different processes gives us values for the background rates for each process along with the probability of an event on one process causing events on other processes. We emphasize that the background rates of the Hawkes processes allows us to also account for the probability of an event caused by external sources of information. Thus, while we are only able to model the specific influences for a limited number of Web communities we study, the resulting probabilities are affirmatively attributable to each of them; the influence of the greater Web is captured by the background rates.

For a discrete-time Hawkes model, time is divided into a series of bins of duration Δt, and events occurring within the same time bin do not interact with each other. The rate of each k-th process, $\lambda_{t,k}$ is given by:

$$\lambda_{t,k} = \lambda_{0,k} + \sum_{k'=1}^{K} \sum_{t'=1}^{t-1} s_{t',k'} \cdot h_{k' \to k}[t - t']$$

where $s \in \mathbb{N}^{T \times K}$ is the matrix of event counts (how many events occur for process k at time t) and $h_{k' \to k}[t - t']$ is an impulse response function that describes the amplitude of influence that events on process k' have on the rate of process k.

Following [27], the impulse response function $h_{k \to k'}[t - t']$ can be decomposed into a scalar weight $W_{k \to k'}$ and a probability mass function $G_{k \to k'}[d]$. The weight specifies the strength of the interaction from process k to process k' and the probability mass function specifies how

the interaction changes over time:

$$h_{k\to k'}[d] = W_{k\to k'} G_{k\to k'}[d]$$

The weight value $W_{k\to k'}$ can be interpreted as the expected number of child events that will be caused on process k' after an event on process k. The probability mass function $G_{k\to k'}$ specifies the probability that a child event will occur at each specific time lag $d\Delta t$, up to a maximum lag Δt_{max}. This interpretation of $W_{k\to k'}$ is useful because it allows us to compare how much influence processes have on each other. For instance, we can examine whether an image posted on Twitter or on Reddit is more likely to cause the same image to be posted on 4chan, or if there is a difference in influence from one platform to another between various images.

Experimental Setup. We assume a Hawkes model that is is fully connected, i.e., it is possible for each process to influence all the others, as well as itself, which describes behavior where participants on a platform see an image and re-post it on the same platform. For example, with Twitter, this value ($W_{\text{Twitter}\to\text{Twitter}}$) would likely be quite high, given that tweets are commonly re-tweeted a number of times: the initial tweet containing an image is likely to cause a number of re-tweets, also containing the image, on the same platform.

For each type of event, we create a matrix $s \in \mathbb{N}^{T\times K}$ containing the number of events per minute for each of the processes (i.e., Web communities). Here, T is the number of minutes from the first recorded post of the event on any process, to the last recorded post of the event on any process (**NB:** this value can be different for each type of event). We select $\Delta t = 1$ minute as a reasonable compromise between accuracy and computational cost.

Next, we fit a Hawkes model for each type of event using the approach described in [27, 28], which uses Gibbs sampling to infer the parameters of the model from the data, including the weights, background rates, and shape of the impulse response functions between the different processes. By setting $\Delta t_{max} = 60 \cdot 12 = 720$ minutes, we say that a given event can cause other events within a 12-hour time window. Experiments with other values (6, 12, 24, and 48 hours) gave similar results. After fitting the models, we have the values for the W matrix – i.e., the weights of the interactions between events on different processes for each type of event. These weights can then be interpreted as the expected number of events. For example, $W_{Twitter\to/pol/} = 0.1$ would mean that an event on Twitter will cause n events on /pol/, where n is drawn from a Poisson distribution with rate parameter 0.1. Finally, for each type of event, we also get the $\lambda_{0,k}$ values for each process, which are the background rates for event arrivals that are *not* caused by other events in the system we model. Again, these background rates

capture events due to some *other* process, e.g., someone posting an image after seeing it on the original site or seeing the image on another site not included in the model, like Facebook.

Metrics. Having obtained the weight matrix W, which specifies the strength of connections between processes for each type of event, we report our influence estimation results using two metrics. First, we measure the absolute influence, which can be interpreted as the expected number of events that are created on a destination process because of events previously seen on a source process. Since the weight values can be interpreted as the expected number of additional events that will be caused a consequence of an event, we can estimate the percentage of events on each process that were caused by each of the other processes by multiplying the weight by the actual number of events that occurred on the source process (e.g., Process A) and dividing by the number of events that occurred on the destination process (e.g., Process B):

$$\text{Influence}_{A\to B} = \frac{\sum_{e\in\text{events}} \left(W_{A\to B} \cdot \sum_{t=1}^{T} s_{t,A}\right)}{\sum_{e\in\text{events}} \sum_{t=1}^{T} s_{t,B}}$$

Second, we measure the *efficiency* of each process in pushing events to the rest of the processes. To do this, we normalize the influence values by the total number of events in the source process (e.g., Process A). This metric allow us to see how much influence each process has, relative to the number of events that are created in the process and is given by:

$$\text{Efficiency}_{A\to B} = \frac{\sum_{e\in\text{events}} \left(W_{A\to B} \cdot \sum_{t=1}^{T} s_{t,A}\right)}{\sum_{e\in\text{events}} \left(\sum_{t=1}^{T} s_{t,A}\right)}$$

Remarks. In this section, we described how we can use Hawkes Processes for modeling the interplay between multiple Web communities and how to quantify the influence that each Web community have to the others. To do this, we leverage Bayesian Inference techniques and data that describes the appearances of events in a set of processes. This allow us to assess the causality of events that happen on multiple Web communities and assess the possible root causes (i.e., the responsible Web community for the creation of the event) for each event. Note that by tweaking what an event is, the proposed framework can be applied to a wide variety of use cases (e.g., an event can be text referring to a specific news story, a specific video, etc.).

Chapter 3

Literature Review

In this chapter, we provide an extensive literature review of work that focus on the false information ecosystem on the Web. First, we present a typology of the various types and actors that are involved in the spread of information on the Web. Then we review the following lines of work: 1) user perception of false information; 2) propagation of false information; 3) detection and containment of false information; 4) false information on the political stage and 5) various other studies that are relevant.

3.1 Typology of the False Information Ecosystem

In this section we present our typology, which we believe will provide a succinct roadmap for future work. The typology is based on [29] and extended to build upon the existing literature. Specifically, we describe the various types of false information that can be found in Online Social Networks (OSNs, Section 3.1.1), the various types of actors that contribute in the distribution of false information (Section 3.1.2), as well as their motives (Section 3.1.3). To extract the proposed typology, we first made an extensive literature review of papers that focus on understanding the spread of false information across the Web. Then, based on all the referenced types of false information, actors and motives, we created the proposed typology. In cases where various papers referenced similar types/actors/motives with various names, we selected the name where it was referenced most of the times in the literature. Note that our typology is different from concurrent work by Kumar and Shah [30] as we provide a fine-grained distinction for the types of false information, the actors, and their motives. Also, note that we make a best effort to cover as many aspects of the false information as per our

knowledge; however, the typology should not be treated as an exhaustive representation of the false information ecosystem.

3.1.1 Types of False Information

False information on the Web can be found in various forms, hence we propose the categorization of false information into eight types as listed below:

- **Fabricated (F) [31].** Completely fictional stories disconnected entirely from real facts. This type is not new and it exists since the birth of journalism. Some popular examples include fabricated stories about politicians and aliens [32] (e.g., the story that Hillary Clinton adopted an alien baby).

- **Propaganda (P) [33].** This is a special instance of the fabricated stories that aim to harm the interests of a particular party and usually has a political context. This kind of false news is not new, as it was widely used during World War II and the Cold War. Propaganda stories are profoundly utilized in political contexts to mislead people with the overarching goal of inflicting damage to a particular political party or nation-state. Due to this, propaganda is a consequential type of false information as it can change the course of human history (e.g., by changing the outcome of an election). Some recent examples of propaganda include stories about the Syria air strikes in 2018 or about specific movements like the BlackLivesMatter (see [34] for more examples).

- **Conspiracy Theories (CT) [35].** Refer to stories that try to explain a situation or an event by invoking a conspiracy without proof. Usually, such stories are about illegal acts that are carried out by governments or powerful individuals. They also typically present unsourced information as fact or dispense entirely with an "evidence" based approach, relying on leaps of faith instead. Popular recent examples of conspiracy theories include the Pizzagate theory (i.e., Clinton's campaign running a pedophile ring) [36] and conspiracies around the murder of Seth Rich [37] (e.g., Seth Rich was involved in the DNC email leaks).

- **Hoaxes (H) [38].** News stories that contain facts that are either false or inaccurate and are presented as legitimate facts. This category is also known in the research community either as half-truth [39] or factoid [40] stories. Popular examples of hoaxes are stories that report the false death of celebrities (e.g., the Adam Sadler death hoax [41]).

- **Biased or one-sided (B).** Refers to stories that are extremely one-sided or biased. In the political context, this type is known as Hyperpartisan news [42] and are stories that are extremely biased towards a person/party/situation/event. Some examples include the wide spread diffusion of false information to the alt-right community from small fringe Web communities like 4chan's /pol/ board [19] and Gab, an alt-right echo chamber [43].

- **Rumors (R) [44].** Refers to stories whose truthfulness is ambiguous or never confirmed. This kind of false information is widely propagated on OSNs, hence several studies have analyzed this type of false information. Some examples of rumors include stories around the 2013 Boston Marathon Bombings like the story that the suspects became citizens on 9/11 or that a Sandy Hook child was killed during the incident [45].

- **Clickbait (CL) [46].** Refers to the deliberate use of misleading headlines and thumbnails of content on the Web. This type is not new as it appeared years before, during the "newspaper era," a phenomenon known as yellow journalism [47]. However, with the proliferation of OSNs, this problem is rapidly growing, as many users add misleading descriptors to their content with the goal of increasing their traffic for profit or popularity [48]. This is one of the least severe types of false information because if a user reads/views the whole content then he can distinguish if the headline and/or the thumbnail was misleading.

- **Satire News (S) [49].** Stories that contain a lot of irony and humor. This kind of news is getting considerable attention on the Web in the past few years. Some popular examples of sites that post satire news are TheOnion [50] and SatireWire [51]. Usually, these sites disclose their satyric nature in one of their pages (i.e., About page). However, as their articles are usually disseminated via social networks, this fact is obfuscated, overlooked, or ignored by users who often take them at face value with no additional verification.

It is extremely important to highlight that there is an overlap in the aforementioned types of false information, thus it is possible to observe false information that may fall within multiple categories. Here, we list two indicative examples to better understand possible overlaps: 1) a rumor may also use clickbait techniques to increase the audience that will read the story; and 2) propaganda stories, which are a special instance of a fabricated story, may also be biased towards a particular party. These examples highlight that the false information ecosystem is extremely complex and the various types of false information need to be considered to mitigate the problem.

3.1.2 False Information Actors

In this section, we describe the different types of actors that constitute the false information propagation ecosystem. We identified a handful of different actors that we describe below.

- **Bots [52].** In the context of false information, bots are programs that are part of a bot network (Botnet) and are responsible for controlling the online activity of several fake accounts with the aim of disseminating false information. Botnets are usually tied to a large number of fake accounts that are used to propagate false information in the wild. A Botnet is usually employed for profit by 3rd party organizations to diffuse false information for various motives (see Section 3.1.3 for more information on their possible motives). Note that various types of bots exist, which have varying capabilities; for instance, some bots only repost content, promote content (e.g., via vote manipulation on Reddit or similar platforms), and others post "original" content. However, this distinction is outside of the scope of this work, which provides a general overview of the information ecosystem on the Web.

- **Criminal/Terrorist Organizations [53].** Criminal gangs and terrorist organizations are exploiting OSNs as the means to diffuse false information to achieve their goals. A recent example is the ISIS terrorist organization that diffuses false information in OSNs for propaganda purposes [53]. Specifically, they widely diffuse ideologically passionate messages for recruitment purposes. This creates an extremely dangerous situation for the community as there are several examples of individuals from European countries recruited by ISIS that ended-up perpetrating terrorist acts.

- **Activist or Political Organizations.** Various organizations share false information in order to either promote their organization, demote other rival organizations, or for pushing a specific narrative to the public. A recent example include the National Rifle Association, a non-profit organization that advocates gun rights, which disseminated false information to manipulate people about guns [54]. Other examples include political parties that share false information, especially near major elections [55].

- **Governments [56].** Historically, governments were involved in the dissemination of false information for various reasons. More recently, with the proliferation of the Internet, governments utilize the social media to manipulate public opinion on specific topics. Furthermore, there are reports that foreign goverments share false information on other countries in order to manipulate public opinion on specific topics that regard

the particular country. Some examples, include the alleged involvement of the Russian government in the 2016 US elections [57] and Brexit referendum [58].

- **Hidden Paid Posters [59] and State-sponsored Trolls [60, 61].** They are a special group of users that are paid in order to disseminate false information on a particular content or targeting a specific demographic. Usually, they are employed for pushing an agenda; e.g., to influence people to adopt certain social or business trends. Similar to bots, these actors disseminate false information for profit. However, this type is substantially harder to distinguish than bots because they exhibit characteristics similar to regular users.

- **Journalists [62].** Individuals that are the primary entities responsible for disseminating information both to the online and to the offline world. However, in many cases, journalists are found in the center of controversy as they post false information for various reasons. For example, they might change some stories so that they are more appealing, in order to increase the popularity of their platform, site, or newspaper.

- **Useful Idiots [63].** The term originates from the early 1950s in the USA as a reference to a particular political party's members that were manipulated by Russia in order to weaken the USA. Useful idiots are users that share false information mainly because they are manipulated by the leaders of some organization or because they are naive. Usually, useful idiots are normal users that are not fully aware of the goals of the organization, hence it is extremely difficult to identify them. Like hidden paid posters, useful idiots are hard to distinguish and there is no study that focuses on this task.

- **"True Believers" and Conspiracy Theorists.** Refer to individuals that share false information because they actually believe that they are sharing the truth and that other people need to know about it. For instance, a popular example is Alex Jones, which is a popular conspiracy theorist that shared false information about the Sandy Hook shooting [64].

- **Individuals that benefit from false information.** Refer to various individuals that will have a personal gain by disseminating false information. This is a very broad category ranging from common persons like an owner of a cafeteria to popular individuals like political persons.

- **Trolls [65].** The term troll is used in great extend by the Web community and refers to users that aim to do things to annoy or disrupt other users, usually for their own

personal amusement. An example of their arsenal is posting provocative or off-topic messages in order to disrupt the normal operation or flow of discussion of a website and its users. In the context of false information propagation, we define trolls as users that post controversial information in order to provoke other users or inflict emotional pressure. Traditionally, these actors use fringe communities like Reddit and 4chan to orcherstrate organized operations for disseminating false information to mainstream communities like Twitter, Facebook, and YouTube [66, 19].

Similarly to the types of false information, overlap may exist in actors too. Some examples include: 1) Bots can be exploited by criminal organizations or political persons to disseminate false information [67]; and 2) Hidden paid posters and state-sponsored trolls can be exploited by political persons or organizations to push false information for a particular agenda [2].

3.1.3 Motives behind false information propagation

False information actors and types have different motives behind them. Below we describe the categorization of motives that we distinguish:

- **Malicious Intent.** Refers to a wide spectrum of intents that drive actors that want to hurt others in various ways. Some examples include inflicting damage to the public image of a specific person, organization, or entity.
- **Influence.** This motive refers to the intent of misleading other people in order to influence their decisions, or manipulate public opinion with respect to specific topics. This motive can be distinguished into two general categories; 1) aiming to get leverage or followers (*power*) and 2) changing the norms of the public by disseminating false information. This is particularly worrisome on political matters [68], where individuals share false information to enhance an individuals' public image or to hurt the public image of opposing politicians, especially during election periods.
- **Sow Discord.** In specific time periods individuals or organizations share false information to sow confusion or discord to the public. Such practices can assist in pushing a particular entity's agenda; we have seen some examples on the political stage where foreign governments try to seed confusion in another country's public for their own agenda [69].

- **Profit.** Many actors in the false information ecosystem seek popularity and monetary profit for their organization or website. To achieve this, they usually disseminate false information that increases the traffic on their website. This leads to increased ad revenue that results in monetary profit for the organization or website, at the expense of manipulated users. Some examples include the use of clickbait techniques, as well as fabricated news to increase views of articles from fake news sites that are disseminated via OSNs [48, 70]

- **Passion.** A considerable amount of users are passionate about a specific idea, organization, or entity. This affects their judgment and can contribute to the dissemination of false information. Specifically, passionate users are blinded by their ideology and perceive the false information as correct, and contribute in its overall propagation [71].

- **Fun.** As discussed in the previous section, online trolls are usually diffusing false information for their amusement. Their actions can sometimes inflict considerable damage to other individuals (e.g., see Doxing [22]), and thus should not be taken lightly.

Again, similarly to Sections 3.1.1 and 3.1.2, we have overlap among the presented motives. For instance, a political person may disseminate false information for political influence and because he is passionate about a specific idea.

3.2 User Perception of False Information

In this section, we describe work that study how users perceive and interact with false information on OSNs. Existing work use the following methodologies in understanding how false information is perceived by users: (i) by analyzing large-scale datasets obtained from OSNs; and (ii) by receiving input from users either from questionnaires, interviews, or through crowdsourcing marketplaces (e.g., Amazon Mechanical Turk, AMT [72]). Table 3.1 summarizes the studies on user perception, as well as their methodology and the considered OSN. Furthermore, we annotate each entry in Table 3.1 with the type of false information that each work considers. The remainder of this section provides an overview of the studies on understanding users' perceptions on false information.

Platform	OSN data analysis	Questionnaires/Interviews	Crowdsourcing platforms
Twitter	Kwon et al. [73] **(R)**, Zubiaga et al. [74] **(R)**, Thomson et al. [75] **(R)**	Morris et al. [76] **(CA)**	Ozturk et al. [77] **(R)**, McCreadie et al. [78] **(R)**
Facebook	Zollo et al. [79] **(CT)**, Zollo et al. [80] **(CT)**, Bessi et al. [81] **(CT)**	Marchi [82] **(B)**	X
Other	Dang et al. [83] **(R)**	Chen et al. [84] **(F)**, Kim and Bock [85] **(R)**, Feldman [86] **(B)**, Brewer et al. [87] **(S)** Winerburg and McGrew [88] **(CA)**	X

Table 3.1: Studies of user perception and interaction with false information on OSNs. The table depicts the main methodology of each paper as well as the considered OSN (if any). Also, where applicable, we report the type of false information that is considered (see bold markers and cf. with Section 3.1.1).

3.2.1 OSN data analysis

Previous work focuses on extracting meaningful insights by analyzing data obtained from OSNs. From Table 3.1 we observe that previous work, leverages data analysis techniques to mainly study how users perceive and interact with rumors and conspiracy theories.

Rumors. Kwon et al. [73] study the propagation of rumors in Twitter, while considering findings from social and psychological studies. By analyzing 1.7B tweets, obtained from [89], they find that: 1) users that spread rumors and non-rumors have similar registration age and number of followers; 2) rumors have a clearly different writing style; 3) sentiment in news depends on the topic and not on the credibility of the post; and 4) words related to social relationships are more frequently used in rumors. Zubiaga et al. [74] analyze 4k tweets related to rumors by using journalists to annotate rumors in real time. Their findings indicate that true rumors resolved faster than false rumors and that the general tendency for users is to support every unverified rumor. However, the latter is less prevalent to reputable user accounts (e.g., reputable news outlets) that usually share information with evidence. Thomson et al. [75] study Twitter's activity regarding the Fukushima Daiichi nuclear power plant disaster in Japan. The authors undertake a categorization of the messages according to their user, location, language, type, and credibility of the source. They observe that anonymous users, as well as users that live far away from the disaster share more information from less credible sources. Finally, Dang et al. [83] focus on the users that interact with rumors on Reddit by studying a popular

false rumor (i.e., Obama is a Muslim). Specifically, they distinguish users into three main categories: the ones that support false rumors, the ones that refute false rumors and the ones that joke about a rumor. To identify these users they built a Naive Bayes classifier that achieves an accuracy of 80% and find that more than half of the users joked about this rumor, 25% refuted the joke and only 5% supported this rumor.

Conspiracy Theories. Zollo et al. [79] study the emotional dynamics around conversations regarding science and conspiracy theories. They do so by collecting posts from 280k users on Facebook pages that post either science or conspiracy theories posts. Subsequently, they use Support Vector Machines (SVMs) to identify the sentiment values of the posts, finding that sentiment is more negative on pages with conspiracy theories. Furthermore, they report that as conversations grow larger, the overall negative sentiment in the comments increases. In another work, Zollo et al. [80] perform a quantitative analysis of 54M Facebook users. finding the existence of well-formed communities for the users that interact with science and conspiracy news. They note that users of each community interact within the community and rarely outside of it. Also, debunking posts are rather inefficient and user exposure to such content increases the overall interest in conspiracy theory posts. Similarly, Bessi et al. [81] study how conspiracy theories and news articles are consumed on Facebook, finding that polarized users contribute more in the diffusion of conspiracy theories, whereas this does not apply for news and their respective polarized users.

3.2.2 Questionnaires/Interviews

To get insights on how users perceive the various types of false information, some of the previous work conducted questionnaires or interviews. The majority of the work aims to understand how younger users (students or teenagers) interact and perceive false information.

Credibility Assessment. Morris et al. [76] highlight that users are influenced by several features related to the author of a tweet like their Twitter username when assessing the credibility of information. Winerburg and McGrew [88] study whether users with different backgrounds have differences in their credibility assessments. To achieve this they conducted experiments with historians, fact-checkers, and undergraduate students, finding that historians and students can easily get manipulated by official-looking logos and domain names.

Biased. Marchi [82] focus on how teenagers interact with news on Facebook by conducting interviews with 61 racially diverse teenagers. The main findings of this study is that teenagers

are not very interested in consuming news (despite the fact that their parents do) and that they demonstrate a preference to news that are opinionated when compared to objective news. Similarly, Feldman [86] focus on biased news and conduct 3 different studies with the participants randomly exposed to 2 biased and 1 non-biased news. The participants were asked to provide information about the news that allowed the authors to understand the perceived bias. They find that participants are capable of distinguishing bias in news articles; however, participants perceived lower bias in news that agree with their ideology/viewpoints.

Fabricated. Chen et al. [84] use questionnaires on students from Singapore with the goal to unveil the reasons that users with no malicious intent share false information on OSNs. They highlight that female students are more prone in sharing false information, and that students are willing to share information of any credibility just to initiate conversations or because the content seems interesting.

Rumors. Kim and Bock [85] study the rumor spreading behavior in OSNs from a psychological point of view by undertaking questionnaires on Korean students. They find that users' beliefs results in either positive or negative emotion for the rumor, which affects the attitude and behavior of the users towards the rumor spreading.

Satire. Brewer et al. [87] indicate that satirical news programs can affect users' opinion and political trust, while at the same time users tend to have stronger opinion on matters that they have previously seen in satirical programs.

3.2.3 Crowdsourcing platforms

Other related work leverages crowdsourcing platform to get feedback from users about false information. We note that, to the best of our knowledge, previous work that used crowdsourcing platforms focused on rumors. **Rumors.** Ozturk et al. [77] study how users perceive health-related rumors and if their are willing to share them on Twitter. For acquiring the rumors, they crawl known health-related websites such as Discovery, Food Networks and National Institute of Health websites. To study the user perceptions regarding these rumors, they use AMT where they query 259 participants about ten handpicked health-related rumors. The participants were asked whether they will share a specific rumor or a message that refutes a rumor or a rumor that had a warning on it (i.e., "this message appeared in a rumor website"). Their results indicate that users are less likely to share a rumor that is accompanied with a warning or a message that refutes a rumor. Through simulations, they demonstrate that this

approach can help in mitigating the spread of rumors on Twitter. Finally, McCreadie et al. [78] use crowdsourcing on three Twitter datasets related to emergency situations during 2014, in order to record users' identification of rumors. Their results note that users were able to label most of the tweets correctly, while they note that tweets that contain controversial information are harder to distinguish.

3.2.4 User Perception - Remarks

The studies discussed in this section aim to shed light on how users *perceive* false information on the Web. Overall the main take-away points from the reviewed related work are: 1) teenagers are not interested in consuming news; 2) students share information of any credibility just to initate conversations; 3) in most cases, adults can identify bias in news and this task is harder when the news are biased towards the reader's ideology; and 4) users can mostly identify rumors except the ones that contain controversial information.

3.3 Propagation of False Information

Understanding the dynamics of false information is of paramount importance as it gives useful insights regarding the problem. Table 3.2 summarizes the studies of false information propagation at OSNs, their methodology, as well as the corresponding type of false information according to the typology in Section 3.1.1. The research community focuses on studying the propagation by either employing data analysis techniques or mathematical and statistical approaches. Furthermore, we note the efforts done on providing systems that visualize the propagation dynamics of false information. Below, we describe the studies that are mentioned in Table 3.2 by dedicating a subsection for each type of methodology.

3.3.1 OSN Data Analysis

Rumors. Mendoza et al. [90] study the dissemination of false rumors and confirmed news on Twitter the days following the 2010 earthquake in Chile. They analyze the propagation of tweets for confirmed news and for rumors finding that the propagation of rumors differs from the confirmed news and that an aggregate analysis on the tweets can distinguish the rumors from the confirmed news. Similarly, Starbird et al. [94] study rumors regarding the 2013 Boston Bombings on Twitter and confirm both findings from Mendoza et al. [90]. In a similar

Platform	OSN data analysis	Epidemic & Statistical Modeling	Systems
Twitter	Mendoza et al. [90] **(R)**, Oh et al. [91] **(R)**, Andrews et al. [92] **(R)**, Gupta et al. [93] **(F)**, Starbird et al. [94] **(R)**, Arif et al. [95] **(R)**, Situngkir [96] **(H)**, Nadamoto et al. [97] **(R)**, Vosoughi et al. [98] **(F)**	Jin et al. [99] **(R)**, Doerr et al. [100] **(R)**, Jin et al. [101] **(R)**	Finn et al. [102] **(R)**, Shao el at. [103] **(F)**
Facebook	Friggeri et al. [104] **(R)**, Del Vicario et al. [105] **(CT)**, Anagnostopoulos et al. [106] **(CT)**	Bessi [107] **(CT)**	X
Other	Ma and Li [108] **(R)**, Zannettou et al. [66] **(B)**	Shah et al. [109] **(R)**, Seo et al. [110] **(R)** , Wang et al. [111] **(R)**	Dang et al. [112] **(R)**
Sina Weibo	X	Nguyen et al. [113] **(R)**	X

Table 3.2: Studies the focus on the propagation of false information on OSNs. The table summarizes the main methodology of each paper as well as the considered OSNs. Also, we report the type of false information that is considered (see bold markers and cf. with Section 3.1.1

notion, Nadamoto et al. [97] analyze the behavior of the Twitter community during disasters (Great East Japan Earthquake in 2011) when compared to a normal time period; finding that the spread of rumors during a disaster situation is different from a normal situation. That is in disaster situations, the hierarchy of tweets is shallow whereas in normal situations the tweets follow a deep hierarchy.

Others focused on understanding how rumors can be controlled and shed light on which types of accounts can help stop the rumor spread. Oh et al. [91] study Twitter data about the 2010 Haiti Earthquake and find that credible sources contribute in rumor controlling, while Andrews et al. [92] find that official accounts can contribute in stopping the rumor propagation by actively engaging in conversations related to the rumors.

Arif et al. [95] focus on the 2014 hostage crisis in Sydney. Their analysis include three main perspectives; (i) volume (i.e., number of rumor-related messages per time interval); (ii) exposure (i.e., number of individuals that were exposed to the rumor) and (iii) content production (i.e., if the content is written by the particular user or if it is a share). Their results highlight all three perspectives are important in understanding the dynamics of rumor propagation. Friggeri et al. [104] use known rumors that are obtained through Snopes [114], a popular site that covers rumors, to study the propagation of rumors on Facebook. Their analysis indicates that rumors' popularity is bursty and that a lot of rumors change over time,

thus creating rumor variants. These variants aim to reach a higher popularity burst. Also, they note that rumors re-shares which had a comment containing a link to Snopes had a higher probability to be deleted by their users.

Finally, Ma and Li [108] study the rumor propagation process when considering a two-layer network; one layer is online (e.g., Twitter) and one layer is offline (e.g., face-to-face). Their simulations indicate that rumor spread is more prevalent in a two-layer network when compared with a single-layer offline network. The intuition is that in an offline network the spread is limited by the distance, whereas this constraint is eliminated in a two-layer network that has an online social network. Their evaluation indicates that in a two-layer network the spreading process on one layer does not affect the spreading process of the other layer; mainly because the interlayer transfer rate is less effective from an offline to an online network when compared with that from an OSN.

Fabricated. Gupta et al. [93] study the propagation of false information on Twitter regarding the 2013 Boston Marathon Bombings. To do so, they collect 7.9M unique tweets by using keywords about the event. Using real annotators, they annotate 6% of the whole corpus that represents the 20 most popular tweets during this crisis situation (i.e., the 20 tweets that got retweeted most times). Their analysis indicate that 29% of the tweets were false and a large number of those tweets were disseminated by reputable accounts. This finding contradicts with the findings of Oh et al. [91], which showed that credible accounts help stop the spread of false information, hence highlighting that reputable accounts can share bad information too. Furthermore, they note that out of the 32K accounts that were created during the crisis period, 19% of them were deleted or suspended by Twitter, indicating that accounts were created for the whole purpose of disseminating false information.

Vosoughi et al. [98] study the diffusion of false and true stories in Twitter over the course of 11 years. They find that false stories propagate faster, farther, and more broadly when compared to true stories. By comparing the types of false stories, they find that these effects were more intensive for political false stories when compared to other false stories (e.g., related to terrorism, science, urban legends, etc.).

Hoaxes. Situngkir [96] observe an empirical case in Indonesia to understand the spread of hoaxes on Twitter. Specifically, they focus on a case where a Twitter user with around 100 followers posted a question of whether a well-known individual is dead. Interestingly, the hoax had a large population spread within 2 hours of the initial post and it could be much larger if a popular mainstream medium did not publicly deny the hoax. Their findings indicate

that a hoax can easily spread to the OSN if there is collaboration between the recipients of the hoax. Again, this work highlights, similarly to Oh et al. [91] that reputable accounts can help in mitigating the spread of false information.

Conspiracy Theories. Del Vicario et al. [105] analyze the cascade dynamics of users on Facebook when they are exposed to conspiracy theories and scientific articles. They analyze the content of 67 public pages on Facebook that disseminate conspiracy theories and science news. Their analysis indicates the formulation of two polarized and homogeneous communities for each type of information. Also, they note that despite the fact that both communities have similar content consumption patterns, they have different cascade dynamics. Anagnostopoulos et al. [106] study the role of homophily and polarization on the spread of false information by analyzing 1.2M Facebook users that interacted with science and conspiracy theories. Their findings indicate that user's interactions with the articles correlate with the interactions of their friends (homophily) and that frequent exposure to conspiracy theories (polarization) determines how viral the false information is in the OSN.

Biased. Zannettou et al. [66], motivated by the fact that the information ecosystem consists of multiple Web communities, study the propagation of news across multiple Web communities. To achieve this, they study URLs from 99 mainstream and alternative news sources on three popular Web communities: Reddit, Twitter, and 4chan. Furthermore, they set out to measure the influence that each Web community has to each other, using a statistical model called Hawkes Processes. Their findings indicate that small fringe communities within Reddit and 4chan have a substantial influence to mainstream OSNs like Twitter.

3.3.2 Epidemic and Statistical Modeling

Rumors. Jin et al. [99] use epidemiological models to characterize cascades of news and rumors in Twitter. Specifically, they use the SEIZ model [115] which divides the user population in four different classes based on their status; (i) Susceptible; (ii) Exposed; (iii) Infected and (iv) Skeptic. Their evaluation indicates that the SEIZ model is better than other models and it can be used to distinguish rumors from news in Twitter. In their subsequent work, Jin et al. [101] perform a quantitative analysis on Twitter during the Ebola crisis in 2014. By leveraging the SEIZ model, they show that rumors spread in Twitter the same way as legitimate news.

Doerr et al. [100] use a mathematical approach to prove that rumors spread fast in OSNs

(similar finding with Vosoughi et al. [98]). For their simulations they used real networks that represent the Twitter and Orkut Social Networks topologies obtained from [89] and SNAP [116], respectively. Intuitively, rumors spread fast because of the combinations of few large-degree nodes and a large number of small-degree nodes. That is, small-degree nodes learn a rumor once one of their adjacent nodes knows it, and then quickly forward the rumor to all adjacent nodes. Also, the propagation allows the diffusion of rumors between 2 large-degree nodes, thus the rapid spread of the rumor in the network.

Several related work focus on finding the source of the rumor. Specifically, Shah et al. [109] focus on detecting the source of the rumor in a network by defining a new rumor spreading model and by forming the problem as a maximum likelihood estimation problem. Furthermore, they introduce a new metric, called *rumor centrality*, which essentially specifies the likelihood that a particular node is the source of the rumor. This metric is evaluated for all nodes in the network by using a simple linear time message-passing algorithm, hence the source of the rumor can be found by selecting the node with the highest rumor centrality. In their evaluation, they used synthetic small-world and scale-free real networks to apply their rumor spreading model and they show that they can distinguish the source of a rumor with a maximum error of 7-hops for general networks, and with a maximum error of 4-hops for tree networks. Seo et al. [110] aim to tackle the same problem by injecting monitoring nodes on the social graph. They propose an algorithm that considers the information received by the monitoring nodes to identify the source. They indicate that with sufficient number of monitoring nodes they can recognize the source with high accuracy. Wang et al. [111] aim to tackle the problem from a statistical point of view. They propose a new detection framework based on rumor centrality, which is able to support multiple snapshots of the network during the rumor spreading. Their evaluation based on small-world and scale-free real networks note that by using two snapshots of the network, instead of one, can improve the source detection. Finally, Nguyen et al. [113] aim to find the k most suspected users where a rumor originates by proposing the use of a reverse diffusion process in conjunction with a ranking process.

Conspiracy Theories. Bessi [107] perform a statistical analysis of a large corpus (354k posts) of conspiracy theories obtained from Facebook pages. Their analysis is based on the Extreme Value Theory branch of statistics [117] and they find that extremely viral posts (greater than 250k shares) follow a Poisson distribution.

3.3.3 Systems

Rumors. Finn et al. [102] propose a web-based tool, called TwitterTrails, which enables users to study the propagation of rumors in Twitter. TwitterTrails demonstrates indications for bursty activity, temporal characteristics of propagation, and visualizations of the re-tweet networks. Furthermore, it offers advanced metrics for rumors such as level of visibility and community's skepticism towards the rumor (based on the theory of h-index [118]). Similarly, Dang et al. [112] propose RumourFlow, which visualizes rumors propagation by adopting modeling and visualization tools. It encompasses various analytical tools like semantic analysis and similarity to assist the user in getting a holistic view of the rumor spreading and its various aspects. To demonstrate their system, they collect rumors from Snopes and conversations from Reddit.

Fabricated. Shao et al. [103] propose Hoaxy, a platform that provides information about the dynamics of false information propagation on Twitter as well as the respective fact checking efforts.

3.3.4 Propagation of False Information - Remarks

In this section, we provided an overview of the existing work that focuses on the propagation of false information on the Web. Some of the main take-aways from the literature review on the propagation of false information are: 1) Accounts on social networks are created with the sole purpose of disseminating false information; 2) False information is more persistent than corrections; 3) The popularity of false information follow a bursty activity; 4) Users on Web communities create polarized communities that disseminate false information; 5) Reputable or credible accounts are usually useful in stopping the spread of false information; however we need to pay particular attention as previous work (see Gupta et al. [93]) has showed that they also share false information; 6) Being able to detect the source of false information is a first step towards stopping the spread of false information on Web communities and several approaches exist that offer acceptable performance.

3.4 Detection and Containment of False Information

3.4.1 Detection of false information

Detecting false information is not a straightforward task, as it appears in various forms, as discussed in Section 3.1. Table 3.3 summarizes the studies that aim to solve the false information detection problem, as well as their considered OSNs and their methodology. Most studies try to solve the problem using handcrafted features and conventional machine learning techniques. Recently, to avoid using handcrafted features, the research community used neural networks to solve the problem (i.e., Deep Learning techniques). Furthermore, we report some systems that aim to inform users about detected false information. Finally, we also note a variety of techniques that are proposed for the detection and containment of false information, such as epidemiological models, multivariate Hawkes processes, and clustering. Below, we provide more details about existing work grouped by methodology and the type of information, according to Table 3.3.

Machine Learning

Credibility Assessment. Previous work leverage machine learning techniques to assess the credibility of information. Specifically, Castillo et al. [119] analyze 2.5k trending topics from Twitter during 2010 to determine the credibility of information. For labeling their data they utilize crowdsourcing tools, namely AMT, and propose the use of conventional machine learning techniques (SVM, Decision Trees, Decision Rules, and Bayes Networks) that take into account message-based, user-based, topic-based and propagation-based features. Gupta and Kumaraguru [120] analyze tweets about fourteen high impact news events during 2011. They propose the use of supervised machine learning techniques with a relevance feedback approach that aims to rank the tweets according to their credibility score. AlRubaian et al. [126] propose the use of a multi-stage credibility assessment platform that consists of a relative importance component, a classification component, and an opinion mining component. The relative importance component requires human experts and its main objective is to rank the features according to their importance. The classification component is based on a Naive Bayes classifier, which is responsible for classifying tweets by taking the output of the relative importance component (ranked features), while the opinion mining component captures the sentiment of the users that interact with the tweets. The output of the three components is then

Platform	Machine Learning	Systems	Other models/algorithms
Twitter	Castillo et al. [119] **(CA)**, Gupta and Kumaraguru [120] **(CA)**, Kwon et al. [121] **(R)**, Yang et al. [122] **(R)**, Liu et al. [123] **(R)**, Wu et al. [124] **(R)**, Gupta et al. [125] **(CA)**, AlRubaian et al. [126] **(CA)**, Hamidian and Diab [127] **(R)**, Giasemidis et al. [128] **(R)** , Kwon et al. [129] **(R)**, Volkova et al. [130] **(CA)**	Resnick et al. [131] **(R)**, Vosoughi et al. [132] **(R)**, Jaho et al. [133] **(CA)**	Qazvinian et al. [134] **(R)** (rumor retrieval model), Zhao el al. [135] **(R)** (clustering), Farajtabar et al. [136] **(F)** (hawkes process), Kumar and Geethakumari [137] **(F)** (algorithm with psychological cues)
Sina Weibo	Yang et al. [138] **(R)**, Wu et al. [139] **(R)**, Liang et al. [140] **(R)**, Zhang et al. [141] **(R)**,	Zhou et al. [142] **(CA)**	X
Twitter and Sina Weibo	Ma et al. [143] **(CA)** Ma et al. [144] **(R)**	X	Jin et al. [145] **(CA)** (graph optimization)
Facebook	Tacchini et al. [146] **(H)**, Conti et al. [147] **(CT)**	X	X
Wikipedia and/or other articles	Qin et al. [148] **(R)**, Rubin et al. [149] **(S)**, Kumar et al. [38] **(H)**, Chen et al. [46] **(CL)**, Chakraborty et al. [150] **(CL)**, Potthast et al. [151] **(CL)**, Biyani et al. [152] **(CL)**, Wang [153] **(F)**, Anand et al. [154] **(CL)**	X	Potthast et al. [42] **(B)** (unmasking)
Other	Afroz et al. [155] **(H)**, Maigrot et al. [156] **(H)**, Zannettou et al. [157] **(CL)**	Vukovic et al. [158] **(H)**	Jin et al. [159] **(CA)** (hierarchical propagation model), Chen et al. [160] **(H)** (Levenshtein Distance)

Table 3.3: Studies that focus on the detection of false information on OSNs. The table demonstrates the main methodology of each study, as well as the considered OSNs. Also, we report the type of false information that is considered (see bold markers and cf. with Section 3.1.1, **CA** corresponds to Credibility Assessment and refers to work that aim to assess the credibility of information).

combined to calculate an overall assessment. Ma et al. [143] observe that typically the features of messages in microblogs vary over time and propose the use of an SVM classifier that is able to consider the messages features in conjunction with how they vary over time. Their experimental evaluation, based on Twitter data provided by [119] and on a Sina Weibo dataset, indicate that the inclusion of the time-varying features increase the performance between 3% and 10%.

All of the aforementioned work propose the use of supervised machine learning techniques. In contrast, Gupta et al. [125] propose a semi-supervised model that ranks tweets according to their credibility in real-time. For training their model, they collect 10M tweets from six incidents during 2013, while they leverage CrowdFlower [161] to obtain groundtruth. Their

system also includes a browser extension that was used by approx. 1.1k users in a 3-month timespan, hence computing the credibility score of 5.4M tweets. Their evaluation indicates that 99% of the users were able to receive credibility scores under 6 seconds. However, feedback from users for approx. 1.2k tweets indicate that 60% of the users disagreed with the predicted score.

Volkova et al. [130] motivated by the performance gains of deep learning techniques, propose the use of neural networks to distinguish news into satire, hoaxes, clickbait, and propaganda news. They collect 130k news posts from Twitter and propose the use of neural networks that use linguistic and network features. Their findings indicate that Recurrent and Convolutional neural networks exhibit strong performance in distinguishing news in the aforementioned categories.

Rumors. Kwon et al. [121] propose the use of Decision Trees, Random Forest, and SVM for detecting rumors on Twitter. Their models leverage temporal, linguistics, and structural features from tweets and can achieve precision and recall scores between 87% and 92%. Yang et al. [122] propose the use of a hot topic detection mechanism that work in synergy with conventional machine learning techniques (Naive Bayes, Logistic Regression and Random Forest). Liu et al. [123] demonstrate the feasibility of a real-time rumoring detection system on Twitter. To achieve real-time debunking of rumors, they propose the use of an SVM classifier that uses beliefs from the users in conjunction with traditional rumor features from [119, 138]. Their evaluation demonstrates that for new rumors (5-400 tweets), the proposed classifier can outperform the models from [119, 138]. Furthermore, they compare their approach with human-based rumor debunking services (Snopes and Emergent), showing that they can debunk 75% of the rumors earlier than the corresponding services. Similarly, Kwon et al. [129] study the rumor classification task with a particular focus on the temporal aspect of the problem, by studying the task over varying time windows on Twitter. By considering user, structural, linguistic, and temporal features, they highlight that depending on the time window, different characteristics are more important than others. For example, at early stages of the rumor propagation, temporal and structural are not available. To this end, they propose a rumor classification algorithm that achieves satisfactory accuracy both on short and long time windows.

Hamidian and Diab [127] propose a supevised model that is based on the Tweet Latent Vector (TLV), which is an 100-dimensional vector, proposed by the authors, that encapsulates the semantics behind a particular tweet. For the classification task, they use an SVM Tree Kernel

model that achieves 97% on two Twitter datasets. Giasemidis et al. [128] study 72 rumors in Twitter by identifying 80 features for classifying false and true rumors. These features include diffusion and temporal dynamics, linguistics, as well as user-related features. For classifying tweets, they use several machine learning techniques and conclude that Decision Trees achieve the best performance with an accuracy of 96%. Yang et al. [138] study the rumor detection problem in the Sina Weibo OSN. For the automatic classification task of the posts they use SVMs that take as input various features ranging from content-based to user- and location-based features. Their evaluation shows that the classifier achieves an accuracy of approximately 78%. Similarly to the aforementioned work, Wu et al. [139] try to tackle the rumor detection problem in the Sina Weibo OSN by leveraging SVMs. Specifically, they propose an SVM classifier which is able to combine a normal radial basis function, which captures high level semantic features, and a random walk graph kernel, which captures the similarities between propagation trees. These trees encompass various details such as temporal behavior, sentiment of re-posts, and user details. Liang et al. [140] study the problem of rumor detection using machine learning solutions that take into account users' behavior in the Sina Weibo OSN. Specifically, they introduce 3 new features that are shown to provide up to 20% improvement when compared with baselines. These features are: 1) average number of followees per day; 2) average number of posts per day; and 3) number of possible microblog sources. Zhang et al. [141] propose various implicit features that can assist in the detection of rumors. Specifically, they evaluate an SVM classifier against the Sina Weibo dataset proposed in [138] with the following features: 1) content-based implicit features (sentiment polarity, opinion on comments and content popularity); 2) user-based implicit features (influence of user to network, opinion re-tweet influence, and match degree of messages) and 3) shallow message features that are proposed by the literature. Their evaluation shows that the proposed sets of features can improve the precision and recall of the system by 7.1% and 6.3%, respectively. Qin et al. [148] propose the use of a new set of features for detecting rumors that aim to increase the detection accuracy; namely novelty-based and pseudo-feedback features. The novelty-based features consider reliable news to find how similar is a particular rumor with reliable stories. The pseudo-feedback features take into account information from historical confirmed rumors to find similarities. To evaluate their approach, they obtain messages from the Sina Weibo OSN and news articles from Xinhua News Agency [162]. They compare an SVM classifier, which encompasses the aforementioned set of features and a set of other features (proposed by the literature), with the approaches proposed by [138, 123]. Their findings indicate that their approach provides an improvement between 17% and 20% in terms

of accuracy. Similarly to [148], Wu et al. [124] propose a system that uses historical data about rumors for the detection task. Their system consists of a feature selection module, which categorizes and selects features, and a classifier. For constructing their dataset they use Snopes and the Twitter API to retrieve relevant tweets, acquiring in total 10k tweets, which are manually verified by annotators. In their evaluation, they compare their system with various baselines finding that the proposed system offers enhanced performance in rumor detection with an increase of 12%-24% for precision, recall, and F1-score metrics. Ma et al. [144] leverage Recurrent neural networks to solve the problem of rumor detection in OSNs. Such techniques are able to learn hidden representations of the input without the need for hand-crafted features. For evaluating their model, they construct two datasets; one from Twitter and one from Sina Weibo. For the labeling of their messages they use Snopes for Twitter and the official rumor-busting service of Sina Weibo's OSN. Their evaluation shows an accuracy of 91% on the Sina Weibo dataset and 88% on the Twitter dataset.

Hoaxes. Tacchini et al. [146] study hoaxes in Facebook and argue that they can accurately discern hoax from non-hoax posts by simply looking at the users that liked the posts. Specifically, they propose the use of Logistic Regression that classifies posts with features based on users' interactions. Their evaluation demonstrate that they can identify hoaxes with an accuracy of 99%. Kumar et al. [38] study the presence of hoaxes in Wikipedia articles by considering 20k hoax articles that are explicitly flagged by Wikipedia editors. They find that most hoaxes are detected quickly and have little impact, however, a small portion of these hoaxes have a significant life-span and are referenced a lot across the Web. By comparing the "successful" hoaxes with failed hoaxes and legitimate articles, the authors highlight that the successful hoaxes have notable differences in terms of structure and content. To this end, they propose the use of a Random Forest classifier to distinguish if articles are hoaxes. Their evaluation reports that their approach achieves an accuracy of 92% and that is able to outperform human judgments by a significant margin (20%). Maigrot et al. [156] propose the use of a multi-modal hoax detection system that fuses the diverse modalities pertaining to a hoax. Specifically, they take into consideration the text, the source, and the image of tweets. They observe higher performance when using only the source or text modality instead of the combination of all modalities.

Conspiracy Theories. Conti et al. [147] focus on identifying conspiracy theories in OSNs by considering only the structural features of the information cascade. The rationale is that such features are difficult to be tampered by malicious users, which aim to avoid detection from classification systems. For their dataset they use data from [81], which consist of scientific

articles and conspiracy theories. For classifying their Facebook data they propose conventional machine learning techniques and they find that it is hard to distinguish a conspiracy theory from a scientific article by only looking at their structural dynamics (F1 -score not exceeding 65%).

Satire. Rubin et al. [149] propose the use of satirical cues for the detection of false information on news articles. Specifically, they propose the use of five new set of features, namely absurdity, humor, grammar, negative affect, and punctuation. Their evaluation shows that by using an SVM algorithm with the aforementioned set of features and others proposed by the literature, they can detect satirical news with 90% precision and 84% recall.

Clickbait. Several studies focus on the detection of clickbait on the Web using machine learning techniques. Specifically, Chen et al. [46] propose tackling the problem using SVMs and Naive Bayes. Also, Chakraborty et al. [150] propose the use of SVM and a browser add-on to offer a system to users for news articles. Potthast et al. [151] proposes the use of Random Forest for detecting clickbait tweets. Moreover, Biyani et al. [152] propose the use of Gradient Boosted Decision Trees for clickbait detection in news articles and show that the degree of informality in the content of the landing page can help in finding clickbait articles. Anand et al. [154] is the first work that suggests the use of deep learning techniques for mitigating the clickbait problem. Specifically, they propose the use of Recurrent Neural Networks in conjunction with word2vec embeddings [12] for identifying clickbait news articles. Similarly, Zannettou et al. [157] use deep learning techniques to detect clickbaits on YouTube. Specifically, they propose a semi-supervised model based on variational autoencoders (deep learning). Their evaluation indicates that they can detect clickbaits with satisfactory performance and that YouTube's recommendation engine does not consider clickbait videos in its recommendations.

Fabricated. Wang [153] presents a dataset that consists of 12.8k manually annotated short statements obtained from PolitiFact. They propose the use of Convolutional neural networks for fusing linguistic features with metadata (e.g., who is the author of the statement). Their evaluation demonstrates that the proposed model outperforms SVM and Logistic Regression algorithms.

Systems

Rumors. Resnick et al. [131] propose a system called RumorLens, which aims to discover rumors in a timely manner, provide insights regarding the rumor's validity, and visualize

a rumor's propagation. To achieve the aforementioned, RumorLens leverages data mining techniques alongside with a visual analysis tool. However, their system raises scalability issues as it highly depends on users' labor, which provide labeling of tweets that are subsequently used for classifying tweets related to a particular rumor. Vosoughi et al. [132] propose a human-machine collaborative system that aims to identify rumors by disposing irrelevant data and ranking the relevant data. Their system consists of two components; the assertion detector and the hierarchical clustering module. The assertion detector is a classifier that uses semantic and syntactic features to find tweets that contain assertions. These tweets are then presented to the clustering module, which clusters the tweets according to the similarity of the assertions. During their evaluation, the authors state that for a particular incident (Boston Marathon Bombings) from a dataset of 20M tweets, their system managed to discard 50% of them using the assertion detector. Furthermore, the 10M relevant tweets are clustered somewhere between 100 and 1000 clusters, something that enables users to quickly search and find useful information easier.

Credibility Assessment. Jaho et al. [133] undertake a statistical analysis by crawling Twitter for 3 months and retrieve a dataset that includes 10M users. They propose a system that is based on contributor-related features (e.g., reputation, influence of source, etc.), content features (e.g., popularity, authenticity, etc.) and context features (e.g., coherence, cross-checking, etc.). Their system combines all the features and outputs a single metric that corresponds to the truthfulness of the message. Zhou et al. [142] note that calculating credibility in the granularity of message is not scalable, therefore they propose the calculation of credibility score per event. To this end, they propose a system that is able to collect related data from Sina Weibo using keywords and detect the credibility of a particular event. The credibility score is calculated by the combination of 3 sub-models; the user model, the propagation model, and the content model. Each one of the sub-models considers one aspect of the news and the overall score is calculated using weighted combination. The system is trained on a dataset that contains 73 real news and 73 fake news from approximately 50k posts. Their evaluation shows that the proposed system provides an accuracy close to 80% and that credibility scores are calculated within 35 seconds.

Hoaxes. Vukovic et al. [158] focus on hoaxes and propose the use of a detection system for email. The proposed system consists of a feed-forward neural network and a self-organizing map (SOM) and it is trained on a corpus of 298 hoax and 1370 regular emails. The system achieves an accuracy of 73% with a ratio of false positives equal to 4.9%. Afroz et al. [155] focus on detecting hoaxes by observing changes in writing style. The intuition is that people

use different linguistic features when they try to obfuscate or change information from users. To detect hoaxes they propose the use of an SVM classifier that takes into account the following set of features: 1) lexical features; 2) syntactic features; 3) content features and 4) lying detection features obtained from [163, 164]. Their evaluation on various datasets indicates that the proposed system can detect hoaxes with an accuracy of 96%.

Other models/algorithms

Rumors. Qazvinian et al. [134] study the rumor detection problem on Twitter by retrieving tweets regarding rumors and leveraging manual inspectors to annotate it. Specifically, the annotators were asked whether tweets contained rumors or not and whether a user endorsed, debunked or was neutral about the rumors. The resulted dataset consists of approximately 10k annotated tweets and was analyzed to demonstrate the effectiveness of the following feature sets in identifying rumors: 1) content-based features; 2) network-based features and 3) Twitter-specific memes (hashtags and URLs). Furthermore, the paper proposes a rumor retrieval model that achieves 95% precision. Zhao et al. [135] are motivated by the fact that identifying false factual claims in each individual message is intractable. To overcome this, they adapt the problem in finding whole clusters of messages that their topic is a disputed factual claim. To do so, they search within posts to find specific phrases that are used from users who want to seek more information or to express their skepticism. For example, some enquiry phrases are ”Is this true?”, ”Really?” and ”What?”. Their approach uses statistical features of the clusters in order to rank them according to the likelihood of including a disputed claim. Their evaluations on real Twitter data indicate that among the top 50 ranked clusters, 30% of them are confirmed rumors.

Fabricated. Farajtabar et al. [136] propose a framework for tackling false information that combines a multivariate Hawkes process and reinforcement learning. Their evaluation highlights that their model shows promising performance in identifying false information in real-time on Twitter. Kumar and Geethakumari [137] measure the diffusion of false information by exploiting cues obtained from cognitive psychology. Specifically, they consider the consistency of the message, the coherency of the message, the credibility of the source, and the general acceptability of the content of the message. These cues are fused to an algorithm that aims to detect the spread of false information as soon as possible. Their analysis on Twitter reports that the proposed algorithm has a 90% True positive rate and a False positive rate less than 10%.

Credibility Assessment. Jin et al. [145] aim to provide verification of news by considering conflicting viewpoints on Twitter and Sina Weibo. To achieve this, they propose the use of a topic model method that identifies conflicting viewpoints. Subsequently they construct a credibility network with all the viewpoints and they formulate the problem as a graph optimization problem, which can be solved with an iterative approach. They compare their approach with baselines proposed in [119, 121], showing that their solution performs better. Jin et al. [159] propose a hierarchical propagation model to evaluate information credibility in microblogs by detecting events, sub-events, and messages. This three-layer network assists in revealing vital information regarding information credibility. By forming the problem as a graph optimization problem, they propose an iterative algorithm, that boosts the accuracy by 6% when compared to an SVM classifier that takes into account only features obtained from the event-level network only.

Biased. Potthast et al. [42] study the writing style of hyperpartisan news (left-wing and right-wing) and mainstream news and how this style can be applied in hyperpartisan news detection. Their dataset consists of 1.6k news articles from three right-wing, three left-wings, and three mainstream news sites. For annotating the dataset they used journalists from Buzzfeed, who rated each article according to its truthfulness. By leveraging the Unmasking approach [165], the paper demonstrates that right-wing and left-wing hyperpartisan news exhibit similar writing style that differentiates from the mainstream news. To this end, they propose the use of Random Forest classifier that aims to distinguish hyperpartisanship. Their evaluation indicates that their style-based classifier can distinguish hyperpartisan news with an accuracy of 75%. However, when the same classifier is used to discern fake or real news, then the accuracy is 55%.

Hoaxes. Chen et al. [160] propose an email hoax detection system by incorporating a text matching method using the Levenshtein distance measure. Specifically, their system maintains a database of hoaxes that is used to calculate the distance between a potential hoax email and the stored hoaxes.

3.4.2 Containment of false information

Several studies focus on containing the diffusion of false information. Our literature review reveals that the majority of previous work on containment of rumors, while we also find one that focus on Hoaxes (see Tambuscio et al. [166]). Below we provide a brief overview of the studies that try to contain the spread of false information, while ensuring that the solutions are

scalable.

Rumors. Tripathy et al. [167] propose a process, called "anti-rumor", which aims to mitigate the spreading of a rumor in a network. This process involves the dissemination of messages, which contradict with a rumor, from agents. The authors make the assumption that once a user receives an anti-rumor message, then he will never believe again the rumor, thus the spreading of a rumor is mitigated. Their evaluation, based on simulations, indicates the efficacy of the proposed approach. Budak et al. [168] formulate the problem of false information spreading as an optimization problem. Their aim is to identify a subset of the users that need to be convinced to spread legitimate messages in contrast with the bad ones that spread rumors. The paper shows that this problem is NP-hard and they propose a greedy solution as well as some heuristics to cope with scalability issues. Fan et al. [169] try to tackle the problem of false information propagation under the assumption that rumors originate from a particular community in the network. Similarly to other work, the paper tries to find a minimum set of individuals, which are neighbors with the rumor community to stop the rumor diffusion in the rest of the network. To achieve this, they propose the use of two greedy-based algorithms, which are evaluated in two real-world networks (Arxiv Hep and Enron). Their experimental results show that the proposed algorithms outperform simple heuristics in terms of the number of infected nodes in the network. However, as noted, the greedy algorithms are time consuming and are not applicable in large-scale networks. Kotnis et al. [170] propose a solution for stopping the spread of false information by training a set of individuals in a network that aim to distinguish and stop the propagation of rumors. This set of individuals is selected based on their degree in the network with the goal to minimize the overarching training costs. For evaluating their solution they create a synthetic network, which takes into account a calculated network degree distribution, based on [171]. Ping et al. [172] leverage Twitter data to demonstrate that sybils presence in OSNs can decrease the effectiveness of community-based rumor blocking approaches by 30%. To this end, they propose a Sybil-aware rumor blocking approach, which finds a subset of nodes to block by considering the network structure in conjunction with the probabilities of nodes being sybils. Their evaluation, via simulations on Twitter data, show that the proposed approach significantly decreases the number of affected nodes, when compared to existing approaches. He et al. [173] argue that existing false information containment approaches have different costs and efficiencies in different OSNs. To this end, they propose an optimization method that combines the spreading of anti-rumors and the block of rumors from influential users. The goal of their approach is to minimize the overarching cost of the method while containing the rumor within an expected

deadline. To achieve this, they use the Pontryagin's maximum principle [174] on the Digg2009 dataset [175]. They find that spreading the truth plays a significant role at the start of the rumor propagation, whereas close to the deadline of containment the blocking of rumors approach should be used extensively. Huang et al. [176] aim to contain the false information spread by finding and decontaminating with good information, the smallest set of influential users in a network. To do so, they propose a greedy algorithm and a community-based heuristic, which takes into consideration the community structure of the underlying network. For evaluating their approach, they used traces from three networks; NetHEPT, NetHEPT_WC and Facebook. Previous studies on false information containment [168, 177] assumed that when true and false information arrive the same time at a particular node, then the true information dominates. Wang et al. [178] state that the dominance of the information should be based on the influence of the neighbors in the network. With this problem formulation in mind, the paper proposes two approaches to find the smallest number of nodes that are required to stop the false information spread. Their evaluation is based on three networks obtained from Twitter, Friendster, and a random synthetic network. Evaluation comparisons with simple heuristics (random and high degree) demonstrate the performance benefits of the proposed approaches. In a similar notion, Tong et al. [179] aim to increase performance motivated by the fact that greedy solutions, which include Monte Carlo simulations, are inefficient as they are computationally intensive. To overcome this, the paper proposes a random-based approach, which utilizes sampling with the aim to be both effective and efficient. The performance evaluations on real-world (obtained from Wikipedia and Epinions [180]) and synthetic networks demonstrate that the proposed solution can provide a 10x speed-up without compromising performance when compared to state-of-the-art approaches. Wang et al. [181] propose a model, called DRIMUX, which aims to minimize the influence of rumors by blocking a subset of nodes while considering users' experience. User experience is defined as a time threshold that a particular node is willing to wait while being blocked. Their model utilizes survival theory and takes into account global rumor popularity features, individual tendencies (how likely is a rumor to propagate between a pair of nodes) as well as the users' experience. Their evaluations on a Sina Weibo network, which consists of 23k nodes and 183k edges, indicate that the proposed model can reduce the overarching influence of false information.

Hoaxes. Tambuscio et al. [166] simulate the spread and debunking of hoaxes on networks. Specifically, they model the problem as a competition between believers (acknowledge the hoax) and fact checkers which reveal the hoax with a specific probability. To study their model they performed simulations on scale-free and random networks finding that a specific threshold

for the probability of fact checkers exists and this indicates that the spread can be stopped with a specific number of fact checkers. However, the paper oversimplifies the problem by assuming all the nodes to have the same probability.

3.4.3 Detection and Containment of False Information - Remarks

The main findings from the literature review of the detection and containment of false information are: 1) Machine learning techniques can assist in identifying false information. However, they heavily rely on handcrafted set of features and it is unclear if they generalize well on other datasets; 2) Containment of false information can be achieved by adding a set of good nodes that disseminate good information or information that refute false; and 3) The problem of detection of false information requires human-machine collaboration for effectively mitigating it.

3.5 False Information in the political stage

Recently, after the 2016 US elections, the problem of false information dissemination got extensive interest from the community. Specifically, Facebook got openly accused for disseminating false information and that affected the outcome of the elections [182]. It is evident that dissemination of false information on the Web is used a lot for political influence. Therefore in this section we review the most relevant studies on the political stage. Table 3.4 reports the reviewed work as well as the main methodology and considered OSN.

3.5.1 Machine Learning

Propaganda. Ratkiewicz et al. [183] study political campaigns on Twitter that use multiple controlled accounts to disseminate support for an individual or opinion. They propose the use of a machine learning-based framework in order to detect the early stages of the spreading of political false information on Twitter. Specifically, they propose a framework that takes into consideration topological, content-based and crowdsourced features of the information diffusion in Twitter. Their experimental evaluation demonstrates that the proposed framework achieves more than 96% accuracy in the detection of political campaigns for data pertaining to the 2010 US midterm elections. Conover et al. [184] study Twitter on a six-week period leading to the 2010 US midterm elections and the interactions between right and left leaning

Platform	Machine Learning	OSN Data Analysis	Other models/algorithms
Twitter	Ratkiewicz et al. [183] **(P)**, Conover et al.[184] **(P)**, Ferrara et al.[185] **(P)**	Wong et al. [186] **(B)**, Golbeck and Hansen [187] **(B)**, Jackson and Welles [188] **(P)**, Hegelich and Janetzko[189] **(P)**, Zannettou et al. [60] **(P)** Howard and Kollanyi[190] **(P)**, Shin et al.[191] **(R)**	An et al. [192] **(B)** (distance model), Al-khateeb and Agarwal [53] **(P)** (social studies) Ranganath et al.[193] **(P)** (exhaustive search), Jin et al. [194] **(R)** (text similarity), Yang et al. [195] **(B)** (agenda-setting tool)
Digg	Zhou et al.[196] **(B)**	X	X
Sina Weibo	X	King et al. [197] **(P)**, Yang et al. [198] **(P)**	X
News articles	Budak et al. [199] **(B)**	Woolley[200] **(P)**	X
Facebook	X	Allcot and Gentzkow[55] **(P)**	X

Table 3.4: Studies on the false information ecosystem on the political stage. The table demonstrates the main methodology of each study as well as the considered OSNs.

communities. They leverage clustering algorithms and manually annotated data to create the re-tweets and mentions networks. Their findings indicate that the re-tweet network has limited connectivity between the right and left leaning communities, whereas this is not the case in the mentions networks. This is because, users try to inject different opinions on users with different ideologies, by using mentions on tweets, so that they change their stance towards a political individual or situation. Ferrara et al. [185] propose the use of a k-nearest neighbor algorithm with a dynamic warping classifier in order to capture promoted campaigns in Twitter. By extracting a variety of features (user-related, timing-related, content-related and sentiment-related features) from a large corpus of tweets they demonstrate that they can distinguish promoted campaigns with an AUC score close to 95% in a timely manner.

Biased. Zhou et al. [196] study Digg, a news aggregator site, and aim to classify users and articles to either liberal or conservative. To achieve this, they propose three semi-supervised propagation algorithms that classify users and articles based on users' votes. The algorithms make use of a few labeled users and articles to predict a large corpus of unlabeled users and articles. The algorithms are based on the assumption that a liberal user is more likely to vote for a liberal article rather than a conservative article. Their evaluations demonstrate that the best algorithm achieves 99% and 96% accuracy on the dataset of users and articles, respectively. Budak et al. [199] use Logistic Regression to identify articles regarding politics from a large corpus of 803K articles obtained from 15 major US news outlets. Their algorithm filtered out

86% of the articles as non-political related, while a small subset of the remainder (approx. 11%) were presented to workers on AMT. The workers were asked to answer questions regarding the topic of the article, whether the article was descriptive or opinionated, the level of partisanship, and the level of bias towards democrats or republicans. Their empirical findings are that on these articles there are no clear indications of partisanship, some articles within the same outlet are left-leaning and some have right-leaning, hence reducing the overall outlet bias. Also, they note that usually bias in news articles is expressed by criticizing the opposed party rather than promoting the supporting party.

3.5.2 OSN Data Analysis

Biased. Wong et al. [186] collect and analyze 119M tweets pertaining to the 2012 US presidential election to quantify political leaning of users and news outlets. By formulating the problem as an ill-posed linear inverse problem, they propose an inference engine that considers tweeting behavior of articles. Having demonstrated their inference engine, the authors report results for the political leaning scores of news sources and users on Twitter. Golbeck and Hansen [187] provide a technique to estimate audience preferences in a given domain on Twitter, with a particular focus on political preferences. Different from methods that assess audience preference based on citation networks of news sources as a proxy, they directly measure the audience itself via their social network. Their technique is composed of three steps: 1) apply ground truth scores (they used Americans for Democratic Action reports as well as DW-Nominate scores) to a set of seed nodes in the network, 2) map these scores to the seed group's followers to create "P-scores", and 3) map the P-scores to the target of interest (e.g., government agencies or think tanks). One important take away from this work is that *Republicans are over-represented on Twitter with respect to their representation in Congress*, at least during the 2012 election cycle. To deal with this, they built a balanced dataset by randomly sampling from bins formed by the number of followers a seed group account had.

Propaganda. Jackson and Welles [188] demonstrate how Twitter can be exploited to organize and promote counter narratives. To do so, they investigate the misuse of a Twitter hashtag (#myNYPD) during the 2014 New York City Police Department public relations campaign. In this campaign, this hashtag was greatly disseminated to promote counter narratives about racism and police misconduct. The authors leverage network and qualitative discourse analysis to study the structure and strategies used for promoting counterpublic narratives.

Hegelich and Janetzko [189] investigate whether bots on Twitter are used as political actors. By

exposing and analyzing 1.7K bots on Twitter, during the Russian/Ukrainian conflict, they find that the botnet has a political agenda and that bots exhibit various behaviors. Specifically, they find that bots try to hide their identity, to be interesting by promoting topics through the use of hashtags and retweets. Howard and Kollanyi [190] focus on the 2016 UK referendum and the role of bots in the conversations on Twitter. They analyze 1.5M tweets from 313K Twitter accounts collected by searching specific hashtags related to the referendum. Their analysis indicates that most of the tweets are in favor of exiting the EU, there are bots with different levels of automation and that 1% of the accounts generate 33% of the overall messages. They also note that among the top sharers, there are a lot of bot accounts that are mostly retweeting and not generating new content. In a similar work, Howard et al. [201] study Twitter behavior during the second 2016 US Presidential Debate. They find that Twitter activity is more pro-Trump and that a lot of activity is driven by bots. However, they note that a substantial amount of tweets is original content posted from regular Twitter users. Woolley [200] analyzes several articles regarding the use of bots in OSNs for political purposes. Specifically, he undertakes a qualitative content analysis on 41 articles regarding political bots from various countries obtained from the Web. One of his main findings is that the use of bots varies from country to country and that some countries (e.g., Argentina, China, Russia, USA, etc.) use political bots on more than one type of event. For example, they report the use of Chinese political bots for elections, for protests and for security reasons.

In the Chinese political stage, during December 2014, an anonymous blogger released an archive of emails pertaining to the employment of Wumao, a group of people that gets paid to disseminate propaganda on social media, from the Chinese government. King et al. [197] analyzed these leaks and found out 43K posts that were posted by Wumao. Their main findings are: 1) by analyzing the time-series of these posts, they find bursty activity, hence signs of coordination of the posters; 2) most of the posters are individuals working for the government; and 3) by analyzing the content of the message, they note that posters usually post messages for distraction rather than discussions of controversial matters (i.e., supporting China's regime instead of discussing an event). Similarly to the previous work, Yang et al. [198] study the Wumao by analyzing 26M posts from 2.7M users on the Sina Weibo OSN, aiming to provide insights regarding the behavior and the size of Wumao. Due to the lack of ground truth data, they use clustering and topic modeling techniques, in order to cluster users that post politics-related messages with similar topics. By manually checking the users on the produced clusters, they conclude that users that post pro-government messages are distributed across multiple clusters, hence there is no signs of coordination of the Wumao on Sina Weibo for the

period of their dataset (August 2012 and August 2013).

Zannettou et al. [60] study Russian state-sponsored troll accounts and measure the influence they had on Twitter and other Web communities. They find that Russian trolls were involved in the discussion of political events, and that they exhibit different behavior when compared to random users. Finally, they show that their influence was not substantial, with the exception of the dissemination of articles from state-sponsored Russian news outlets like Russia Today (RT). Allcot and Gentzkow [55] make a large scale analysis on Facebook during the period of the 2016 US election. Their results provide the following interesting statistics about the US election: 1) 115 pro-Trump fake stories are shared 30M times, whereas 41 pro-Clinton fake stories are shared 7.6M times. This indicates that fake news stories that favor Trump are more profound in Facebook. 2) The aforementioned 37.6M shares translates to 760M instances of a user clicking to the news articles. This indicates the high reachability of the fake news stories to end-users. 3) By undertaking a 1200-person survey, they highlight that a user's education, age and overall media consumption are the most important factors that determine whether a user can distinguish false headlines.

Rumors. Shin et al. [191] undertake a content-based analysis on 330K tweets pertaining to the 2012 US election. Their findings agree with existing literature, noting that users that spread rumors are mostly sharing messages against a political person. Furthermore, they highlight the resilience of rumors despite the fact that rumor debunking evidence was disseminated in Twitter; however, this does not apply for rumors that originate from satire websites.

3.5.3 Other models/algorithms

Biased. An et al. [192] study the interactions of 7M followers of 24 US news outlets on Twitter, in order to identify political leaning. To achieve this, they create a distance model, based on co-subscription relationships, that maps news sources to a dimensional dichotomous political spectrum. Also, they propose a real-time application, which utilizes the underlying model, and visualizes the ideology of the various news sources. Yang et al. [195] investigate the topics of discussions on Twitter for 51 US political persons, including President Obama. The main finding of this work is that Republicans and Democrats are similarly active on Twitter with the difference that Democrats tend to use hashtags more frequently. Furthermore, by utilizing a graph that demonstrates the similarity of the agenda of each political person, they highlight that Republicans are more clustered. This indicates that Republicans tend to share more tweets regarding their party's issues and agenda.

Propaganda. Al-khateeb and Agarwal [53] study the dissemination of propaganda on Twitter from terrorist organizations (namely ISIS). They propose a framework based on social studies that aim to identify social and behavioral patterns of propaganda messages disseminated by a botnet. Their main findings are that bots exhibit similar behaviors (i.e., similar sharing patterns, similar usernames, lot of tweets in a short period of time) and that they share information that contains URLs to other sites and blogs. Ranganath et al. [193] focus on the detection of political advocates (individuals that use social media to strategically push a political agenda) on Twitter. The authors note that identifying advocates is not a straightforward task due to the nuanced and diverse message construction and propagation strategies. To overcome this, they propose a framework that aims to model all the different propagation and message construction strategies of advocates. Their evaluation on two datasets on Twitter regarding gun rights and elections demonstrate that the proposed framework achieves good performance with a 93% AUC score.

Rumors. Jin et al. [194] study the 2016 US Election through the Twitter activity of the followers of the two presidential candidates. For identifying rumors, they collect rumor articles from Snopes and then they use text similarity algorithms based on: 1) Term frequency-inverse document frequency (TF-IDF); 2) BM25 proposed in [202] 3) Word2Vec embeddings [12]; 4) Doc2Vec embeddings [203]; 5) Lexicon used in [135]. Their evaluation indicates that the best performance is achieved using the BM25-based approach. This algorithm is subsequently used to classify the tweets of the candidates' followers. Based on the predictions of the algorithm, their main findings are: 1) rumors are more prevalent during election period; 2) most of the rumors are posted by a small group of users; 3) rumors are mainly posted to debunk rumors that are against their presidential candidate, or to inflict damage on the other candidate; and 4) rumor sharing behavior increases in key points of the presidential campaign and in emergency events.

3.5.4 False information in political stage - Remarks

The main insights from the review of work that focus on the political stage are: 1) Temporal analysis can by leveraged to assess coordination of bots, state-sponsored actors, and orchestrated efforts on disseminating political false information; 2) Bots are extensively used for the dissemination of political false information; 3) Machine learning techniques can assist in detecting political false information and political leaning of users. However, there are concerns about the generalization of such solutions on other datasets/domains; and 4) Political

campaigns are responsible for the substantial dissemination of political false information in mainstream Web communities.

3.6 Other related work

In this section we present work that is relevant to the false information ecosystem but does not fit in the aforementioned lines of work. Specifically, we group these studies in the following categories: 1) General Studies; 2) Systems; and 3) Use of images on the false information ecosystem.

3.6.1 General Studies

Credibility Assessment. Buntain and Golbeck [204] compare the accuracy of models that use features based on journalists assessments and crowdsourced assessments. They indicate that there is small overlap between the two features sets despite the fact that they provide statistically correlated results. This indicates that crowdsourcing workers discern different aspects of the stories when compared to journalists. Finally, they demonstrate that models that utilize features from crowdsourcing outperform the models that utilize features from journalists. Zhang et al. [205] present a set of indicators that can used to assess the credibility of articles. To find these indicators they use a diverse set of experts (coming from multiple disciplines), which analyzed and annotated 40 news articles. Despite the low number of annotated articles, this inter-disciplinary study is important as it can help in defining standards for assessing the credibility of content on the Web. Mangolin et al. [206] study the interplay between fact-checkers and rumor spreaders on social networks finding that users are more likely to correct themselves if the correction comes from a user they follow when compared to a stranger.

Conspiracy Theories. Starbird [207] performs a qualitative analysis on Twitter regarding shooting events and conspiracy theories. Using graph analysis on the domains linked from the tweets, she provides insight on how various websites work to promote conspiracy theories and push political agendas.

Fabricated. Horne and Adah [208] focus on the headline of fake and real news. Their analysis on three datasets of news articles highlight that fake news have substantial differences in their structure when compared with real news. Specifically, they report that generally the structure

of the content and the headline is different. That is, fake news are smaller in size, use simple words, and use longer and "clickbaity" headlines. Potts et al. [209] study Reddit and 4chan and how their interface is a part of their culture that affects their information sharing behavior. They analyze the information shared on these two platforms during the 2013 Boston Marathon bombings. Their findings highlight that users on both sites tried to find the perpetrator of the attack by creating conversations for the attack, usually containing false information. Bode and Vraga [210] propose a new function on Facebook, which allow users to observe related stories that either confirm or correct false information; they highlight that using this function users acquire a better understanding of the information and its credibility. Finally, Pennycook and Rand [211] highlight that by attaching warnings to news articles can help users to better assess the credibility of articles, however news articles that are not attached with warnings are considered as validated, which is not always true, hence users are tricked.

Propaganda. Chen et al. [59] study the behavior of hidden paid posters on OSNs. To better understand how these actors work, an author of this work posed as a hidden paid poster for a site[212] that gives users the option to be hidden paid posters. This task revealed valuable information regarding the organization of such sites and the behavior of the hidden paid posters, who are assigned with missions that need to be accomplished within a deadline. For example, a mission can be about posting articles of a particular content on different sites. A manager of the site can verify the completion of the task and then the hidden paid poster gets paid. To further study the problem, they collect data ,pertaining to a dispute between two big Chinese IT companies, from users of 2 popular Chinese news sites (namely Sohu [213] and Sina [214]). During this conflict there were strong suspicions that both companies employed hidden paid posters to disseminate false information that aimed to inflict damage to the other company. By undertaking statistical and semantic analysis on the hidden paid posters' content they uncover a lot of useful features that can be used in identifying hidden paid posters. To this end, they propose the use of SVMs in order to detect such users by taking into consideration statistical and semantic features; their evaluation show that they can detect users with 88% accuracy.

Rumors. Starbird et al. [215] study and identify various types of expressed uncertainty within posts in OSN during a rumor's lifetime. To analyze the uncertainty degree in messages, the paper acquires 15M tweets related to two crisis incidents (Boston Bombings and Sydney Siege). They find that specific linguistic patterns are used in rumor-related tweets. Their findings can be used in future detection systems in order to detect rumors effectively in a timely manner. Zubiaga et al. [216] propose a different approach in collecting and preparing

datasets for false information detection. Instead of finding rumors from busting websites and then retrieving data from OSNs, they propose the retrieval of OSN data that will subsequently annotated by humans. In their evaluation, they retrieve tweets pertaining to the Ferguson unrest incident during 2014. They utilize journalists that act as annotators with the aim to label the tweets and their conversations. Specifically, the journalists annotated 1.1k tweets, which can be categorized into 42 different stories. Their findings show that 24.6% of the tweets are rumorous. FInally, Spiro et al. [217] undertake a quantitative analysis on tweets pertaining to the 2010 Deepwater Horizon oil spill. They note that media coverage increased the number of tweets related to the disaster. Furthermore, they observe that retweets are more commonly transmitted serially when they have event-related keywords.

3.6.2 Systems

Biased. Park et al. [218] note that biased information is profoundly disseminated in OSNs. To alleviate this problem, they propose NewsCube: a service that aims to provide end-users with all the different aspects of a particular story. In this way, end-users can read and understand the stories from multiple perspectives hence assisting in the formulation of their own unbiased view for the story. To achieve this, they perform structure-based extraction of the different aspects that exist in news stories. These aspects are then clustered in order to be presented to the end-users. To evaluate the effectiveness of their system, they undertake several user studies that aim to demonstrate the effectiveness in terms of the ability of the users to construct balanced views when using the platform. Their results indicate that 16 out of 33 participants stated that the platform helped them formulate a balanced view of the story, 2 out of 33 were negative, whereas the rest were neutral.

Credibility Assessment. Hassan et al. [219] propose FactWatcher, a system that reports facts that can be used as leads in stories. Their system is heavily based on a database and offers useful features to it's users such as ranking of the facts, keyword-based search and fact-to-statement translation. Ennals et al. [220] describe the design and implementation of Dispute Finder, which is a browser extension that allows users to be warned about claims that are disputed by sources that they might trust. Dispute Finder maintains a database with well-known disputed claims which are used to inform end-users in real-time while they are reading stories. Users are also able to contribute to the whole process by explicitly flagging content as disputed, or as evidence to dispute other claims. In the case of providing evidence, the system requires a reference to a trusted source that supports the user's actions, thus ensuring

the quality of user's manual annotations. Mitra and Gilbert [221] propose CREDBANK that aims to process large datasets by combining machine and human computations. The former is used to summarize tweets in events, while the latter is responsible for assessing the credibility of the content. Pirolli et al. [222] focus on Wikipedia and develop and system that presents users an interactive dashboard, which includes the history of article content and edits. The main finding is that users can better judge the credibility of an article, given that they are presented with the history of the article and edits through an interactive dashboard.

3.6.3 Use of images on the false information ecosystem

Information can be disseminated via images on the Web. The use of images increases the credibility of the included information, as users tend to believe more information that is substantiated with an image. However, nowadays, images can be easily manipulated, hence used for the dissemination of false information. In this section, we provide an overview of the papers that studied the problem of false information on the Web, while considering images.

Fabricated. Boididou et al. [223, 224] focus on the use of multimedia in false information spread in OSNs. In [224] they prepare and propose a dataset of 12K tweets, which are manually labeled as fake, true, or unknown. A tweet is regarded as true if the image is referring to a particular event and fake if the image is not referring to a particular event. The authors argue that this dataset can help researchers in the task of automated identification of fake multimedia within tweets. In [223] they study the challenges that exist in providing an automated verification system for news that contain multimedia. To this end, they propose the use of conventional classifiers with the aim to discern fake multimedia pertaining to real events. Their findings demonstrate that generalizing is extremely hard as their classifiers perform poorly (58% accuracy) when they are trained with a particular event and they are tested with another. Diego Saez-Trumper [225] proposes a Web application, called Fake Tweet Buster, that aims to warn users about tweets that contain false information through images or users that habitually diffuse false information. The proposed approach is based on the reverse image search technique (using Google Images) in order to determine the origin of the image, its age and its context. Furthermore, the application considers user attributes and crowdsourcing data in order to find users that consistently share tweets that contain false information on images. Pasquini et al. [226] aim to provide image verification by proposing an empirical system that seeks visually and semantically related images on the web. Specifically, their system utilizes news articles metadata in order to search, using Google's search engine, for relevant news

articles. These images are then compared with the original's article images in order to identify whether the images were tampered. To evaluate their approach, they created dummy articles with tampered images in order to simulate the whole procedure.

Jin et al. [227] emphasize the importance of images in news articles for distinguishing its truthfulness. They propose the use of two sets of features extracted from images in conjunction with features that are proposed by [119, 121]. For the image features, they define a set of visual characteristics as well as overall image statistics. Their data is based on a corpus obtained from the Sina Weibo that comprises 50K posts and 26K images. For evaluating the image feature set, they use conventional machine learning techniques: namely SVM, Logistic Regression, KStar, and Random Forest. They find that the proposed image features increase the accuracy by 7% with an overall accuracy of 83%. In a follow-up work, Jin et al. [228] leverage deep neural networks with the goal of distinguishing the credibility of images. They note that this task is extremely difficult as images can be misleading in many ways. Specifically, images might be outdated (i.e., old images that are falsely used to describe a new event), inaccurate, or even manipulated. To assess the image credibility, they train a Convolutional Neural Network (CNN) using a large-scale auxiliary dataset that comprises 600K labeled fake and real images. Their intuition is that the CNN can extract useful hyperparameters that can be used to detect eye-catching and visually striking images, which are usually used to describe false information. Their evaluation indicates that the proposed model can outperform several baselines in terms of the precision, recall, F1, and accuracy scores. Gupta et al. [229] focus on the diffusion of fake images in Twitter during Hurricane Sandy in 2012. They demonstrate that the use of automated techniques (i.e., Decision Trees) can assist in distinguishing fake images from real ones. Interestingly, they note that the 90% of the fake images came from the top 0.3% of the users.

Chapter 4

Understanding the Spread Of Information Through The Lens Of Multiple Web Communities

In this chapter, we present our work that helps in better understanding the spread of information across the Web and how web communities influence each other. We focus on understanding the spread of news and image-based memes across multiple Web communities, namely, Twitter, Reddit, 4chan, and Gab.

4.1 Understanding How Web Communities Influence Each Other Through the Lens of News Sources

4.1.1 Motivation

Over the past few years, several conspiracy theories and false stories have spread on the Web. Some examples include the Boston Marathon bombings in 2013, where a large number of tweets started to claim that the bombings were a "false flag" perpetrated by the goverment of the United States. More recently, the Pizzagate conspiracy [36] – a debunked theory connecting a restaurant and members of the US Democratic Party to a child sex ring – led to a shooting in a Washinghton DC restaurant [230]. These stories were all propagated, in no small part, via the use of "alternative" news sites like Infowars and "fringe" Web communities

like 4chan. This is mainly because the barrier of entry for such alternative news sources has been greatly reduced by the Web and large social networks. Due to the negligible cost of distributing information over social media, fringe sites can quickly gain traction with large audiences.

Although previous works have studied the dissemination of false information on the Web, as discussed in Chapter 3, very little work provides a holistic view of the modern information ecosystem. This knowledge, however, is crucial for understanding the alternative news world and for designing appropriate detection and mitigation strategies. Anecdotal evidence and press coverage suggest that alternative news dissemination might start on fringe sites, eventually reaching mainstream online social networks and news outlets [20, 4]. Nevertheless, this phenomenon has not been measured and no thorough analysis has focused on how news moves from one online service to another, sort of forming an interconnected centipede of Web Communities.

In this work, we address this gap by performing the first thorough large-scale measurement on how mainstream and alternative news flows through three Web Communities; namely Twitter, Reddit, and 4chan. We focus on these three platforms because of: 1) they are fundamentally different and they drive substantial portions of the online world; 2) there is anecdotal evidence that suggests that specific communities within Reddit and 4chan act as generators [20] and incubators [231] of false information; and 3) they are able to have a substantial impact in forming and manipulating peoples' opinions by constantly circulating false information [230].

Contributions. First, we undertake a large-scale measurement and comparison of the occurrence of mainstream and alternative news sources across three social media platforms (4chan, Reddit, and Twitter). Then, we provide an understanding of the temporal dynamics of how URLs from news sites are posted on the different social networks. Finally, we present a measurement of the influence between the platforms that provides insight into how information spreads throughout the greater Web. Overall, our findings indicate that Twitter, Reddit, and 4chan are used quite extensively for the dissemination of both alternative and mainstream news. Using a statistical model for influence – namely, Hawkes processes – we show that each of the platforms have varying degrees of in influence on each other, and this influence differs with respect to mainstream and alternative news sources.

4.1.2 Datasets

Our analysis uses a set of news websites that can confidently be labeled as either "mainstream" or "alternative" news. More specifically, we create a list of 99 news sites including 45 mainstream and 54 alternative ones.[1] For the former, we select 45 from the Alexa top 100 news sites, leaving out those based on user-generated content, those serving specialized content (e.g., finance news), as well as non-English sites. For the latter, we use Wikipedia [232] and FakeNewsWatch [233]. We also add two state-sponsored alternative news domains: `sputniknews.com` and `rt.com`, as they have recently attracted public attention due to their posting of controversial, and seemingly agenda-pushing stories [234].

We gather information from posts, threads, and comments on Twitter, Reddit, and 4chan that contain URLs from the 99 news sites. With a few gaps (see below), our datasets cover activity on the three platforms between June 30, 2016 and February 28, 2017. Table 4.1 shows the total number of posts/comments crawled and the percentage of posts that contains links to URLs from the aforementioned news domains. We observe that mainstream news URLs are present in a greater percentage of posts on 4chan and Reddit than on Twitter, while alternative ones are about twice as likely to appear in posts on 4chan than on Twitter or Reddit. Table 4.2 provides a summary of our datasets, which we present in more detail below. Note that we break Reddit and 4chan datasets into two different instances, as further discussed.

Platform	Total Posts	% Alt.	% Main.
Twitter	587M	0.022%	0.070%
Reddit (posts + comments)	332M	0.023%	0.181%
4chan	42M	0.050%	0.197%

Table 4.1: Total number of posts crawled and percentage of posts that contain URLs to our list of alternative and mainstream news sites.

Twitter. We collect the 1% of all publicly available tweets with URLs from the aforementioned news domains between June 30, 2016 and February 28, 2017 using the Twitter Streaming API [235]. In total, we gather 487K tweets containing 279K unique URLs pointing to mainstream or alternative news sites. Since tweets are retrieved at the time they are posted, we do not get information such as the number of times they are re-tweeted or liked. Therefore, between March and May 2017, we re-crawled each tweet to retrieve this data. Basic statistics

[1] The complete list of the 99 sites is available at `https://drive.google.com/open?id=0ByP5a__khV0dM1ZSY3YxQWF2N2c`

Platform	Posts/Comments	Alt. URLs	Main. URLs
Twitter	486,700	42,550	236,480
Reddit (six selected subreddits)	620,530	40,046	301,840
Reddit (all other subreddits)	1,228,105	24,027	726,948
4chan (/pol/)	90,537	8,963	40,164
4chan (/int/, /sci/, /sp/)	7,131	615	5,513

Table 4.2: Overview of our datasets with the number of posts/comments that contain a URL to one of our information sources, as well as the number of unique URLs linking to alternative and mainstream news sites in our list.

are summarized in Table 4.3. Due to a failure in our collection infrastructure, we have some gaps in the Twitter dataset, specifically between Oct 28–Nov 2 and Nov 5–16, 2016, as well as Nov 22, 2016 – Jan 13, 2017, and Feb 24–28, 2017.

	Tweets	Retrieved (%)	Avg. Retweets	Avg. Likes
Alternative	110,629	92,104 (83.2%)	$341 \pm 1{,}228$	0.82 ± 15.6
Mainstream	376,071	329,950 (87.7%)	$404 \pm 2{,}146$	0.96 ± 55.6

Table 4.3: Basic statistics of the occurrence of alternative and mainstream news URLs in the tweets in our dataset.

Reddit. We obtain all posts and comments on Reddit between June 30, 2016 and February 28, 2017, using data made available on Pushshift [236]. We collect approximately 42M posts, 390M comments, and 300K subreddits. Once again, we filter posts and comments that contain URLs from one of the 99 news sites, which yields a dataset of 1.8M posts/comments and approximately 1.1M URLs.

4chan. For 4chan, we use all threads and posts made on the Politically Incorrect (/pol/) board, as well as /sp/ (Sports), /int/ (International), and /sci/ (Science) boards for comparison, using the same methodology as [19]. We opt to select both not safe for work boards (i.e., /pol/) and safe for work boards (i.e., /sp/, /int/, and /sci/) to observe how these compare to each other with respect to the dissemination of news. The resulting dataset includes 97K posts and replies, including 56K alternative and mainstream news URLs, between June 30, 2016 and February 28, 2017. We have some small gaps due to our crawler failing, specifically, Oct 15–16 and Dec 16–25, 2016 as well as Jan 10–13, 2017.

Subreddit (Alt.)	(%)	Subreddit (Alt.)	(%)	Subreddit (Main.)	(%)	Subreddit (Main.)	(%)
The_Donald	35.37 %	KotakuInAction	1.04 %	politics	12.9 %	EnoughTrumpSpam	1.20 %
politics	8.21 %	HillaryForPrison	0.94 %	worldnews	6.24 %	NoFilterNews	1.16 %
news	3.85 %	TheOnion	0.94 %	The_Donald	4.53 %	BreakingNews24hr	1.07 %
conspiracy	3.84 %	AskTrumpSupporters	0.84 %	news	4.23 %	conspiracy	0.89 %
Uncensored	2.66 %	POLITIC	0.81 %	TheColorIsBlue	3.06 %	todayilearned	0.83 %
Health	2.10 %	rss_theonion	0.67 %	TheColorIsRed	2.48 %	thenewsrightnow	0.78 %
PoliticsAll	1.54 %	the_Europe	0.67 %	willis7737_news	2.27 %	europe	0.77 %
Conservative	1.45 %	new_right	0.6 %	news_etc	1.94 %	ReddLineNews	0.75 %
worldnews	1.41 %	AskReddit	0.59 %	AskReddit	1.37 %	hillaryclinton	0.73 %
WhiteRights	1.21 %	AnythingGoesNews	0.51 %	canada	1.31 %	nottheonion	0.73 %

Table 4.4: Top 20 subreddits w.r.t. mainstream and alternative news URLs occurrence and their percentage in Reddit (all subreddits).

4.1.3 General Characterization

In this section, we present a general characterization of the mainstream and alternative news URLs found on the three platforms.

Reddit. We start by identifying news and politics communities. In Table 4.4, we report the top 20 subreddits with the most URLs, along with their percentage. Note that we omit automated ones (e.g., /r/AutoNewspaper/) where news articles are posted without user intervention. Many of the subreddits are indeed related to news and politics – e.g., 'The_Donald' is mostly a community of Donald Trump supporters, while 'worldnews' is focused around globally relevant events. We also find the presence of the 'conspiracy' subreddit, which has been involved in disinformation campaigns including Pizzagate, as well as 'AskReddit,' where both mainstream and alternative news sources are used to answer questions submitted by users. Although the latter is intended for open-ended questions that spark discussion, it is evident that commenters often try to push their agenda even in non-political threads. In the end, based on their propensity to include news URLs of both types, we single out the follow top six subreddits for further exploration: The_Donald, politics, conspiracy, news, worldnews, and AskReddit.

In order to get a better view of the popularity of news sites on the six subreddits, we study the occurrence of each news outlet. Specifically, we find 76K URLs (40K unique) from alternative news and 600K (301K unique) from mainstream news domains. Table 4.5 reports the top 20 mainstream/alternative news sites and their percentage in the six subreddits. The top 20 domains for mainstream news account for 89% of all mainstream news URLs in our data,

Domain (Alt.)	(%)	Domain (Alt.)	(%)	Domain (Main.)	(%)	Domain (Main.)	(%)
breitbart.com	55.58 %	prntly.com	0.49 %	nytimes.com	14.07 %	nbcnews.com	2.86 %
rt.com	19.18 %	dccclothesline.com	0.4 %	cnn.com	11.23 %	time.com	2.57 %
infowars.com	8.99 %	worldnewsdailyreport.com	0.36 %	theguardian.com	8.86 %	washinghtontimes.com	2.52 %
sputniknews.com	3.95 %	therealstrategy.com	0.3 %	reuters.com	6.67 %	bloomberg.com	2.5 %
beforeitsnews.com	2.34 %	disclose.tv	0.23 %	huffingtonpost.com	5.67 %	wsj.com	2.31 %
lifezette.com	2.28 %	clickhole.com	0.2 %	thehill.com	5.15 %	cbsnews.com	2.26 %
naturalnews.com	1.54 %	libertywritersnews.com	0.2 %	foxnews.com	4.89 %	thedailybeast.com	2.05 %
activistpost.com	1.45 %	worldtruth.tv	0.14 %	bbc.com	4.76 %	forbes.com	1.87 %
veteranstoday.com	1.11 %	thelastlineofdefence.org	0.07 %	abcnews.go.com	2.94 %	nypost.com	1.85 %
redflagnews.com	0.63 %	nodisinfo.com	0.05 %	usatoday.com	2.87 %	cncb.com	1.54 %

Table 4.5: Top 20 mainstream and alternative domains and their percentage in the six selected subreddits.

Domain (Alt.)	(%)	Domain (Alt.)	(%)	Domain (Main.)	(%)	Domain (Main.)	(%)
breitbart.com	46.04 %	activistpost.com	0.41 %	theguardian.com	19.04 %	usatoday.com	2.02 %
rt.com	17.56 %	disclose.tv	0.39 %	nytimes.com	10.07 %	thedailybeast.com	2.02 %
infowars.com	17.25 %	prntly.com	0.26 %	bbc.com	8.99 %	nbcnews.com	1.96 %
therealstrategy.com	5.63 %	worldtruth.tv	0.25 %	forbes.com	6.24 %	nypost.com	1.95 %
sputniknews.com	4.11 %	libertywriternews.com	0.15 %	thehill.com	4.95 %	cbsnews.com	1.89 %
beforeitsnews.com	2.26 %	worldnewsdailyreport.com	0.06 %	cbc.ca	4.82 %	abcnews.go.com	1.78 %
redflagnews	2.04 %	mediamass.net	0.04 %	foxnews.com	4.79 %	time.com	1.71 %
dccclothesline.com	1.37 %	newsbiscuit.com	0.03 %	wsj.com	4.04 %	cnbc.com	1.40 %
naturalnews.com	1.29 %	react365.com	0.02 %	bloomberg.com	3.48 %	washingtontimes.com	1.34 %
clickhole.com	0.53 %	the-daily.buzz	0.02 %	reuters.com	2.85 %	washingtonexaminer.com	1.33 %

Table 4.6: Top 20 mainstream and alternative news sites in the Twitter dataset and their percentage.

while for alternative domains the percentage is 99%. Known alt-right news outlets, such as `breitbart.com` and `infowars.com`, are predominantly present, as well as state-sponsored alternative domains like `sputniknews.com` and `rt.com`, which have recently been in the spotlight for disseminating false information and propaganda [234]. The fact that many such URLs appear in our dataset may indeed be an indication that the six subreddits significantly contribute to the dissemination of controversial stories.

Twitter. In our Twitter dataset, we find 129K (42K unique) URLs of alternative news domains and 413K (236K unique) URLs of mainstream ones. Recall that we re-crawl tweets to get the number of retweets and likes, and a small percentage of them are no longer available as they were either deleted or the associated account was suspended. This percentage is slightly higher for tweets with URLs from alternative news, possibly due to the fact that some users tend to remove controversial content when a particular false story is debunked [104]. Also, alternative and mainstream news tend to get a significant number of retweets, at about the

Domain (Alt.)	(%)	Domain (Alt.)	(%)	Domain (Main.)	(%)	Domain (Main.)	(%)
breitbart.com	53.00 %	activistpost.com	0.38 %	theguardian.com	14.10 %	wsj.com	2.82 %
rt.com	28.22 %	dccclothesline.com	0.29 %	nytimes.com	10.07 %	washinghtontimes.com	2.77 %
infowars.com	9.12 %	redflagnews.com	0.20 %	cnn.com	9.90 %	bloomberg.com	2.75 %
sputniknews.com	3.36 %	libertywritersnews.com	0.16 %	bbc.com	5.45 %	cbc.ca	2.66 %
veteranstoday.com	1.07 %	therealstrategy.com	0.16 %	foxnews.com	5.35 %	nypost.com	2.65 %
beforeitsnews.com	0.91 %	clickhole.com	0.11 %	reuters.com	5.10 %	cbsnews.com	2.44 %
lifezette.com	0.86 %	disclose.tv	0.10 %	time.com	3.42 %	nbcnews.com	2.32 %
naturalnews.com	0.61 %	now8news.com	0.06 %	abcnews.go.com	3.40 %	usatoday.com	2.25 %
worldnewsdailyreport.com	0.46 %	firebrandleft.com	0.05 %	huffingtonpost.com	3.29 %	cnbc.com	2.13 %
prntly.com	0.41 %	nodisinfo.com	0.05 %	thehill.com	3.04 %	forbes.com	1.68 %

Table 4.7: Top 20 mainstream and alternative news sites in the /pol/ dataset and their percentage.

same rate (on average, 404 and 341 retweets per tweet, respectively). A similar pattern is observed for likes (see Table 4.3).

In Table 4.6, we report the top 20 mainstream and alternative news domains, and their percentage, in our Twitter dataset. These cover, respectively, 86% and 99% of all URLs. Similar to Reddit, there are many popular alt-right and state-sponsored news outlets.

4chan. In our /pol/ dataset, we find 21K (9K unique) URLs to alternative news outlets and 82K (40K unique) to mainstream news. Table 4.7 reports the percentage of URLs of the top 20 domains for each type of news. These cover 87% and 99% of mainstream and alternative news URLs, respectively. Again, we observe that, by far, the most popular alternative news domains are `breitbart.com`, `rt.com`, `infowars.com`, and `sputniknews.com`. For the mainstream news, we observe that `theguardian.com` is the most frequently posted, followed by `nytimes.com`, `cnn.com`, and `bbc.com`. We also obtained similar statistics for domain popularity in the other boards of 4chan, but we omit them for brevity.

To get a better view of the platforms' URL posting behavior, Fig. 4.1 plots the CDF of URL appearances (i.e., how many times a specific URL appears) within a particular platform. We observe that a substantial portion of the URLs appear only once for both alternative and mainstream news, and that, on Twitter, alternative news tends to appear more times than mainstream news. For /pol/ and the six subreddits, we observe a similar behavior for both mainstream and alternative news.

Next, in Fig. 4.2, we compare how popular domains, in both categories, appear on the three platforms (i.e., Twitter, the six subreddits, and /pol/). We find that the top 4 alternative domains – `breitbart.com`, `rt.com`, `infowars.com`, `sputniknews.com` – influence the three platforms more or less in the same way. However, some outlets appear predominantly in

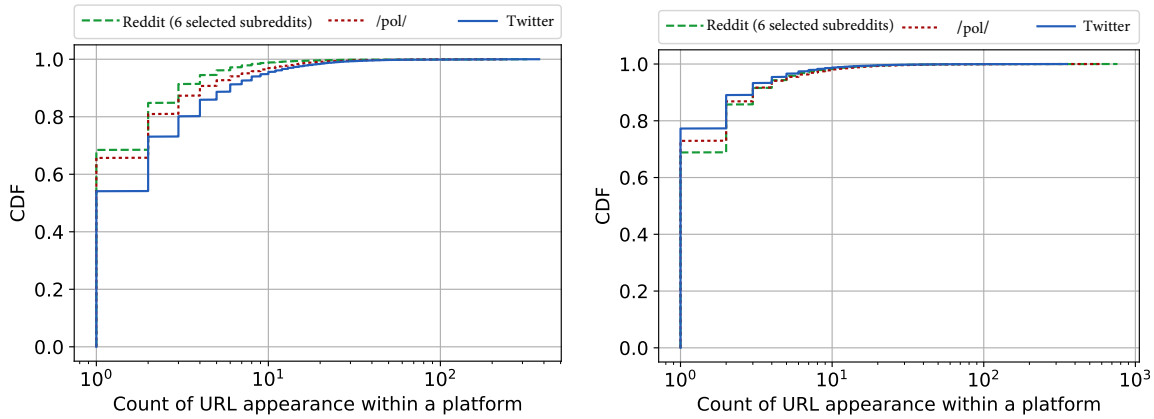


Figure 4.1: CDF of the counts of URL appearance within a particular platform: (a) alternative news and (b) mainstream news.

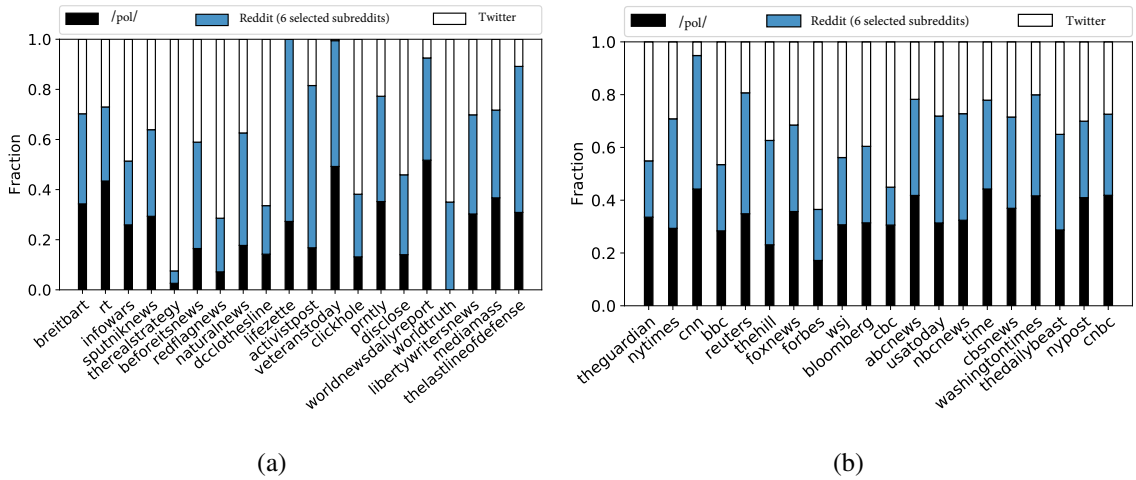
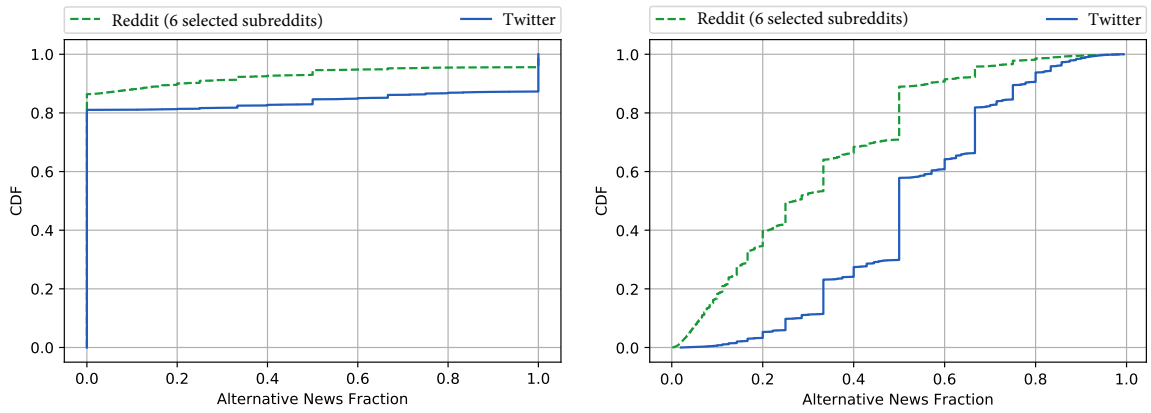


Figure 4.2: Top 20 domains and each platform's fraction for (a) alternative and (b) mainstream news.

some platforms but not in others; e.g., `therealstrategy.com` is popular only on Twitter, while `lifezette.com` and `veteranstoday.com` are popular on the 6 subreddits and /pol/, but not on Twitter. We believe the primary reason for this has to do with Twitter bots. We cannot exclude with certainty that bots do not exist on 4chan, while bots are actually acceptable on Reddit (as long as they follow the rules of Reddit's API [237]), however, they are certainly more prevalent on Twitter. Thus, if a particular domain is popular on Twitter because of the influence of bots, then it might not be popular on Reddit and 4chan. We have also considered ways to factor out posting behavior from bots, especially for Twitter, such as the one proposed in [238]. However, we have not removed this activity due to: 1) posting behavior from bots can affect real users' posting behavior, hence this activity is part of the



(a) (b)

Figure 4.3: CDF of the fraction of URLs from alternative news and overall news URLs for (a) all users in our Twitter and Reddit datasets, and (b) users that shared URLs from both mainstream and alternative news.

overall news dissemination ecosystem and needs to be accounted for; and 2) the satisfactory performance of such approaches is yet to be proven.

We also measure the fraction of news URLs that are alternative, *per user*, in Fig. 4.3. We report this fraction only for Reddit and Twitter users, since on 4chan posts are anonymous. We find that 80% of the users of both platforms share only URLs from mainstream news, while, 13% of Twitter users – which are likely bots [239] – exclusively post URLs to alternative news. We observe from Fig. 4.3(b), which shows the fraction for users sharing URLs from both categories, that there is a wide distribution, especially on the six selected subreddits, between people that rarely share alternative news (fraction close to 0) and those who share them almost all the time (fraction close to 1). Moreover, we find that Twitter users share more alternative news: just 5% of these users have a fraction below 0.2, which might be also attributed to the presence of bots.

4.1.4 Temporal Analysis

In this section, we present the results of a cross-platform temporal analysis of the way news are posted on Twitter, Reddit, and 4chan.

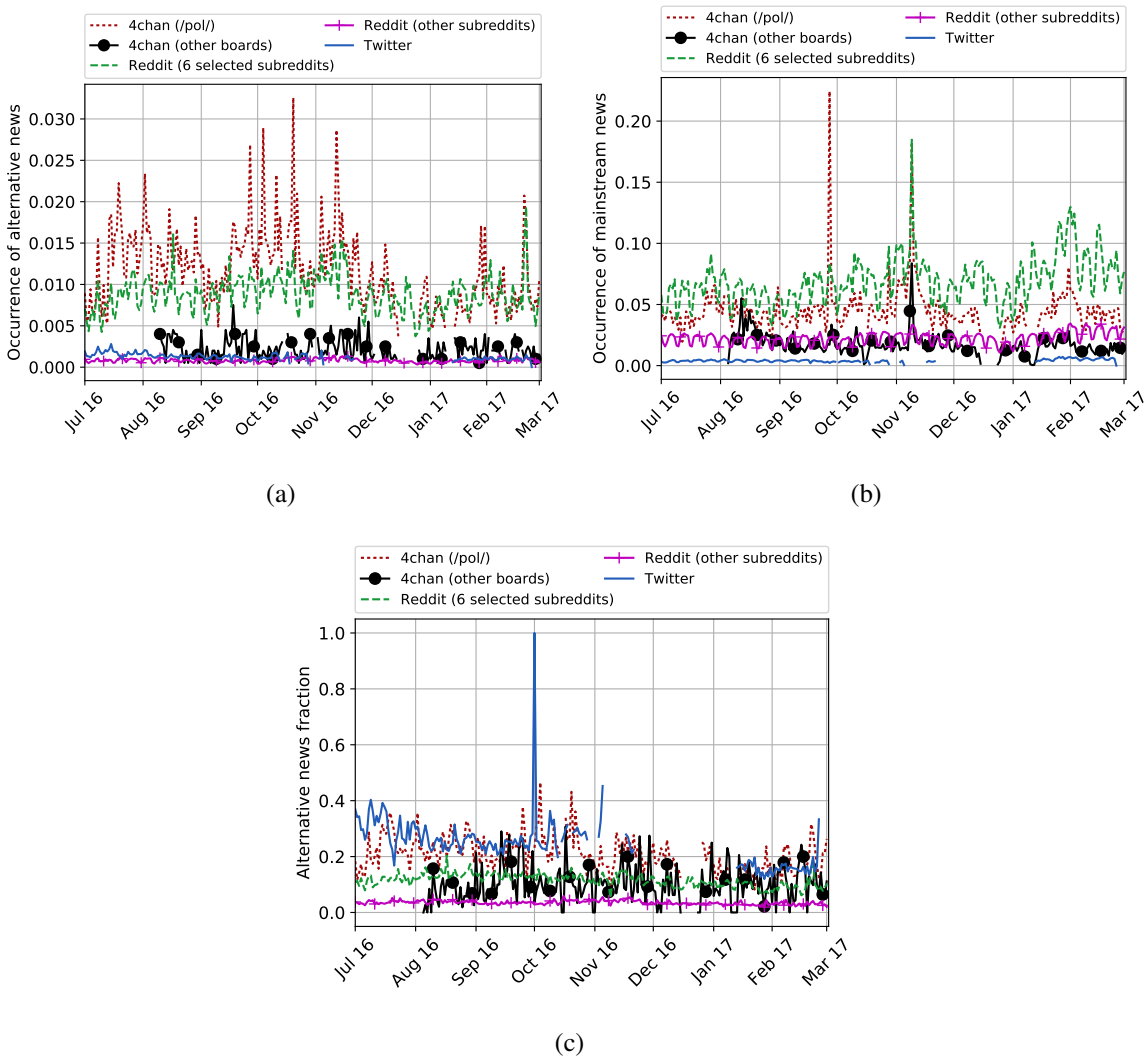
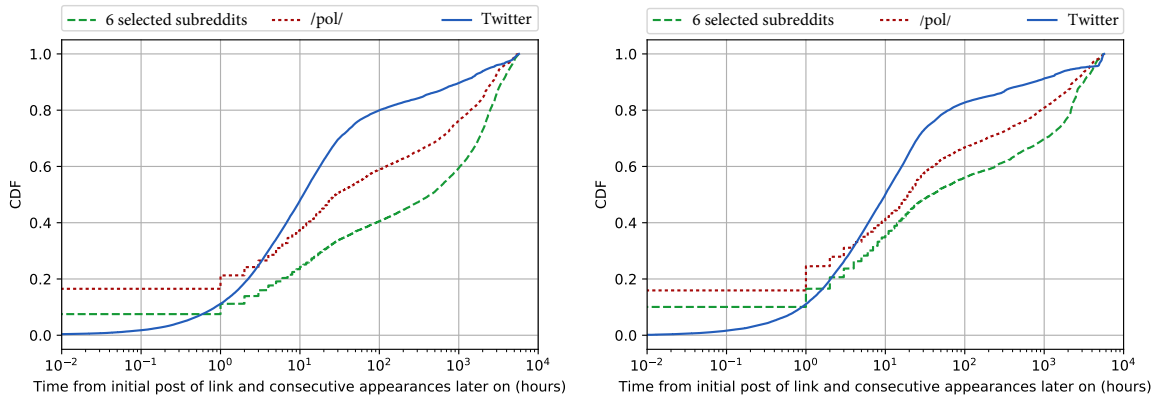


Figure 4.4: Normalized daily occurrence of URLs for (a) alternative news, (b) mainstream news, and (c) fraction of alternative news over all news.

URL Occurrence

In Fig. 4.4, we measure the daily occurrence of news URLs over the three platforms normalized by the average daily number of URLs shared in each community.[2] We find that /pol/ and the six selected subreddits exhibit a much higher percentage of occurrences of alternative news compared to the other communities (Fig. 4.4(a)), whereas, for mainstream news, the sharing behavior is more similar across platforms (Fig. 4.4(b)). There are also some interesting spikes, likely related to the 2016 US elections, on the date of the first presidential debate and election day itself. These findings indicate that the selected sub-communities are heavily utilized for

[2]Gaps in the plot correspond to gaps in our dataset due to crawler failure.



(a) (b)

Figure 4.5: CDF of time difference (in hours) between the first occurrence of a URL and its next occurrences on each platform for (a) alternative and (b) mainstream news.

the dissemination of alternative news. We also study the fraction of alternative news URLs with respect to overall news URLs (Fig. 4.4(c)), highlighting that mainstream news URLs are overall more "popular" than the alternative news URLs. Note that the Twitter spike in Fig. 4.4(c) appears to be an artifact of a failure in our collection infrastructure.

As some users repost the same URL many times within the same platform, we next study such reposting behavior and extract insights while comparing platforms. In Fig. 4.5, we plot the CDF of the time difference between the first occurrence of a URL and its next occurrences on the same platform. Both alternative and mainstream news URLs are recycled over time within the platform (even after several months), but Twitter exhibits a smaller lag between the first occurrence and later ones compared to the other two platforms. In all three platforms, there is an inflection point at the 24h period, which probably signifies the day-to-day behavior of news propagation within a platform, and this is true for both alternative and mainstream news. Finally, mainstream news seem to propagate faster in these platforms than alternative news, especially on the six subreddits; for Twitter and /pol/ the difference is not evident.

We also study the inter-arrival time of reposted URLs. Fig. 4.6 shows the CDF of the mean inter-arrival time of URLs that appear more than one time in each platform. Each platform exhibits unique behavior, confirmed by a two sample Kolmogorov-Smirnov test showing significant differences between the distributions ($p < 0.01$ for each pairwise comparison). However, /pol/ and the six subreddits exhibit similar time-related sharing behavior for both mainstream and alternative news URLs, and Twitter has smaller mean inter-arrival time overall. Interestingly, the six subreddits appear to have a duality in reposting behavior: for URLs with

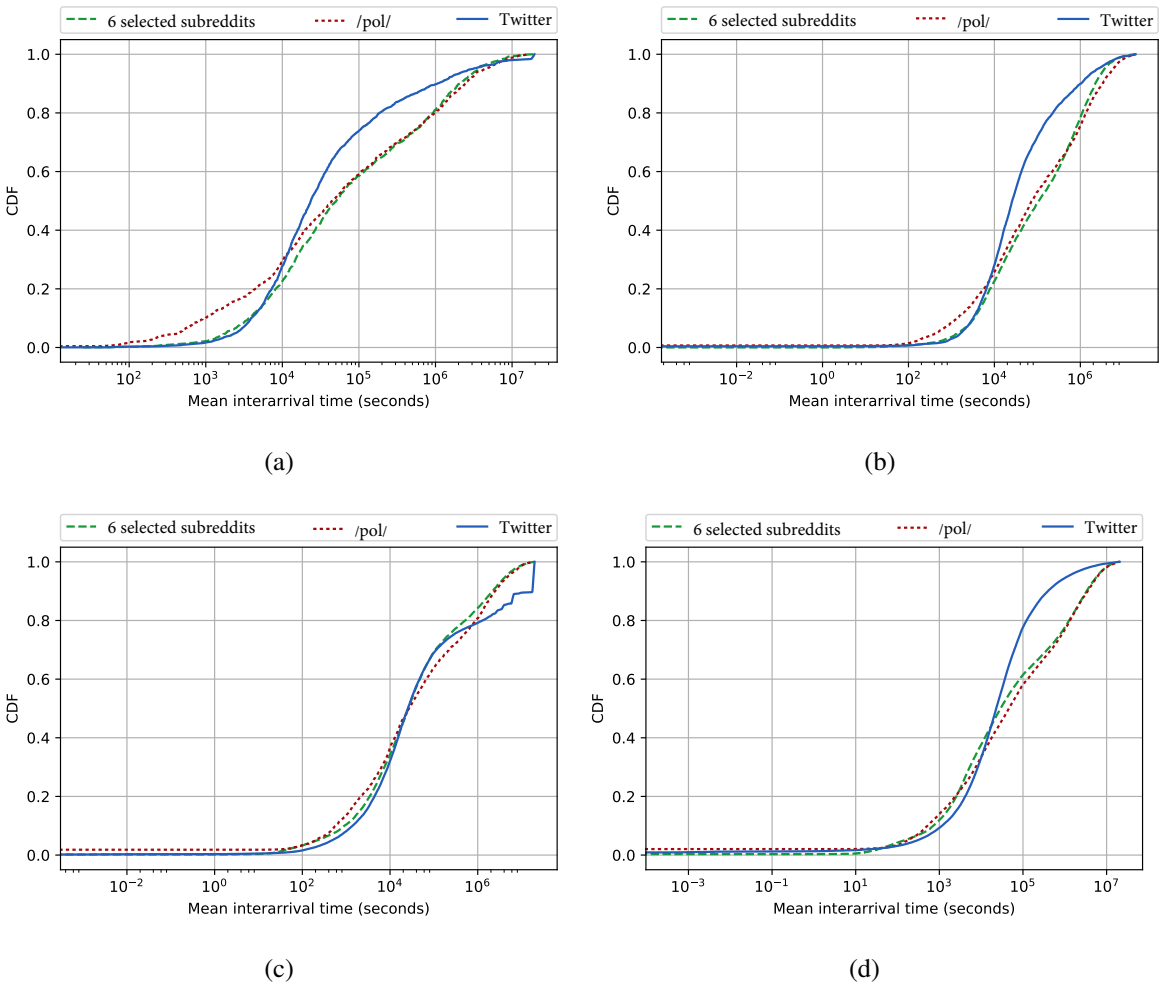


Figure 4.6: CDF for mean inter-arrival time for the URLs that occur more than once for (a) common alternative news URLs; (b) common mainstream news URLs; (c) all alternative news URLs, and (d) all mainstream news URLs.

small inter-arrival time, it follows the faster pace of Twitter, whereas, for URLs with longer inter-arrival times, it follows /pol/.

Cross Platform Analysis

We now look at URLs that appear on more than one platform and study the time at which they are shared. Fig. 4.7 plots the CDF of the time difference (in seconds) between the first occurrence of a URL on pairs of platforms, while Table 4.8 reports the numbers of URLs involved in each comparison.

We make the following observations: first, when comparing pairs of distributions for a given

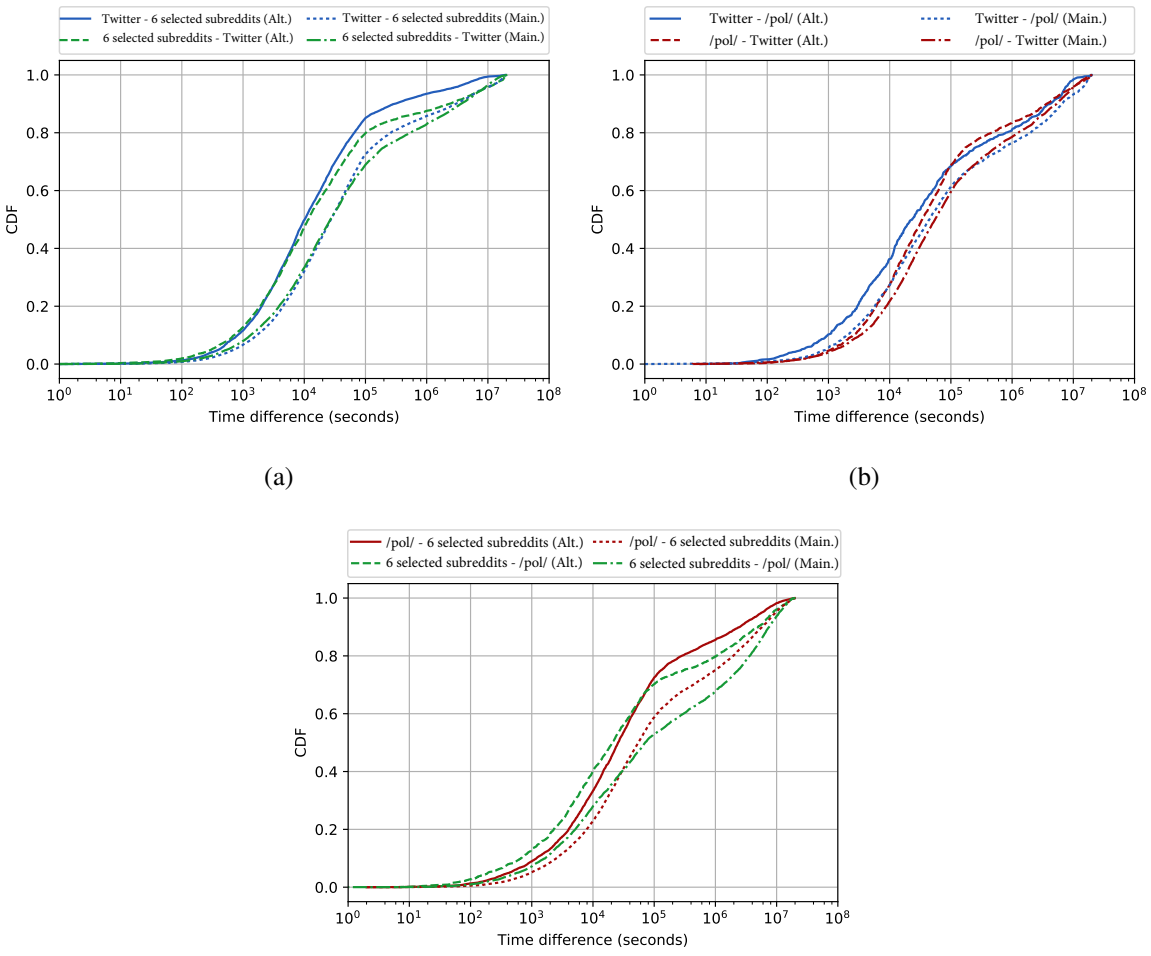


Figure 4.7: CDF of the difference between the first occurrence of a URL between (a) six selected subreddits and Twitter, (b) /pol/ and Twitter, and (c) /pol/ and six selected subreddits.

category of URLs, they are statistically different (a two sample Kolmogorov-Smirnov test rejects the null hypothesis with $p < 10^{-4}$). Second, alternative news appear on multiple platforms faster than mainstream news. This is consistent regardless of the pair of platforms we consider, and the sequence of appearances (i.e., first in platform A and then B, vs. first in B and then in A). Third, we notice the presence of a "turning point" with respect to the delay between URL appearance on each platform, which seems to be consistent across all pairs of platforms and types of news, and matches the 24h period observed earlier. Finally, there is a cross point when comparing URLs first posted on platform A and then on B, and URLs which were posted first in B and then on A (i.e., when the lines for the same type of URLs cross). Such a point represents which portion of URLs appear faster in one platform than the other. For the Twitter-six selected subreddits comparison, alternative (mainstream)

Comparison	Type of News	#URLs where platform 1 is faster	#URLs where platform 2 is faster
Reddit vs Twitter	Mainstream	18,762	11,416
	Alternative	5,232	4,301
/pol/ vs Twitter	Mainstream	2,938	4,700
	Alternative	778	2,099
/pol/ vs Reddit	Mainstream	5,382	14,662
	Alternative	1,455	3,695

Table 4.8: Statistics of URLs for the comparisons of time difference between platforms. Reddit refers to the six selected subreddits.

news appear faster on Twitter than the six subreddits 80% of the time (50%), with these URLs exhibiting slower propagation, since the turning point is at $\sim$1 hour (5 hours). Similarly, for the Twitter-/pol/ comparison, alternative (mainstream) news appear faster on Twitter than /pol/ 70% (5%) of the time, with the turning point at 1 day (2 days). Finally, for the six selected subreddits-/pol/ comparison, alternative (mainstream) news appear faster on the six subreddits than /pol/ for 65% (40%) of the time, with the turning point around 18 hours (12 hours).

Next, given the set of unique URLs across all platforms and the time they appear for the first time, we analyze their appearance in one, two, or three platforms, and the order in which this happens. For each URL, we find the first occurrence on each platform and build corresponding "sequences," e.g., if a URL first appears on the six subreddits (Reddit) and subsequently on /pol/ (4chan), the sequence is Reddit$\rightarrow$ 4chan (R$\rightarrow$4). Table 4.9 reports the distribution of the sequences of appearances considering only the first hop, i.e., up to the first two platforms in the sequence. The majority of URLs only appear on one platform: 82% of alternative URLs and 89% of mainstream news URLs. Also, both alternative and mainstream news URLs tend to appear on the six subreddits first and later appear on either Twitter or /pol/, and on Twitter before /pol/.

We also study the temporal dynamics of URLs that appear on all three platforms, with triplets of sequences. Table 4.10 reports the distribution of these sequences. The most common sequences are similar for both alternative and mainstream news URLs: R$\rightarrow$T$\rightarrow$4, R$\rightarrow$4$\rightarrow$T, and T$\rightarrow$R$\rightarrow$4 are the top three sequences. As already mentioned, the six selected subreddits "outperform" both other platforms in terms of the speed of sharing mainstream and alternative news URLs, as evidenced by the fact that it is at the head of the sequence for 51% and 59% of alternative and mainstream news URLs, respectively.

Sequence	Alternative (%)	Mainstream (%)
4 only	3,236 (4.4%)	18,654 (3.7%)
4→R	1,118 (1.5%)	4,606 (0.9%)
4→T	315 (0.5%)	861 (0.17%)
R only	24,292 (33.3%)	230,602(46.1%)
R→4	2,181 (3.0%)	11,307 (2.3%)
R→T	4,769 (6.5%)	16,685 (3.35%)
T only	32,443 (44.5%)	204,836 (41%)
T→4	585 (0.8%)	1,345 (0.26%)
T→R	3,964 (5.5%)	10,640 (2.12%)

Table 4.9: Distribution of URLs according to the sequence of first appearance within platforms for all URLs, considering only the first hop. "4" stands for /pol/ (4chan), "R" for the six selected subreddits (Reddit), and "T" for Twitter.

Finally, we analyze the source of the URLs for each of the three platforms, as follows. We create two directed graphs, one for each type of news, $G = (V, E)$, where V represents alternative or mainstream domains, as well as the three platforms, and E the set of sequences that consider only the first-hop of the platforms. For example, if a `breitbart.com` URL appears first on Twitter and later on the six selected subreddits, we add an edge from `breitbart.com` to Twitter, and from Twitter to the six selected subreddits. We also add weights on these edges based on the number of such unique URLs. By examining the paths we can discern which domains' URLs tend to appear first on each of the platforms.

Fig. 4.8 shows the graphs built for alternative and mainstream domains. Comparing the thickness of the outgoing edges, one can see that `breitbart.com` URLs appear first in the six selected subreddits more often than on Twitter and more frequently than on /pol/. However, for other popular alternative domains, such as `infowars.com`, `rt.com`, and `sputniknews.com`, URLs appear first on Twitter more often than the six selected subreddits and /pol/. Also, /pol/ is rarely the platform where a URL first shows up. For the mainstream news domains, we note that URLs from `nytimes.com` and `cnn.com` tend to appear first more often on the selected subreddits than Twitter and /pol/, however, URLs from other domains like `bbc.com` and `theguardian.com` tend to appear first more often on Twitter than the selected subreddits. Similar to the alternative domains graph, there is no domain where /pol/ dominates in terms of first URL appearance.

Sequence	Alternative (%)		Mainstream (%)	
4→R→T	128	(5.5%)	552	(8.9%)
4→T→R	145	(6.2%)	290	(4.7%)
R→4→T	335	(14.4%)	1,525	(24.5%)
R→T→4	841	(36.3%)	2,189	(35.3%)
T→4→R	192	(8.2%)	486	(7.8%)
T→R→4	673	(29%)	1,166	(18.8%)

Table 4.10: Distribution of URLs according to the sequence of first appearance within a platform for URLs common to all platforms. "4" stands for /pol/ (4chan), "R" for the six selected subreddits (Reddit), and "T" for Twitter.

4.1.5 Influence Estimation

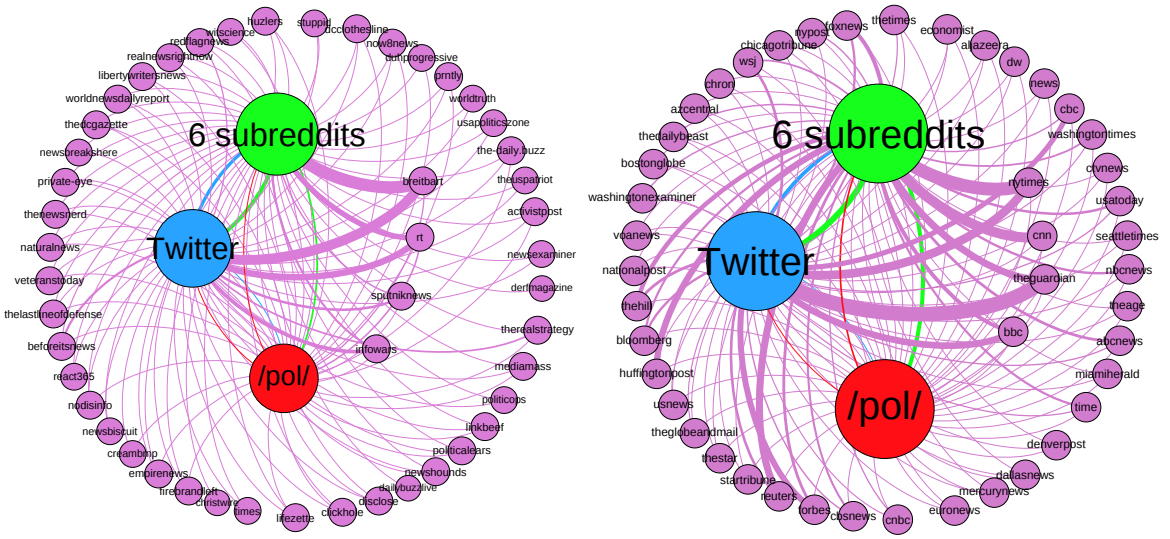
Thus far, our measurements have shown relative differences in how news media is shared on Reddit, Twitter, and 4chan. In this section, we provide meaningful evidence of how the individual platforms influence the media shared on other platforms. We do so by using a mathematical technique known as Hawkes processes. These statistical models can be used for modeling the dissemination of information in Web communities [136] as well as measuring social influence [240]. For more details regarding the Hawkes Processes and the general methodology used we refer the interested reader to Section 2.2.

Methodology

We now provide more details about our experiments, once again, considering 4chan (/pol/), Twitter, and the six subreddits. We study Hawkes processes at the subreddit granularity to get a better understanding of the various platforms and particular subreddits.

We aim to examine how these platforms and subreddits influence each other, so we model the arrival of URLs, in posts or tweets, with a Hawkes model with $K = 8$ point processes—one for Twitter, one for /pol/, and one for each of the subreddits. The model is fully connected, i.e., it is possible for each process to influence all the others, as well as itself, which describes behavior where participants on a platform see a URL and re-post it on the same platform.

We select URLs that have at least one event in Twitter, /pol/, and at least one of the subreddits, and we model each URL individually. The missing Twitter data affects 3,1K (37%) URLs. One way to mitigate the impact of this missing data is to remove events for which it has a



(a) (b)

Figure 4.8: Graph representation of news ecosystem (a) alternative news domains and (b) mainstream news domains. Edges are colored the same as their source node.

larger impact. E.g., if an event spans 100 days, the missing Twitter data has less of an effect than if the event only spanned two days. Thus, we examine URLs from other platforms that overlap with any of the missing days and remove the 10% of URLs (895) with the shortest total duration from the first event recorded until the last event recorded. This results in the missing data making up a smaller portion of the overall duration of the events.

		The_Donald	worldnews	politics	news	conspiracy	AskReddit	/pol/	Twitter
URLs	Mainstream	3,097	2,523	3,578	2,584	907	841	5,589	5,589
	Alternative	2,008	252	813	362	321	100	2,136	2,136
	Total	5,105	2,775	4,391	2,946	1,228	941	7,725	7,725
Events	Mainstream	12,312	7,517	26,160	5,794	1,995	2,302	19,746	36,250
	Alternative	7,797	458	2,484	586	497	176	7,322	23,172
	Total	20,109	7,975	28,644	6,380	2,492	2,478	27,068	59,422
Mean λ_0	Mainstream	0.001502	0.001382	0.001265	0.001392	0.000501	0.000107	0.001564	0.002330
	Alternative	0.001627	0.000619	0.000696	0.000553	0.000423	0.000034	0.001525	0.002803

Table 4.11: Total URLs with at least one event in Twitter, /pol/, and at least one of the subreddits; total events for mainstream and alternative URLs, and the mean background rate (λ_0) for each platform/subreddit.

The number of remaining URLs and events included for each platform are shown in Table 6.15. Next, we fit a Hawkes model for each URL and calculate the influence results using the approach described in Section 2.2.

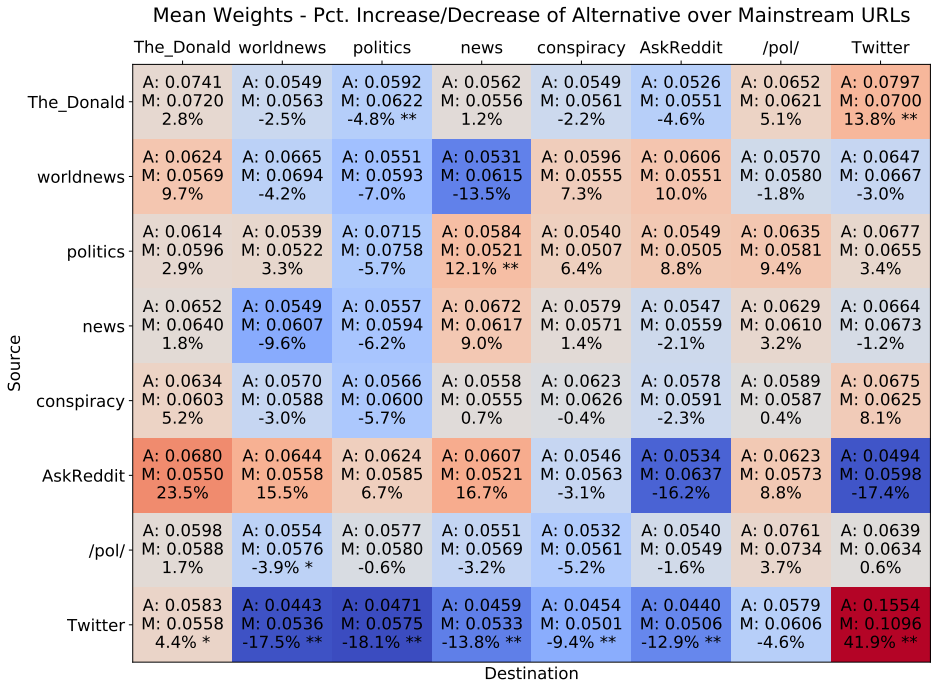


Figure 4.9: The mean weights for alternative URLs (A), the mean weights for mainstream URLs (M), and the percent increase/decrease between mainstream and alternative (also indicated by the coloration). The stars on the cells indicate the level of statistical significance (p-value) between the weight distributions: no stars indicate no statistical significance, whereas * and ** indicate statistical significance with $p < 0.05$ and $p < 0.01$ respectively.

Results

Looking at the number of URLs in Table 4.11, we note that there are substantially more events for mainstream than alternative news URLs. However, for Twitter, /pol/, and The_Donald, the ratios of events to URLs for alternative news URLs are similar to or greater than the ratios for mainstream ones. These high ratios explain the high background rates (see Table 4.11) for alternative news URLs for these platforms despite the lower number of events.

From the Hawkes models for each URL, we obtain the weight matrix W which specifies the strength of the connections between the different platforms and subreddits. The mean weight values over all URLs for alternative and mainstream news URLs, as well as the percentage difference between them are presented in Fig. 4.9. First, we look at Twitter. Background rates are high for both mainstream and alternative news URLs, which is not surprising given the large number of users on the platform. The values for $W_{\text{Twitter}\to\text{Twitter}}$ are also substantially higher than all other weights: 0.1096 for mainstream news URLs and 0.1554 for alternative news URLs. This reflects the ease and common practice of re-tweeting: a URL in a tweet is

likely to generate other events as users re-tweet it. There are different possible explanations for why the Twitter to Twitter rate for alternative news URLs is much greater than the rate for mainstream news URLs. The first is bot activity—if automated Twitter bots are used to spread alternative news URLs, it could result in a much higher rate of tweeting and re-tweeting. Another possible explanation is the behavior of users who read news stories from alternative sources; they might be more inclined to re-tweet the URL [229].

Looking at the weights for Twitter to the other platforms, except The_Donald, they are all greater for mainstream news URLs, meaning that the average tweet containing a mainstream URL is more likely to cause a subsequent post on the other platforms than the average tweet containing an alternative URL. The next communities most likely to cause events on others are The_Donald and /pol/. It is worth noting that The_Donald is the only platform/subreddit that has greater alternative URL weights for all of its inputs. Assuming that the population of The_Donald users that also read, say, worldnews is the same for both alternative and mainstream news URLs—which is reasonable—then the difference in weights implies that the users have a stronger preference for re-posting alternative news URLs back to The_Donald than for mainstream news URLs. The opposite can be seen for worldnews and politics, where most of the input weights are stronger for mainstream news. However, despite the higher weights for alternative news URLs, The_Donald is also, interestingly, influenced more strongly by mainstream news URLs than alternative news URLs on all platforms, with the exception of Twitter. This is in part because of the greater number of mainstream URL events, but The_Donald also has a higher background rate for alternative news URLs than mainstream news URLs, which implies that a lot of the alternative news URLs on the platform are coming from other sources.

To assess the statistical significance of the results, we perform two sample Kolmogorov-Smirnov tests on the weight distributions of mainstream and alternative news URLs for each source-destination pair (depicted as stars in Fig. 4.9). This allow us to assess whether the distributions of the weights for mainstream and alternative news URLs have statistically significant differences, hence indicating whether mainstream news URLs spread differently compared to alternative news URLs across the Web communities we study. Unsurprisingly, many of the source-destination pairs have no significant difference. However, in most cases where Twitter is the source community there *is* a significant statistical difference with $p < 0.01$. I.e., for some communities, Twitter is used not just to disseminate news, but to disseminate news from a specific *type* of source.

Fig. 4.10 illustrates the estimated total influence of the different platforms on each other, for

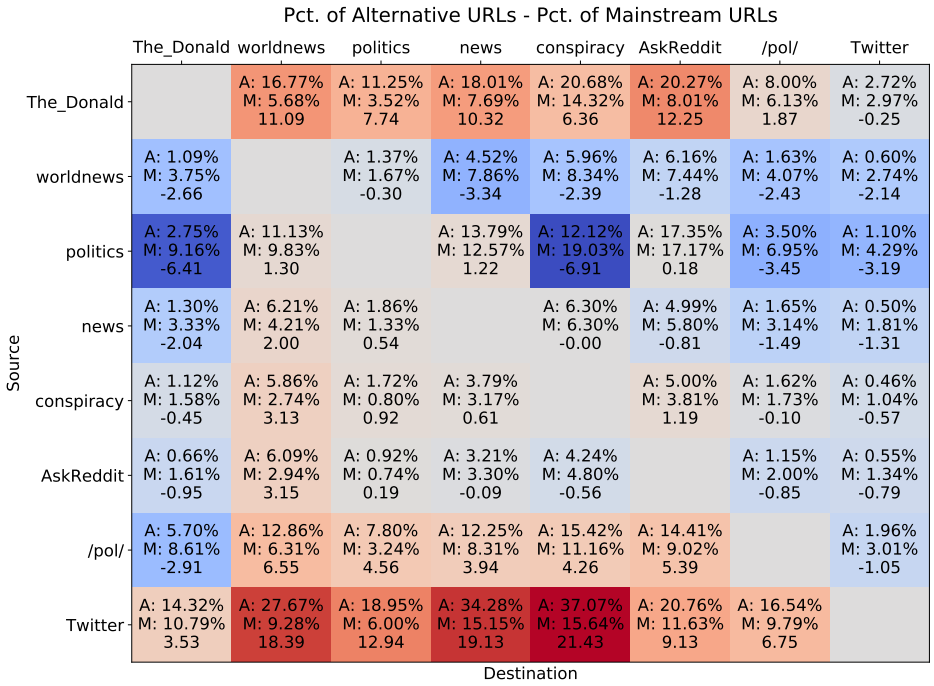


Figure 4.10: The estimated mean percentage of alternative URL events caused by alternative news URL events (A), the estimated mean percentage of mainstream news URL events caused by mainstream news URL events (M), and the difference between alternative and mainstream news (also indicated by the coloration).

both mainstream and alternative news URLs. Twitter contributes heavily to both types of events on the other platforms—and is in fact the most influential single source for most of the other platforms. Despite Twitter's lower weights for alternative news URLs, it actually has a greater influence on alternative than mainstream news URLs, in terms of percentage of events caused, on all the other platforms/subreddits. This is due to the fact that, even though it has lower weights, the largest proportion of alternative URL events are on Twitter. After Twitter, The_Donald and /pol/ also have a strong influence on the alternative news URLs that get posted on other platforms. The_Donald has a stronger effect for alternative news URLs on all platforms except Twitter—although it still has the largest alternative influence on Twitter, causing an estimated 2.72% of alternative news URLs tweeted. Interestingly, The_Donald causes 8% of /pol/'s alternative news URLs, while /pol/'s influence on The_Donald is less, at 5.7%. For the mainstream news URLs the strength of influence is reversed. Specifically, /pol/'s influence on The_Donald is 8.61% whereas The_Donald's influence on /pol/ is 6.13%.

In descending order, the influences on Twitter for mainstream news URLs are politics (4.29%), /pol/ (3.01%), The_Donald (2.97%), worldnews (2.74%), news (1.81%), AskReddit (1.34%), and conspiracy (1.04%). The strongest influences for alternative news URLs are, unsur-

prisingly, The_Donald (2.72%) and /pol/ (1.96%), followed by politics (1.10%), worldnews (0.60%), AskReddit (0.55%), news (0.50%), and conspiracy (0.46%). Twitter influences the alternative news URLs on other platforms to a large degree—but the largest alternative URL inputs to Twitter are The_Donald and /pol/. While we are only looking at a closed system of 8 different platforms and subreddits, we note that Twitter is undoubtedly effective at propagating information. Thus the influence these two communities have on Twitter is likely to have a disproportional impact on the greater Web compared to their relatively minuscule userbase.

4.1.6 Remarks

In this work, we explored how mainstream and fringe Web communities share mainstream and alternative news sources with a particular focus on how communities influence each other. We collected millions of posts from Twitter, Reddit, and 4chan, and analyzed the occurrence and temporal dynamics of news shared from 45 mainstream and 54 alternative news sites. We found that users on various platforms prefer distinct news sources, especially when it comes to alternative ones. We also explored complex temporal dynamics and we discovered, for example, that Twitter and Reddit users tend to post the same stories within a relatively short period of time, with 4chan posts lagging behind both of them. However, when a story becomes popular after a day or two, it is usually the case it was posted on 4chan first, lending some credence to 4chan's supposed influence on the Web.

Using Hawkes processes, we also modeled the influence the individual platforms have on each other, while also taking into account influence that comes from external sources of information. We found that the interplay between platforms manifests in subtle, yet meaningful ways. For example, of all the platforms and subreddits, Twitter by far has the most influence in terms of the number of URLs it causes to be posted to other platforms, and contributes to the share of alternative news URLs on the other platforms to a much greater degree than to the share of mainstream news URLs. After Twitter, The_Donald subreddit and /pol/ are the next most influential when it comes to alternative news URLs. For such URLs, The_Donald is less influenced by the other platforms than /pol/, and has a higher background rate, i.e., more of the URLs posted there come from other sources.

To the best of our knowledge, our analysis constitutes the first attempt to characterize the dissemination of mainstream and alternative news across multiple social media platforms, and to estimate a quantifiable influence between them. Overall, our findings shed light on how Web communities influence each other and can be extremely useful to better understand and

detect false information as well as informing the design of systems that aim to trace the origins of fake stories and mitigate their dissemination.

4.2 Detecting and Understanding the Spread of Memes Across Multiple Web Communities

4.2.1 Motivation

The Web has become one of the most impactful vehicles for the propagation of ideas and culture. Images, videos, and slogans are created and shared online at an unprecedented pace. Some of these, commonly referred to as *memes*, become viral, evolve, and eventually enter popular culture. The term "meme" was first coined by Richard Dawkins [241], who framed them as cultural analogues to genes, as they too self-replicate, mutate, and respond to selective pressures [242]. Numerous memes have become integral part of Internet culture, with well-known examples including the Trollface [243], Bad Luck Brian [244], and Rickroll [245].

While most memes are generally ironic in nature, used with no bad intentions, others have assumed negative and/or hateful connotations, including outright racist and aggressive undertones [246]. These memes, often generated by fringe communities, are being "weaponized" and even becoming part of political and ideological propaganda [247]. For example, memes were adopted by candidates during the 2016 US Presidential Elections as part of their iconography [248]; in October 2015, then-candidate Donald Trump retweeted an image depicting him as Pepe The Frog, a controversial character considered a hate symbol [249]. In this context, polarized communities within 4chan and Reddit have been working hard to create new memes and make them go viral, aiming to increase the visibility of their ideas—a phenomenon known as "attention hacking" [250].

Despite their increasingly relevant role, we have very little measurements and computational tools to understand the origins and the influence of memes. The online information ecosystem is very complex; social networks do not operate in a vacuum but rather influence each other as to how information spreads [66]. However, previous work has mostly focused on social networks in an isolated manner.

In this work, we aim to bridge these gaps by identifying and addressing a few research questions, which are oriented towards fringe Web communities: 1) How can we characterize memes, and how do they evolve and propagate? 2) Can we track meme propagation across

multiple communities and measure their influence? 3) How can we study variants of the same meme? 4) Can we characterize Web communities through the lens of memes?

Our work focuses on four Web communities: Twitter, Reddit, Gab, and 4chan's Politically Incorrect board (/pol/), because of their impact on the information ecosystem [66] and anecdotal evidence of them disseminating weaponized memes [251]. We design a processing pipeline and use it over 160M images posted between July 2016 and July 2017. Our pipeline relies on perceptual hashing (pHash) and clustering techniques; the former extracts representative feature vectors from the images encapsulating their visual peculiarities, while the latter allow us to detect groups of images that are part of the same meme. We design and implement a custom distance metric, based on both pHash and meme metadata, obtained from Know Your Meme (KYM), and use it to understand the interplay between the different memes. Finally, using Hawkes processes, we quantify the reciprocal influence of each Web community with respect to the dissemination of image-based memes.

Findings. Some of our findings (among others) include:

1. Our influence estimation analysis reveals that /pol/ and The_Donald are influential actors in the meme ecosystem, despite their modest size. We find that /pol/ substantially influences the meme ecosystem by posting a large number of memes, while The_Donald is the most *efficient* community in pushing memes to both fringe and mainstream Web communities.
2. Communities within 4chan, Reddit, and Gab use memes to share hateful and racist content. For instance, among the most popular cluster of memes, we find variants of the anti-semitic "Happy Merchant" meme [252] and the controversial Pepe the Frog [253].
3. Our custom distance metric effectively reveals the phylogenetic relationships of clusters of images. This is evident from the graph that shows the clusters obtained from /pol/, Reddit's The_Donald subreddit, and Gab available for exploration at [254].

Contributions. First, we develop a robust processing pipeline for detecting and tracking memes across multiple Web communities. Based on pHash and clustering algorithms, it supports large-scale measurements of meme ecosystems, while minimizing processing power and storage requirements. Second, we introduce a custom distance metric, geared to highlight hidden correlations between memes and better understand the interplay and overlap between them. Third, we provide a characterization of multiple Web communities (Twitter, Reddit, Gab, and /pol/) with respect to the memes they share, and an analysis of their reciprocal influence using the Hawkes Processes statistical model. Finally, we release our processing pipeline and

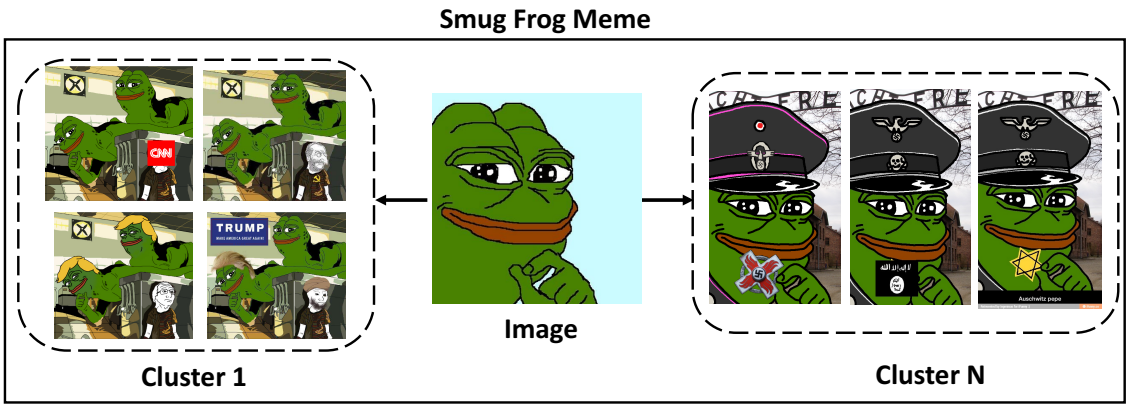


Figure 4.11: An example of a meme (Smug Frog) that provides an intuition of what an image, a cluster, and a meme is.

datasets[3], in the hope to support further measurements in this space.

4.2.2 Methodology

In this section, we present our methodology for measuring the propagation of memes across Web communities.

Overview

Memes are high-level concepts or ideas that spread within a culture [241]. In Internet vernacular, a *meme* usually refers to variants of a particular image, video, cliché, etc. that share a common theme and are disseminated by a large number of users. In this thesis, we focus on their most common incarnation: *static images*.

To gain an understanding of how memes propagate across the Web, with a particular focus on discovering the communities that are most influential in spreading them, our intuition is to build *clusters* of visually similar images, allowing us to track variants of a meme. We then group clusters that belong to the same meme to study and track the meme itself. In Figure 4.11, we provide a visual representation of the Smug Frog meme [255], which includes many variants of the same image (a "smug" Pepe the Frog) and several clusters of variants. Cluster 1 has variants from a Jurassic Park scene, where one of the characters is hiding from two velociraptors behind a kitchen counter: the frogs are stylized to look similar to velociraptors, and the character hiding varies to express a particular message. For example, in the image in

[3] https://github.com/memespaper/memes_pipeline

the top right corner, the two frogs are searching for an anti-semitic caricature of a Jew (itself a meme known as the Happy Merchant [252]). Cluster N shows variants of the smug frog wearing a Nazi officer military cap with a photograph of the infamous "Arbeit macht frei" slogan from the distinctive curved gates of Auschwitz in the background. In particular, the two variants on the right display the death's head logo of the SS-Totenkopfverbände organization responsible for running the concentration camps during World War II. Overall, these clusters represent the branching nature of memes: as a new variant of a meme becomes prevalent, it often branches into its own sub-meme, potentially incorporating imagery from other memes.

Processing Pipeline

Our processing pipeline is depicted in Figure 4.12. As discussed above, our methodology aims at identifying clusters of similar images and assign them to higher level groups, which are the actual memes. Note that the proposed pipeline is not limited to image macros and can be used to identify any image. We first discuss the types of data sources needed for our approach, i.e., meme annotation sites and Web communities that post memes (dotted rounded rectangles in the figure). Then, we describe each of the operations performed by our pipeline (Steps 1-7, see regular rectangles).

Data Sources. Our pipeline uses two types of data sources: 1) sites providing meme annotation and 2) Web communities that disseminate memes. In this thesis, we use Know Your Meme for the former, and Twitter, Reddit, /pol/, and Gab for the latter. We provide more details about our datasets in Section 4.2.3. Note that our methodology supports any annotation site and any Web community, and this is why we add the "*Generic*" sites/communities notation in Figure 4.12.

pHash Extraction (Step 1). We use the Perceptual Hashing (pHash) algorithm [256] to calculate a fingerprint of each image in such a way that any two images that look similar to the human eye map to a "similar" hash value. pHash generates a feature vector of 64 elements that describe an image, computed from the Discrete Cosine Transform among the different frequency domains of the image. Thus, visually similar images have minor differences in their vectors, hence allowing to search for and detect visually similar images. For example, the string representation of the pHashes obtained from the images in cluster N (see Figure 4.11) are 55352b0b8d8b5b53, 55952b0bb58b5353, and 55952b2b9da58a53, respectively. The algorithm is also robust against changes in the images, e.g., signal processing operations and direct manipulation [257], and effectively reduces the dimensionality of the raw images.

Clustering via pairwise distance calculation (Steps 2-3). Next, we cluster images from one

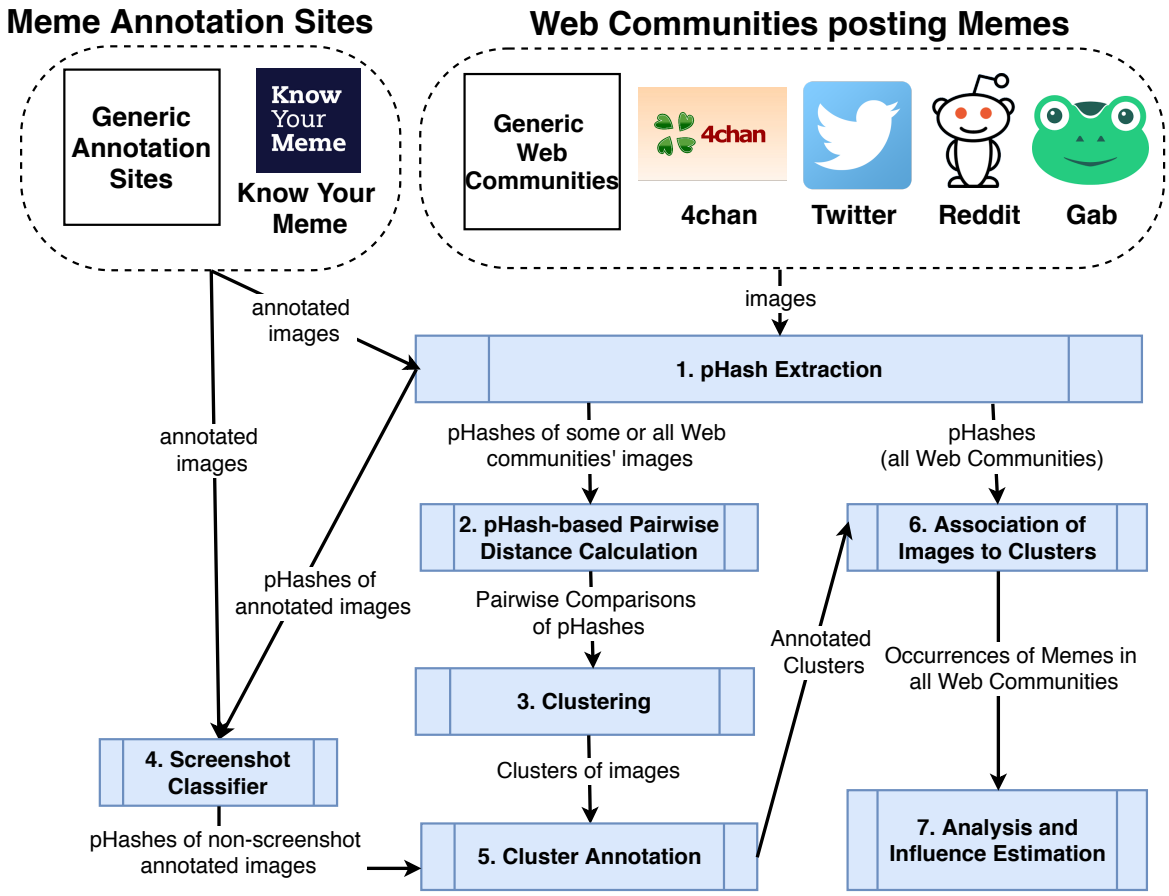


Figure 4.12: High-level overview of our processing pipeline.

or more Web Communities using the pHash values. We perform a pairwise comparison of all the pHashes using Hamming distance (Step 2). To support large numbers of images, we implement a highly parallelizable system on top of TensorFlow [258], which uses multiple GPUs to enhance performance. Images are clustered using a density-based algorithm (Step 3). Our current implementation uses DBSCAN [16], mainly because it can discover clusters of arbitrary shape and performs well over large, noisy datasets. Nonetheless, our architecture can be easily tweaked to support any clustering algorithm and distance metric.

We also perform an analysis of the clustering performance and the rationale for selecting the clustering threshold. Our implementation uses the DBSCAN algorithm with a clustering threshold equal to 8. To select this threshold, we perform the clustering step while varying the distances. Table 4.12 shows the number of clusters and the percentage of images that are regarded as noise by the clustering algorithm for varying distances. We observe that, for distances 2-4, we have a substantially larger percentage of noise, while with distance 10 we have the least percentage of noise. With distances between 6 and 8 we observe that we get a

Distance	#Clusters	%Noise
2	30,327	82.9%
4	34,146	78.5%
6	37,292	73.0%
8	38,851	62.8%
10	30,737	27.8%

Table 4.12: Number of clusters and percentage of noise for varying clustering distances.

larger number of clusters than the other distances, while the noise percentages are 73% and 63%, respectively.

To further evaluate the clustering performance for varying distances, we randomly select 200 clusters and manually calculate the number of images that are false positives within each cluster. Figure 4.13 shows the CDF of the false positive fraction in the random sample of clusters for distances 6, 8, and 10 (we disregard distances 2-4 due to the high percentage of noise). Distance 10 yields a high number of false positives, while distances 6-8 the overall false positives are below 3%. Therefore, we investigate the impact of these false positives in the overall dataset, looking at all posts that contain false and true positives in the random sample of 200 clusters, using distance 8. We find that the false positives have little impact as they occur substantially fewer times than true positives: the percentage of true positives over the set of false positives and true positives is 99.4%. Thus, due to the larger number of clusters, the acceptable false positive performance, and the smaller percentage of noise (when compared to distances 2-6), we elect to use as a threshold the perceptual distance that is equal to 8.

Screenshots Removal (Step 4). Meme annotation sites like KYM often include, in their image galleries, screenshots of social network posts that are not variants of a meme but just comments about it. Hence, we discard social-network screenshots from the annotation sites data sources using a deep learning classifier. Below, we provide more details about our screenshot removal classifier.

Dataset. Table 4.13 summarizes the dataset used for training the classifier. It includes 28.8K images that depict posts from Twitter, 4chan, Reddit, Facebook, and Instagram, which we collect from public sources. First, we download images from specific subreddits that only allow screenshots from a particular community. For example, the 4chan subreddit require all

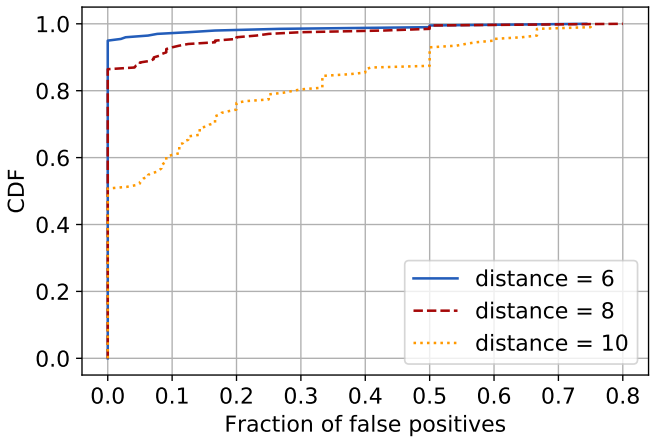


Figure 4.13: Fraction of false positives in clusters with varying clustering distance.

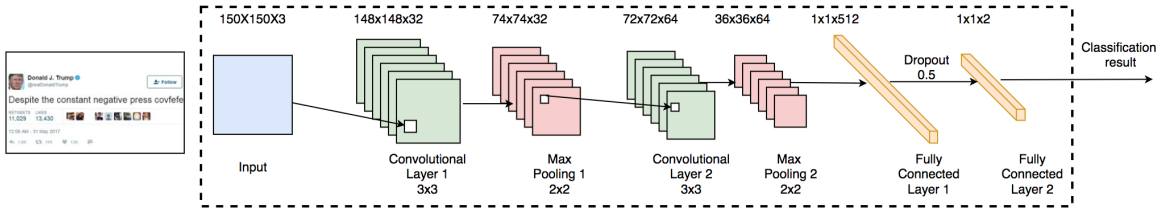


Figure 4.14: Architecture of the deep learning model for detecting screenshots from Twitter, /pol/, Reddit, Instagram, and Facebook.

submissions to be of a screenshot of a 4chan thread. Next, we use the Pinterest platform to download specific boards that contain mostly screenshots from the communities we study. Also, we search and obtain image datasets that are publicly available on Web archiving services like the Wayback Machine. We then manually filter out images that were misplaced. Finally, we include 10K random images posted on /pol/ (i.e., a subset of the 4.3M images collected for our measurements).

Classifier. To detect screenshots that contain images from one of the social networks included in our dataset, we use Convolutional Neural Networks. Figure 4.14 provides an overview of our classifier's architecture. It includes two Convolutional Neural Networks, each followed by a max-pooling layer. The output of these layers is fed to a fully-connected dense layer comprising 512 units. Finally, we have another fully-connected layer with two units, which outputs the probability that a particular image is a screenshot from one of the five social networks and the probability that an image is a random one. To avoid overfitting on the two last fully-connected layers, we apply Dropout with $d = 0.5$ [259]. This means that, while

Platform	Twitter	4chan	Reddit	Facebook	Instagram	Other
# images	14,602	10,127	2,181	1,414	497	10,630

Table 4.13: Curated dataset used to train the screenshot classifier.

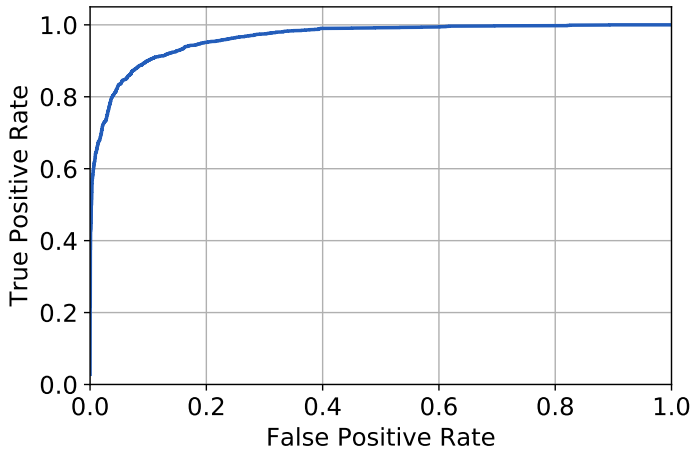


Figure 4.15: ROC curve of the screenshot classifier.

training, 50% of the units are randomly omitted from updating their parameters.

Experimental Evaluation. Our implementation uses Keras [260] with TensorFlow as the backend [258]. To train our model, we randomly select 80% of the images and evaluate based on the rest 20% out-of-sample dataset. Figure 4.15 shows the ROC curve of the model. We observe that the devised classifier exhibits acceptable performance with an Area Under the Curve (AUC) of 0.96. We also evaluate our model in terms of accuracy, precision, recall, and F1-score, which amount to 91.3%, 94.3%, 93.5%, and 93.9%, respectively.

Cluster Annotation (Steps 5). Clustering annotation uses the *medoid* of each cluster, i.e., the element with the minimum square average distance from all images in the cluster. In other words, the medoid is the image that best represents the cluster. The clusters' medoids are compared with all images from meme annotation sites, by calculating the Hamming distance between each pair of pHash vectors. We consider that an image matches a cluster if the distance is less than or equal to a threshold θ, which we set to 8, as it allows us to capture the diversity of images that are part of the same meme while maintaining a low number of false positives.

As the annotation process considers all the images of a KYM entry's image gallery, it is likely we will get multiple annotations for a single cluster. To find the representative KYM entry for

each cluster, we select the one with the largest proportion of matches of KYM images with the cluster medoid. In case of ties, we select the one with the minimum average Hamming distance.

While KYM might not be a household name, the site is seemingly the largest curated collection of memes on the Web, i.e., KYM is as close to an "authority" on memes as there is. That said, crowdsourcing *is* an aspect of how KYM works, and thus there might be questions as to how "legitimate" some of the content is. To this end, we set out to measure the quality of KYM by sampling a number of pages and manually examining them. This is clearly a subjective task, and a fully specified definition of what makes a valid meme is approximately as difficult as defining "art." Nevertheless, the authors of this work have, for better or worse, collectively spent thousands of hours immersed in the communities we explore; thus, while we are not confident in providing a strict definition of a meme, we are in claiming that we know a meme when we see it.

Using the same randomly selected 200 clusters as mentioned in Steps2-3 above, we visited each KYM page the cluster was tagged with and noted whether or not it properly documented what we consider an "actual" meme. The 200 clusters were mapped to 162 unique KYM pages, and of these 162 pages, 3 (1.85%) we decided were "bad." This is mainly due to the lack of completeness and relatively high number of random images in the gallery (see [261, 262] for some examples of "bad" KYM entries).

Next, we set out to determine whether the label (i.e., KYM page) assigned to each of our randomly sampled clusters was appropriate. Using three annotators, for each cluster we examined the KYM page, the medoid of the cluster, and the images in the cluster itself and noted whether the label does in fact apply to the cluster. Here, again, there is a great degree of subjectivity. To reign some of the subjectivity in, we used the following guidelines:

1. If the exact image(s) in the cluster appear in the KYM gallery, then the label is correct.
2. For images that do not appear in the KYM gallery, if the label is *appropriate*, then it is a correct labeling.

There are some important caveats with these guidelines. First, KYM galleries are crowdsourced, and while curated to some extent, the possibility for what amounts to random images in a gallery *does* exist; however, based on our assessment of KYM page validity, this occurs with low probability. Second, we considered a label correct if it was *appropriate*, even if it was not necessarily the *best* possible label. For example, as our results show, many memes are related, and many images mix and match pieces of various memes. While it is definitely

true that there might be better labels that exist for a given cluster, this straightforward and comprehensible labeling process is sufficient for our purposes. We leave a more in-depth study of the subjective nature of memes for future work. Finally, it is important to note that memes are a *cultural* phenomenon, and thus the potential for cultural bias in our annotation is possible. Note that our annotators were born in three different countries (USA, Italy, and Cyprus), only one is a native English speaker, and two have spent substantial time in the US.

After annotating clusters, we compute the Fleis agreement score (κ). With our cluster samples, we achieve κ=0.67, which is considered "substantial" agreement. Finally, for each cluster we obtain the majority agreement of all annotators to assess the accuracy of our annotation process; we find that 89% of the clusters had a legitimate annotation to a specific KYM entry.

Association of images to memes (Step 6). To associate images posted on Web communities (e.g., Twitter, Reddit, etc.) to memes, we compare them with the clusters' medoids, using the same threshold θ. This is conceptually similar to Step 5, but uses images from Web communities instead of images from annotation sites. This lets us identify memes posted in generic Web communities and collect relevant metadata from the posts (e.g., the timestamp of a tweet). Note that we track the propagation of memes in generic Web communities (e.g., Twitter) using a *seed* of memes obtained by clustering images from other (fringe) Web communities. More specifically, our seeds will be memes generated on three fringe Web communities (/pol/, The_Donald subreddit, Gab); nonetheless, our methodology can be applied to any community.

Analysis and Influence Estimation (Step 7). We analyze all relevant clusters and the occurrences of memes, aiming to assess: 1) their popularity and diversity in each community; 2) their temporal evolution; and 3) how communities influence each other with respect to meme dissemination.

Distance Metric

To better understand the interplay and connections between the clusters, we introduce a custom distance metric, which relies on both the visual peculiarities of the images (via pHash) and data available from annotation sites. The distance metric supports one of two modes: 1) one for when both clusters are annotated (*full-mode*), and 2) another for when one or none of the clusters is annotated (*partial-mode*).

Definition. Let c be a cluster of images and F a set of features extracted from the clusters.

The custom distance metric between cluster c_i and c_j is defined as:

$$\mathsf{distance}(c_i, c_j) = 1 - \sum_{\mathsf{f} \in \mathsf{F}} w_{\mathsf{f}} \times \mathtt{r}_{\mathsf{f}}(c_i, c_j) \tag{4.1}$$

where $\mathtt{r}_{\mathsf{f}}(c_i, c_j)$ denotes the similarity between the features of type $\mathsf{f} \in \mathsf{F}$ of cluster c_i and c_j, and w_f is a weight that represents the relevance of each feature. Note that $\sum_{\mathsf{f}} w_{\mathsf{f}} = 1$ and $\mathsf{r}_f(c_i, c_j) = \{x \in \mathbb{R} \mid 0 \leq x \leq 1\}$. Thus, $\mathsf{distance}(c_i, c_j)$ is a number between 0 and 1.

Features. We consider four different features for $\mathtt{r}_{\mathsf{f} \in \mathsf{F}}$, specifically, $\mathsf{F} = \{perceptual, meme, people, culture\}$; see below.

$r_{perceptual}$: this feature is the similarity between two clusters from a perceptual viewpoint. Let h be a pHash vector for an image m in cluster c, where m is the medoid of the cluster, and d_{ij} the Hamming distance between vectors h_i and h_j (see Step 5 in Section 4.2.2). We compute d_{ij} from c_i and c_j as follows. First, we obtain obtain the medoid m_i from cluster c_i. Subsequently, we obtain h_i=pHash(m_i). Finally, we compute d_{ij}=Hamming(h_i, h_j). We simplify notation and use d instead of d_{ij} to denote the distance between two medoid images and refer to this distance as the Hamming *score*.

We define the perceptual similarity between two clusters as an exponential decay function over the Hamming score d:

$$r_{perceptual}(d) = 1 - \frac{d}{\tau \times e^{\mathtt{max}/\tau}} \tag{4.2}$$

where $\mathtt{max}$ represents the maximum pHash distance between two images and τ is a constant parameter, or *smoother*, that controls how fast the exponential function decays for all values of d (recall that $\{d \in \mathbb{R} \mid 0 \leq d \leq \mathtt{max}\}$). Note that $\mathtt{max}$ is bound to the precision given by the pHash algorithm. Recall that each pHash has a size of $|d|$=64, hence $\mathtt{max}$=64. Intuitively, when $\tau << 64$, $r_{perceptual}$ is a high value only with perceptually indistinguishable images, e.g., for τ=1, two images with d=0 have a similarity $r_{perceptual}$=1.0. With the same τ, the similarity drops to 0.4 when d=1. By contrast, when τ is close to 64, $r_{perceptual}$ decays almost linearly. For example, for τ=64, $r_{perceptual}(d$=0)=1.0 and $r_{perceptual}(d$=1)=0.98. Figure 4.16 shows how $r_{perceptual}$ performs for different values of τ. As mentioned above, we observe that pairs of images with scores between d=0 and d=8 are usually part of the same variant (see Step 5 in Section 4.2.2). In our implementation, we set τ=25 as $r_{perceptual}$ returns high values up to d=8, and rapidly decays thereafter.

r_{meme}, $r_{culture}$, and r_{people}: the annotation process (Step 5) provides contextualized information

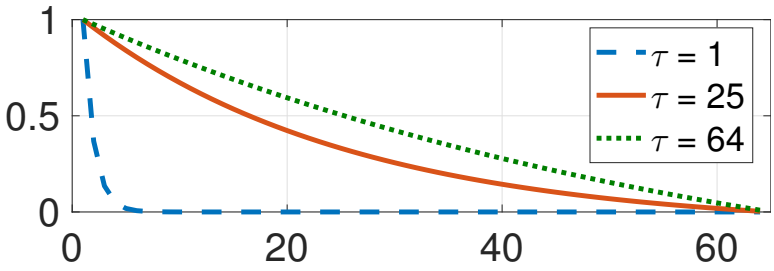


Figure 4.16: Different values of $r_{perceptual}$ (y-axis) for all possible inputs of d (x-axis) with respect to the smoother τ.

about the cluster medoid, including the name (i.e., the main identifier) given to a meme, the associated culture (i.e., high-level group of meme), and people that are included in a meme. Note that we use all the annotations for each category and not only the representative one (see Step 5). Therefore, we model a different similarity for each of the these categories, by looking at the overlap of all the annotations among the medoids of both clusters (m_i, m_j, for c_i and c_j, respectively). Specifically, for each category, we calculate the Jaccard index between the annotations of both medoids, for memes, cultures, and people, thus acquiring r_{meme}, $r_{culture}$, r_{people}, respectively.

Modes. Our distance metric measures how similar two clusters are. If both clusters are annotated, we operate in "full-mode," and in "partial-mode" otherwise. For each mode, we use different weights for the features in Eq. 4.1, which we set empirically as we lack the ground-truth data needed to automate the computation of the optimal set of thresholds.

Full-mode. In full-mode, we set weights as follows. 1) The features from the perceptual and meme categories should have higher relevance than people and culture, as they are intrinsically related to the definition of meme (see Section 4.2.2). The last two are non-discriminant features, yet are informative and should contribute to the metric. Also, 2) r_{meme} should not outweigh $r_{perceptual}$ because of the relevance that visual similarities have on the different variants of a meme. Likewise, $r_{perceptual}$ should not dominate over r_{meme} because of the branching nature of the memes. Thus, we want these two categories to play an equally important weight. Therefore, we choose $w_{perceptual}$=0.4, w_{meme}=0.4, w_{people}=0.1, $w_{culture}$=0.1.

This means that when two clusters belong to the same meme and their medoids are perceptually similar, the distance between the clusters will be small. In fact, it will be at most $0.2 = 1-(0.4+0.4)$ if people and culture do not match, and 0.0 if they also match. Note that our metric also assigns small distance values for the following two cases: 1) when two clusters

are part of the same meme variant, and 2) when two clusters use the same image for different memes.

Partial-mode. In this mode, we associate unannotated images with any of the known clusters. This is a critical component of our analysis (Step 6), allowing us to study images from generic Web communities where annotations are unavailable. In this case, we rely entirely on the perceptual features. We once again use Eq. 4.1, but simply set all weights to 0, except for $w_{perceptual}$ (which is set to 1). That is, we compare the image we want to test with the medoid of the cluster and we apply Eq. 4.2 as described above.

4.2.3 Datasets

We now present the datasets used in our measurements.

Web Communities

As mentioned earlier, our data sources are Web communities that post memes and meme annotation sites. For the former, we focus on four communities: Twitter, Reddit, Gab, and 4chan (more precisely, 4chan's Politically Incorrect board, /pol/). This provides a mix of mainstream social networks (Twitter and Reddit) as well as fringe communities that are often associated with the alt-right and have an impact on the information ecosystem (Gab and /pol/) [66].

There are several other platforms playing important roles in spreading memes, however, many are "closed" (e.g., Facebook) or do not involve memes based on static images (e.g., YouTube, Giphy). In future work, we plan to extend our measurements to communities like Instagram and Tumblr, as well as to GIF and video memes. Nonetheless, we believe our data sources already allow us to elicit comprehensive insights into the meme ecosystem.

Table 4.14 reports the number of posts and images processed for each community. Note that the number of images is lower than the number of posts with images because of duplicate image URLs and because some images get deleted. Next, we discuss each dataset.

Twitter. Our Twitter dataset is based on tweets made available via the 1% Streaming API, between July 1, 2016 and July 31, 2017. In total, we parse 1.4B tweets: 242M of them have at least one image. We extract all the images, ultimately collecting 114M images yielding 74M unique pHashes.

Platform	#Posts	#Posts with Images	#Images	#Unique pHashes
Twitter	1,469,582,378	242,723,732	114,459,736	74,234,065
Reddit	1,081,701,536	62,321,628	40,523,275	30,441,325
/pol/	48,725,043	13,190,390	4,325,648	3,626,184
Gab	12,395,575	955,440	235,222	193,783
KYM	15,584	15,584	706,940	597,060

Table 4.14: Overview of our datasets.

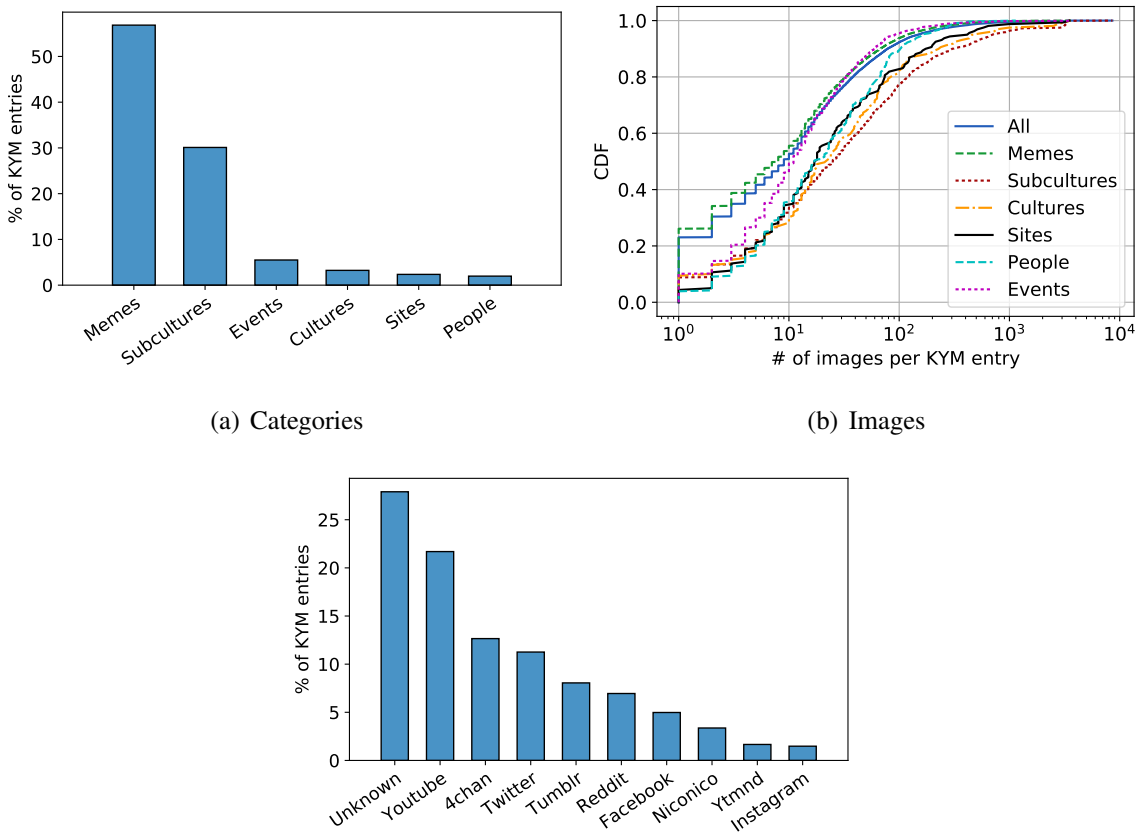
Reddit. We gather images from Reddit using publicly available data from Pushshift [263]. We parse all submissions and comments between July 1, 2016 and July, 31 2017, and extract 62M posts that contain at least one image. We then download 40M images producing 30M unique pHashes.

4chan. We obtain all threads posted on /pol/, between July 1, 2016 and July 31, 2017, using the same methodology of [19]. Since all threads (and images) are removed after a week, we use a public archive service called 4plebs [264] to collect 4.3M images, thus yielding 3.6M unique pHashes.

Gab. We collect 12M posts, posted on Gab between August 10, 2016 and July 31, 2017, and 955K posts have at least one image, using the same methodology as in [43]. Out of these, 235K images are unique, producing 193K unique pHashes. Note that our Gab dataset starts one month later than the other ones, since Gab was launched in August 2016.

Meme Annotation Site

Know Your Meme (KYM). We choose KYM as the source for meme annotation as it offers a comprehensive database of memes. KYM is a sort of encyclopedia of Internet memes: for each meme, it provides information such as its origin (i.e., the platform on which it was first observed), the year it started, as well as descriptions and examples. In addition, for each entry, KYM provides a set of keywords, called *tags*, that describe the entry. Also, KYM provides a variety of higher-level categories that group meme entries; namely, cultures, subcultures, people, events, and sites. "Cultures" and "subcultures" entries refer to a wide variety of topics ranging from video games to various general categories. For example, the Rage Comics *subculture* [265] is a higher level category associated with memes related to comics like Rage



(a) Categories

(b) Images

(c) Origins

Figure 4.17: Basic statistics of the KYM dataset.

Guy [266] or LOL Guy [267], while the Alt-right *culture* [268] gathers entries from a loosely defined segment of the right-wing community. The rest of the categories refer to specific individuals (e.g., Donald Trump [269]), specific *events (e.g.,#CNNBlackmail [270]), and* sites (e.g., /pol/ [271]), respectively. It is also worth noting that KYM moderates all entries, hence entries that are wrong or incomplete are marked as so by the site.

As of May 2018, the site has 18.3K entries, specifically, 14K memes, 1.3K subcultures, 1.2K people, 1.3K events, and 427 websites [272]. We crawl KYM between October and December 2017, acquiring data for 15.6K entries; for each entry, we also download all the images related to it by crawling all the pages of the image gallery. In total, we collect 707K images corresponding to 597K unique pHashes. Note that we obtain 15.6K out of 18.3K entries, as we crawled the site several months before May 2018.

Getting to know KYM. We also perform a general characterization of KYM. First, we look at the distribution of entries across categories: as shown in Figure 4.17(a), as expected, the

majority (57%) are memes, followed by subcultures (30%), cultures (3%), websites (2%), and people (2%).

Next, we measure the number of images per entry: as shown in Figure 4.17(b), this varies considerably (note log-scale on x-axis). KYM entries have as few as 1 and as many as 8K images, with an average of 45 and a median of 9 images. Larger values may be related to the meme's popularity, but also to the "diversity" of image variants it generates. Upon manual inspection, we find that the presence of a large number of images for the same meme happens either when images are visually very similar to each other (e.g., Smug Frog images *within* the two clusters in Figure 4.11), or if there are actually remarkably different variants of the same meme (e.g., images in 'cluster 1' *vs.* images in 'cluster N' in the same figure). We also note that the distribution varies according to the category: e.g., higher-level concepts like cultures include more images than more specific entries like memes.

We then analyze the origin of each entry: see Figure 4.17(c). Note that a large portion of the memes (28%) have an unknown origin, while YouTube, 4chan, and Twitter are the most popular platforms with, respectively, 21%, 12%, and 11%, followed by Tumblr and Reddit with 8% and 7%. This confirms our intuition that 4chan, Twitter, and Reddit, which are among our data sources, play an important role in the generation and dissemination of memes. As mentioned, we do not currently study video memes originating from YouTube, due to the inherent complexity of video-processing tasks as well as scalability issues. However, a large portion of YouTube memes actually end up being morphed into image-based memes (see, e.g., the Overly Attached Girlfriend meme [273]).

Running the pipeline on our datasets

For all four Web communities (Twitter, Reddit, /pol/, and Gab), we perform Step 1 of the pipeline (Figure 4.12), using the ImageHash library.[4] After computing the pHashes, we delete the images (i.e., we only keep the associated URL and pHash) due to space limitations of our infrastructure. We then perform Steps 2-3 (i.e., pairwise comparisons between all images and clustering), for all the images from /pol/, The_Donald subreddit, and Gab, as we treat them as fringe Web communities. Note that, we exclude mainstream communities like the rest of Reddit and Twitter as our main goal is to obtain clusters of memes from fringe Web communities and later characterize all communities by means of the clusters. Next, we go through Steps 4-5 using all the images obtained from meme annotation websites (specifically,

[4] https://github.com/JohannesBuchner/imagehash

Platform	#Images	Noise	#Clusters	#Clusters with KYM tags (%)
/pol/	4,325,648	63%	38,851	9,265 (24%)
T_D	1,234,940	64%	21,917	2,902 (13%)
Gab	235,222	69%	3,083	447 (15%)

Table 4.15: Statistics obtained from clustering images from /pol/, The_Donald, and Gab.

Know Your Meme, see Section 4.2.3) and the medoid of each cluster from /pol/, The_Donald, and Gab. Finally, Steps 6-7 use all the pHashes obtained from Twitter, Reddit (all subreddits), /pol/, and Gab to find posts with images matching the annotated clusters. This is an integral part of our process as it allows to characterize and study mainstream Web communities not used for clustering (i.e., Twitter and Reddit).

4.2.4 Analysis

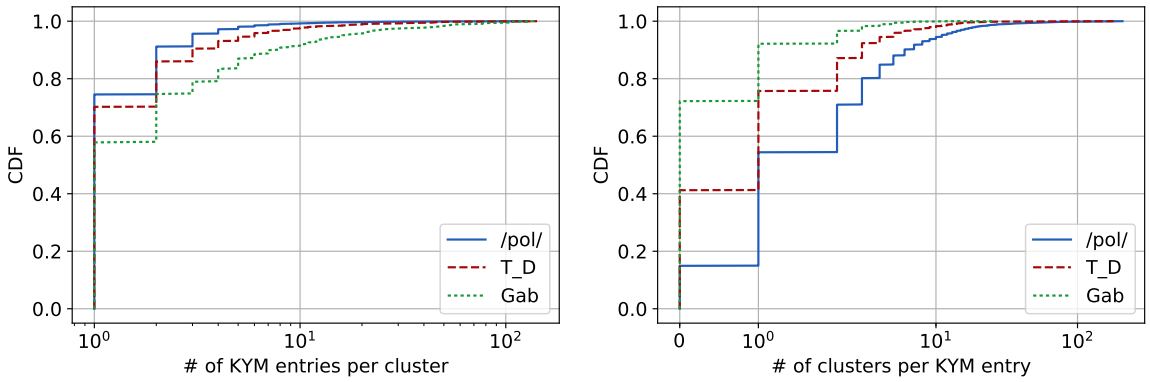
In this section, we present a cluster-based measurement of memes and an analysis of a few Web communities from the "perspective" of memes. We measure the prevalence of memes across the clusters obtained from fringe communities: /pol/, The_Donald subreddit (T_D), and Gab. We also use the distance metric introduced in Eq. 4.1 to perform a *cross-community* analysis, then, we group clusters into broad, but related, categories to gain a macro-perspective understanding of larger communities, including Reddit and Twitter.

Cluster-based Analysis

We start by analyzing the 12.6K annotated clusters consisting of 268K images from /pol/, The_Donald, and Gab (Step 5 in Figure 4.12). We do so to understand the *diversity* of memes in each Web community, as well as the interplay between *variants* of memes. We then evaluate how clusters can be grouped into higher structures using hierarchical clustering and graph visualization techniques.

Clusters

Statistics. In Table 4.15, we report some basic statistics of the clusters obtained for each Web community. A relatively high percentage of images (63%–69%) are not clustered, i.e.,



(a) (b)

Figure 4.18: CDF of (a) KYM entries per cluster and (b)clusters per KYM entry.

are labeled as noise. While in DBSCAN "noise" is just an instance that does not fit in any cluster (more specifically, there are less than 5 images with perceptual distance ≤ 8 from that particular instance), we note that this likely happens as these images are not memes, but rather "one-off images." For example, on /pol/ there is a large number of pictures of random people taken from various social media platforms.

Overall, we have 2.1M images in 63.9K clusters: 38K clusters for /pol/, 21K for The_Donald, and 3K for Gab. 12.6K of these clusters are successfully annotated using the KYM data: 9.2K from /pol/ (142K images), 2.9K from The_Donald (121K images), and 447 from Gab (4.5K images).

We also present some examples of clusters showcasing how the proposed pipeline can effectively detect and group images that belong to the same meme.

Specifically, Figure 4.19 shows a subset of the images from the Dubs Guy/Check Em meme [274], Figure 4.20 a subset of images that belong to the Nut Button meme [275], while Figure 4.21 – to the Goofy's Time meme [276]. Note that all these images are obtained from /pol/ clusters.

In all clusters, we observe similar variations, i.e., variations of Donald Trump, Adolf Hitler, The Happy Merchant, and Pepe the Frog appear in all examples. Once again, this emphasizes the overlap that exists among memes.

As for the un-annotated clusters, manual inspection confirms that many include miscellaneous images unrelated to memes, e.g., similar screenshots of social networks posts (recall that we only filter out screenshots from the KYM image galleries), images captured from video games,

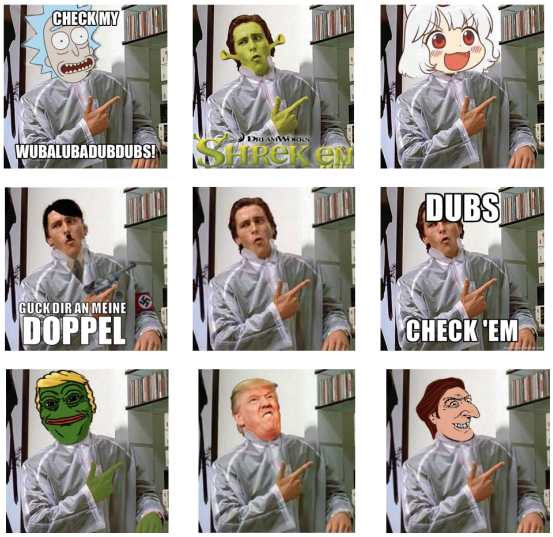


Figure 4.19: Images that are part of the Dubs Guy/Check Em Meme.

Figure 4.20: Images that are part of the Nut Button Meme.

etc.

KYM entries per cluster. Each cluster may receive multiple annotations, depending on the KYM entries that have at least one image matching that cluster's medoid. As shown in Figure 4.18(a), the majority of the annotated clusters (74% for /pol/, 70% for The_Donald, and 58% for Gab) only have a single matching KYM entry. However, a few clusters have a large number of matching entries, e.g., the one matching the Conspiracy Keanu meme [277] is annotated by 126 KYM entries (primarily, other memes that add text in an image associated with that meme). This highlights that memes do overlap and that some are highly influenced by other ones.

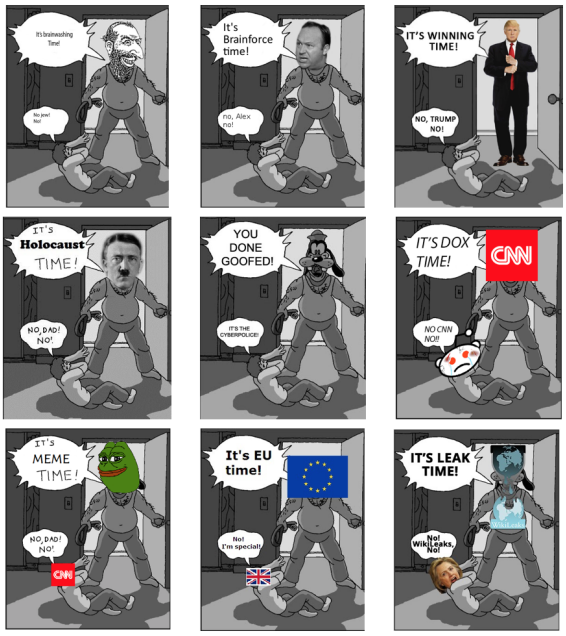


Figure 4.21: Images that are part of the Goofy's Time Meme.

Clusters per KYM entry. We also look at the number of clusters annotated by the *same* KYM entry. Figure 4.18(b) plots the CDF of the number of clusters per entry. About 40% only annotate a single /pol/ cluster, while 34% and 20% of the entries annotate a single The_Donald and a single Gab cluster, respectively. We also find that a small number of entries are associated to a large number of clusters: for example, the Happy Merchant meme [252] annotates 124 different clusters on /pol/. This highlights the *diverse* nature of memes, i.e., memes are mixed and matched, not unlike the way that genetic traits are combined in biological reproduction.

Top KYM entries. Because the majority of clusters match only one or two KYM entries (Figure 4.18(a)), we simplify things by giving all clusters a *representative annotation* based on the most prevalent annotation given to the medoid, and, in the case of ties the average distance between all matches (see Section 4.2.2). *Thus, in the rest of this thesis, we report our findings based on the representative annotation for each cluster.*

In Table 4.16, we report the top 20 KYM entries with respect to the number of clusters they annotate. These cover 17%, 23%, and 27% of the clusters in /pol/, The_Donald, and Gab, respectively, hence covering a relatively good sample of our datasets. Donald Trump [269], Smug Frog [255], and Pepe the Frog [253] appear in the top 20 for all three communities, while the Happy Merchant [252] only in /pol/ and Gab. In particular, Donald Trump annotates the most clusters (207 in /pol/, 177 in The_Donald, and 25 in Gab). In fact, politics-related

/pol/			T_D			Gab		
Entry	**Category**	**Clusters (%)**	**Entry**	**Category**	**Clusters (%)**	**Entry**	**Category**	**Clusters (%)**
Donald Trump	People	207 (2.2%)	Donald Trump	People	177 (6.1%)	Donald Trump	People	25 (5.6%)
Happy Merchant	Memes	124 (1.3%)	Smug Frog	Memes	78 (2.7%)	Happy Merchant	Memes	10 (2.2%)
Smug Frog	Memes	114 (1.2%)	Pepe the Frog	Memes	63 (2.1%)	Demotivational Posters	Memes	7 (1.5%)
Computer Reaction Faces	Memes	112 (1.2%)	Feels Bad Man/ Sad Frog	Memes	61 (2.1%)	Pepe the Frog	Memes	6 (1.3%)
Feels Bad Man/ Sad Frog	Memes	94 (1.0%)	Make America Great Again	Memes	50 (1.7%)	#Cnnblackmail	Events	6 (1.3%)
I Know that Feel Bro	Memes	90 (1.0%)	Bernie Sanders	People	31 (1.0%)	2016 US election	Events	6 (1.3%)
Tony Kornheiser's Why	Memes	89 (1.0%)	2016 US Election	Events	27 (0.9%)	Know Your Meme	Sites	6 (1.3%)
Bait/This is Bait	Memes	84 (0.9%)	Counter Signal Memes	Memes	24 (0.8%)	Tumblr	Sites	6 (1.3%)
#TrumpAnime/Rick Wilson	Events	76 (0.8%)	#Cnnblackmail	Events	24 (0.8%)	Feminism	Cultures	5 (1.1%)
Reaction Images	Memes	73 (0.8%)	Know Your Meme	Sites	20 (0.7%)	Barack Obama	People	5 (1.1%)
Make America Great Again	Memes	72 (0.8%)	Angry Pepe	Memes	18 (0.6%)	Smug Frog	Memes	5 (1.1%)
Counter Signal Memes	Memes	72 (0.8%)	Demotivational Posters	Memes	18 (0.6%)	rwby	Subcultures	5 (1.1%)
Pepe the Frog	Memes	65 (0.7%)	4chan	Sites	16 (0.5%)	Kim Jong Un	People	5 (1.1%)
Spongebob Squarepants	Subcultures	61 (0.7%)	Tumblr	Sites	15 (0.5%)	Murica	Memes	5 (1.1%)
Doom Paul its Happening	Memes	57 (0.6%)	Gamergate	Events	15 (0.5%)	UA Passenger Removal	Events	5 (1.1%)
Adolf Hitler	People	56 (0.6%)	Colbertposting	Memes	15 (0.5%)	Make America Great Again	Memes	4 (0.9%)
pol	Sites	53 (0.6%)	Donald Trump's Wall	Memes	15 (0.5%)	Bill Nye	People	4 (0.9%)
Dubs Guy/Check'em	Memes	53 (0.6%)	Vladimir Putin	People	15 (0.5%)	Trolling	Cultures	4 (0.9%)
Smug Anime Face	Memes	51 (0.6%)	Barack Obama	People	15 (0.5%)	4chan	Sites	4 (0.9%)
Warhammer 40000	Subcultures	51 (0.6%)	Hillary Clinton	People	15 (0.5%)	Furries	Cultures	3 (0.7%)
Total		**1,638 (17.7%)**			**695 (23.9%)**			**121 (27.1%)**

Table 4.16: Top 20 KYM entries appearing in the clusters of /pol/, The_Donald, and Gab. We report the number of clusters and their respective percentage (per community). Each item contains a hyperlink to the corresponding entry on the KYM website.

entries appear several times in the Table, e.g., Make America Great Again [278] as well as political personalities like Bernie Sanders, Barack Obama, Vladimir Putin, and Hillary Clinton.

When comparing the different communities, we observe the most prevalent categories are memes (6 to 14 entries in each community) and people (2-5). Moreover, in /pol/, the 2nd most popular entry, related to people, is Adolf Hilter, which supports previous reports of the community's sympathetic views toward Nazi ideology [19]. Overall, there are several memes with hateful or disturbing content (e.g., holocaust). This happens to a lesser extent in The_Donald and Gab: the most popular people after Donald Trump are contemporary politicians, i.e., Bernie Sanders, Vladimir Putin, Barack Obama, and Hillary Clinton.

Finally, image posting behavior in fringe Web communities is greatly influenced by real-world events. For instance, in /pol/, we find the #TrumpAnime controversy event [279], where a political individual (Rick Wilson) offended the alt-right community, Donald Trump supporters, and anime fans (an oddly intersecting set of interests of /pol/ users). Similarly, on The_Donald and Gab, we find the #Cnnblackmail [270] event, referring to the (alleged) blackmail of the Reddit user that created the infamous video of Donald Trump wrestling the CNN.

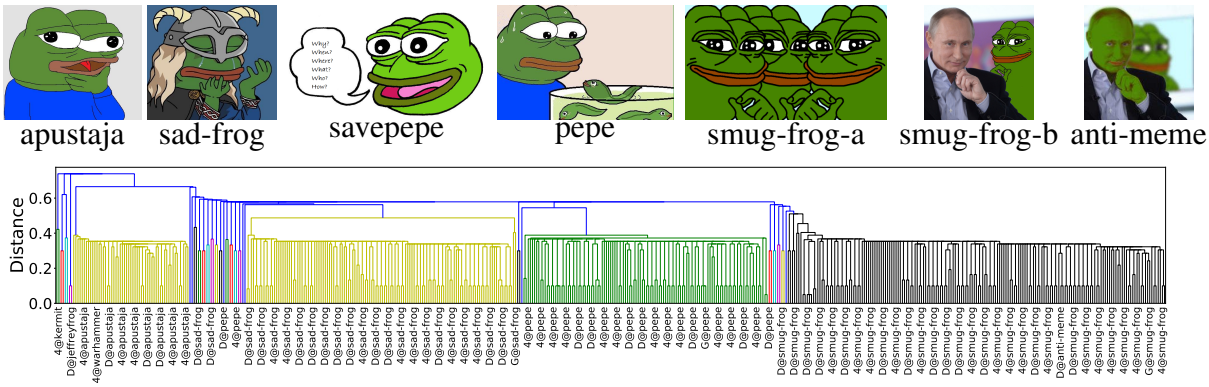


Figure 4.22: Inter-cluster distance between all clusters with frog memes. Clusters are labeled with the origin (*4* for 4chan, *D* for The_Donald, and *G* for Gab) and the meme name. To ease readability, we do not display all labels, abbreviate meme names, and only show an excerpt of all relationships.

Memes' Branching Nature

Next, we study how memes *evolve* by looking at *variants* across different clusters. Intuitively, clusters that look alike and/or are part of the same meme are grouped together under the same branch of an evolutionary tree. We use the custom distance metric introduced in Section 4.2.2, aiming to infer the phylogenetic relationship between variants of memes. Since there are 12.6K annotated clusters, we only report on a subset of variants. In particular, we focus on "frog" memes (e.g., Pepe the Frog [253]); as discussed later in Section 4.2.4, this is one of the most popular memes in our datasets.

The dendrogram in Figure 4.22 shows the hierarchical relationship between groups of clusters of memes related to frogs. Overall, there are 525 clusters of frogs, belonging to 23 different memes. These clusters can be grouped into four large categories, dominated by Apu Apustaja [280], Feels Bad Man/Sad Frog [281], Pepe the Frog [253], and Smug Frog [255]. The different memes express different ideas or messages: e.g., Apu Apustaja depicts a simple-minded non-native speaker using broken English, while the Feels Bad Man/Sad Frog (ironically) expresses dismay at a given situation, often accompanied with text like "You will never do/be/have X." The dendrogram also shows a variant of Smug Frog (*smug-frog-b*) related to a variant of the Russian Anti Meme Law [282] (*anti-meme*) as well as relationships between clusters from Pepe the Frog and Isis meme [283], and between Smug Frog and Brexit-related clusters [284], as shown in Section 4.2.5.

The distance metric quantifies the similarity of any two variants of *different* memes; however, recall that two clusters can be close to each other even when the medoids are perceptually different (see Section 4.2.2), as in the case of Smug Frog variants in the *smug-frog-a* and

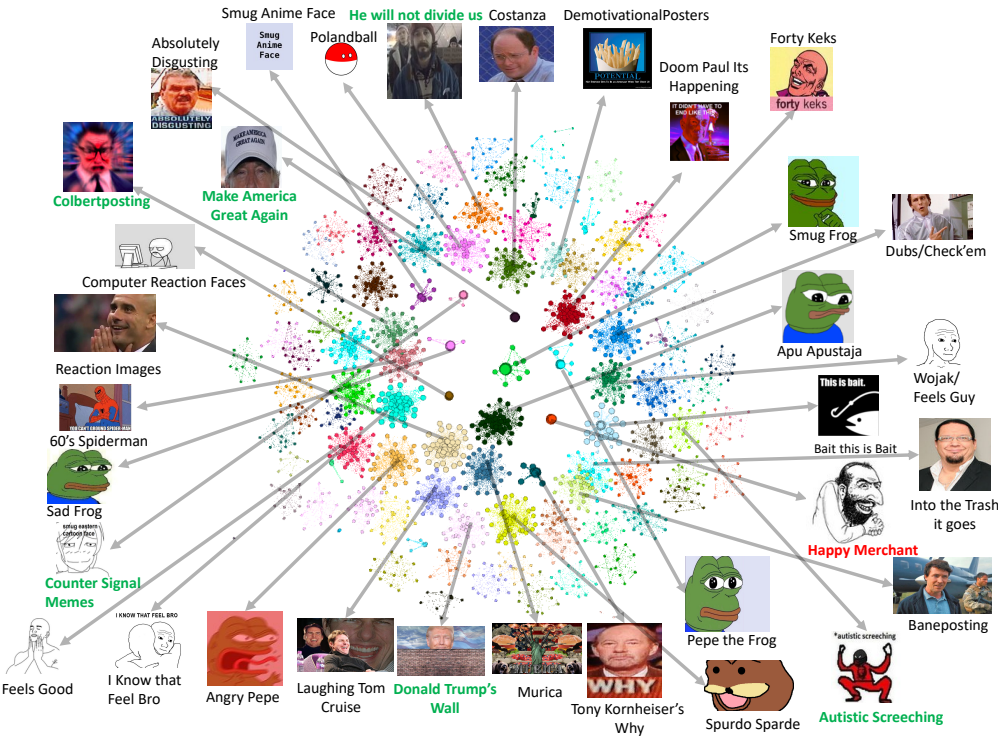


Figure 4.23: Visualization of the obtained clusters from /pol/, The_Donald, and Gab. Note that memes with red labels are annotated as racist, while memes with green labels are annotated as politics (see Section 4.2.4 for the selection criteria).

smug-frog-b clusters (top of Figure 4.22). Although this analysis is limited to a single "family" of memes, our distance metric can actually provide useful insights regarding the phylogenetic relationships of any clusters. In fact, more extensive analysis of these relationships (through our pipeline) can facilitate the understanding of the diffusion of ideas and information across the Web, and provide a rigorous technique for large-scale analysis of Internet culture.

Meme Visualization

We also use the custom distance metric (see Eq. 4.1) to visualize the clusters with annotations. We build a graph $G = (V, E)$, where V are the medoids of annotated clusters and E the connections between medoids with distance under a threshold κ. Figure 4.23 shows a snapshot of the graph for $\kappa = 0.45$, chosen based on the frogs analysis above. In particular, we select this threshold as the majority of the clusters from the same meme (note coloration in Figure 4.22) are hierarchically connected with a higher-level cluster at a distance close to 0.45. To ease readability, we filter out nodes and edges that have a sum of in- and out-degree less than 10,

which leaves 40% of the nodes and 92% of the edges. Nodes are colored according to their KYM annotation. NB: the graph is laid out using the OpenOrd algorithm [14] and the distance between the components in it does not exactly match the actual distance metric. We observe a large set of disconnected components, with each component containing nodes of primarily one color. This indicates that our distance metric is indeed capturing the peculiarities of different memes. Finally, note that an interactive version of the full graph is publicly available from [254].

Web Community-based Analysis

We now present a macro-perspective analysis of the Web communities through the lens of memes. We assess the presence of different memes in each community, how popular they are, and how they evolve. To this end, we examine the *posts* from all four communities (Twitter, Reddit, /pol/, and Gab) that contain *images* matching *memes* from fringe Web communities (/pol/, The_Donald, and Gab).

Meme Popularity

Memes. We start by analyzing clusters grouped by KYM 'meme' entries, looking at the number of posts for each meme in /pol/, Reddit, Gab, and Twitter.

In Table 4.17, we report the top 20 memes for each Web community sorted by the number of posts. We observe that Pepe the Frog [253] and its variants are among the most popular memes for every platform. While this might be an artifact of using fringe communities as a "seed" for the clustering, recall that the goal of this work is in fact to gain an understanding of how fringe communities disseminate memes and influence mainstream ones. Thus, we leave to future work a broader analysis of the wider meme ecosystem.

Sad Frog [281] is the most popular meme on /pol/ (4.9%), the 3rd on Reddit (1.3%), the 10th on Gab (0.8%), and the 12th on Twitter (0.5%). We also find variations like Smug Frog [255], Apu Apustaja [280], Pepe the Frog [253], and Angry Pepe [285]. Considering that Pepe is treated as a hate symbol by the Anti-Defamation League [249] and that is often used in hateful or racist, this likely indicates that polarized communities like /pol/ and Gab do use memes to incite hateful conversation. This is also evident from the popularity of the anti-semitic Happy Merchant meme [252], which depicts a "greedy" and "manipulative" stereotypical caricature of a Jew (3.8% on /pol/ and 1.1% on Gab).

/pol/		Reddit		Gab		Twitter	
Entry	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts(%)**
Feels Bad Man/Sad Frog	64,367 (4.9%)	Manning Face	12,540 (2.2%)	Jesusland **(P)**	454 (1.6%)	Roll Safe	55,010 (5.9%)
Smug Frog	63,290 (4.8%)	That's the Joke	7,626 (1.3%)	Demotivational Posters	414 (1.5%)	Evil Kermit	50,642 (5.4%)
Happy Merchant **(R)**	49,608 (3.8%)	Feels Bad Man/ Sad Frog	7,240 (1.3%)	Smug Frog	392 (1.4%)	Arthur's Fist	37,591 (4.0%)
Apu Apustaja	29,756 (2.2%)	Confession Bear	7,147 (1.3%)	Based Stickman **(P)**	391 (1.4%)	Nut Button	13,598 (1,5%)
Pepe the Frog	25,197 (1.9%)	This is Fine	5,032 (0.9%)	Pepe the Frog	378 (1.3%)	Spongebob Mock	11,136 (1,2%)
Make America Great Again **(P)**	21,229 (1.6%)	Smug Frog	4,642 (0.8%)	Happy Merchant **(R)**	297 (1.1%)	Reaction Images	9,387 (1.0%)
Angry Pepe	20,485 (1.5%)	Roll Safe	4,523 (0.8%)	Murica	274 (1.0%)	Conceited Reaction	9,106 (1.0%)
Bait this is Bait	16,686 (1.2%)	Rage Guy	4,491 (0.8%)	And Its Gone	235 (0.9%)	Expanding Brain	8,701 (0.9%)
I Know that Feel Bro	14,490 (1.1%)	Make America Great Again **(P)**	4,440 (0.8%)	Make America Great Again **(P)**	207 (0.8%)	Demotivational Posters	7,781 (0.8%)
Cult of Kek	14,428 (1.1%)	Fake CCG Cards	4,438 (0.8%)	Feels Bad Man/ Sad Frog	206 (0.8%)	Cash Me Ousside/Howbow Dah	5,972 (0.6%)
Laughing Tom Cruise	14,312 (1.1%)	Confused Nick Young	4,024 (0.7%)	Trump's First Order of Business **(P)**	192 (0.7%)	Salt Bae	5,375 (0.6%)
Awoo	13,767 (1.0%)	Daily Struggle	4,015 (0.7%)	Kekistan	186 (0.6%)	Feels Bad Man/ Sad Frog	4,991 (0.5%)
Tony Kornheiser's Why	13,577 (1.0%)	Expanding Brain	3,757 (0.7%)	Picardia **(P)**	183 (0.6%)	Math Lady/Confused Lady	4,722 (0.5%)
Picardia **(P)**	13,540 (1.0%)	Demotivational Posters	3,419 (0.6%)	Things with Faces (Pareidolia)	156 (0.5%)	Computer Reaction Faces	4,720 (0.5%)
Big Grin / Never Ever	12,893 (1.0%)	Actual Advice Mallard	3,293 (0.6%)	Serbia Strong/Remove Kebab	149 (0.5%)	Clinton Trump Duet **(P)**	3,901 (0.4%)
Reaction Images	12,608 (0.9%)	Reaction Images	2,959 (0.5%)	Riot Hipster	148 (0.5%)	Kendrick Lamar Damn Album Cover	3,656 (0.4%)
Computer Reaction Faces	12,247 (0.9%)	Handsome Face	2,675 (0.5%)	Colorized History	144 (0.5%)	What in tarnation	3,363 (0.3%)
Wojak / Feels Guy	11,682 (0.9%)	Absolutely Disgusting	2,674 (0.5%)	Most Interesting Man in World	140 (0.5%)	Harambe the Gorilla	3,164 (0.3%)
Absolutely Disgusting	11,436 (0.8%)	Pepe the Frog	2,672 (0.5%)	Chuck Norris Facts	131 (0.4%)	I Know that Feel Bro	3,137 (0.3%)
Spurdo Sparde	9,581 (0.7%)	Pretending to be Retarded	2,462 (0.4%)	Roll Safe	131 (0.4%)	This is Fine	3,094 (0.3%)
Total	**445,179 (33.4%)**		**94,069 (16.7%)**		**4,808 (17.0%)**		**249,047 (26.4%)**

Table 4.17: Top 20 KYM entries for memes that we find our datasets. We report the number of posts for each meme as well as the percentage over all the posts (per community) that contain images that match one of the annotated clusters. The (R) and (P) markers indicate whether a meme is annotated as racist or politics-related, respectively (see Section 4.2.4 for the selection criteria).

By contrast, mainstream communities like Reddit and Twitter primarily share harmless/neutral memes, which are rarely used in hateful contexts. Specifically, on Reddit the top memes are Manning Face [286] (2.2%) and That's the Joke [287] (1.3%), while on Twitter the top ones are Roll Safe [288] (5.9%) and Evil Kermit [289] (5.4%).

Once again, we find that users (in all communities) post memes to share politics-related information, possibly aiming to enhance or penalize the public image of politicians (see Section 4.2.5 for an example of such memes). For instance, we find Make America Great Again [278], a meme dedicated to Donald Trump's US presidential campaign, among the top memes in /pol/ (1.6%), in Reddit (0.8%), and Gab (0.8%). Similarly, in Twitter, we find the Clinton Trump Duet meme [290] (0.4%), a meme inspired by the 2nd US presidential debate.

People. We also analyze memes related to people (i.e., KYM entries with the people category). Table 4.18 reports the top 15 KYM entries in this category. We observe that, in all Web Communities, the most popular person portrayed in memes is Donald Trump: he is depicted in 4.6% of /pol/ posts that contain annotated images, while for Reddit, Gab, and Twitter the percentages are 6.1%, 6.1%, and 1.3%, respectively. Other popular personalities, in all platforms, include several politicians. For instance, in /pol/, we find Mike Pence (0.3%), Jeb Bush (0.3%), Vladimir Putin (0.2%), while, in Reddit, we find Steve Bannon (0.6%), Chelsea Manning (0.6%), and Bernie Sanders (0.3%), in Gab, Mitt Romney (1.7%) and Barack

/pol/		Reddit		Gab		Twitter	
Entry	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts (%)**	**Entry**	**Posts(%)**
Donald Trump	60,611 (4.6%)	Donald Trump	34,533 (6.1%)	Donald Trump	1,665 (6.1%)	Donald Trump	10,208 (1.3%)
Adolf Hitler	8,759 (0.6%)	Steve Bannon	3,733 (0.6%)	Mitt Romney	455 (1.7%)	Barack Obama	5,187 (0.6%)
Mike Pence	4,738 (0.3%)	Stephen Colbert	3,121 (0.6%)	Bill Nye	370 (1.3%)	Chelsea Manning	4,173 (0.5%)
Jeb Bush	4,217 (0.3%)	Chelsea Manning	2,261 (0.4%)	Adolf Hitler	106 (0.4%)	Kim Jong Un	3,271 (0.4%)
Vladimir Putin	3,218 (0.2%)	Ben Carson	2,148 (0.4%)	Barack Obama	104 (0.4%)	Anita Sarkeesian	2,764 (0.3%)
Alex Jones	3,206 (0.2%)	Bernie Sanders	1,757 (0.3%)	Isis Daesh	92 (0.3%)	Bernie Sanders	2,277 (0.3%)
Ron Paul	3,116 (0.2%)	Ajit Pai	1,658 (0.3%)	Death Grips	91 (0.3%)	Vladimir Putin	1,733 (0.2%)
Bernie Sanders	3,022 (0.2%)	Barack Obama	1,628 (0.3%)	Eminem	89 (0.3%)	Billy Mays	1,454 (0.2%)
Massimo D'alema	2,725 (0.2%)	Gabe Newell	1,518 (0.3%)	Kim Jong Un	87 (0.3%)	Adolf Hitler	1,304 (0.2%)
Mitt Romney	2,468 (0.2%)	Bill Nye	1,478 (0.3%)	Ajit Pai	76 (0.3%)	Kanye West	1,261 (0.2%)
Chelsea Manning	2,403 (0.2%)	Hillary Clinton	1,468 (0.3%)	Pewdiepie	73 (0.3%)	Bill Nye	968 (0.2%)
Hillary Clinton	2,378 (0.2%)	Death Grips	1,463 (0.3%)	Bernie Sanders	71 (0.3%)	Mitt Romney	923 (0.1%)
A. Wyatt Mann	2,110 (0.2%)	Adolf Hitler	1,449 (0.3%)	Alex Jones	70 (0.3%)	Filthy Frank	777 (0.1%)
Ben Carson	1,780 (0.1%)	Mitt Romney	1,294 (0.2%)	Hillary Clinton	59 (0.2%)	Hillary Clinton	758 (0.1%)
Filthy Frank	1,598 (0.1%)	Eminem	1,274 (0.2%)	Anita Sarkeesian	54 (0.2%)	Ajit Pai	715 (0.1%)

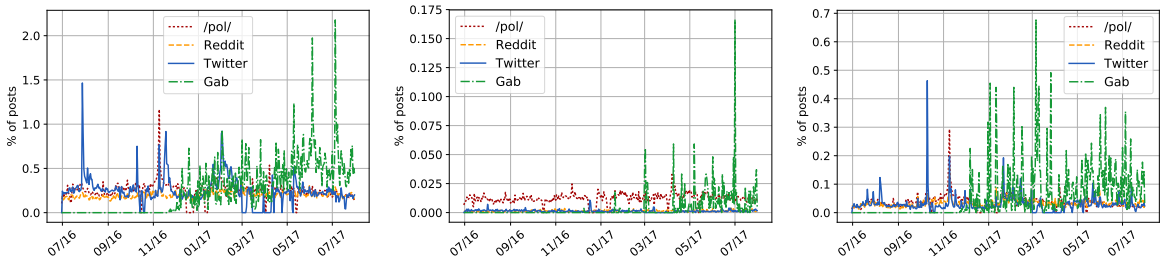
Table 4.18: Top 15 KYM entries about people that we find in each of our dataset. We report the number of posts and the percentage over all the posts (per community) that match a cluster with KYM annotations.

Obama (0.4%), and, in Twitter, Barack Obama (0.6%), Kim Jong Un (0.5%), and Chelsea Manning (0.4%). This highlights the fact that users on these communities utilize memes to share information and opinions about politicians, and possibly try to either enhance or harm public opinion about them. Finally, we note the presence of Adolf Hitler memes on all Web Communities, i.e., /pol/ (0.6%), Reddit (0.3%), Gab (0.4%), and Twitter (0.2%).

We further group memes into two high-level groups, racist and politics-related. We use the *tags* that are available in our KYM dataset, i.e., we assign a meme to the politics-related group if it has the "politics," "2016 us presidential election," "presidential election," "trump," or "clinton" tags, and to the racism-related one if the tags include "racism," "racist," or "antisemitism," obtaining 117 racist memes (4.4% of all memes that appear in our dataset) and 556 politics-related memes (21.2% of all memes that appear on our dataset). In the rest of this section, we use these groups to further study the memes, and later in Section 4.2.5 to estimate influence.

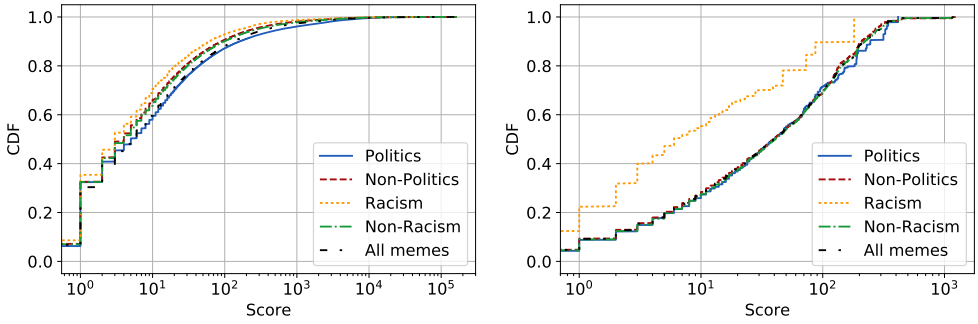
Temporal Analysis

Next, we study the temporal aspects of posts that contain memes from /pol/, Reddit, Twitter, and Gab. In Figure 4.24, we plot the percentage of posts per day that include memes. For all memes (Figure 4.24(a)), we observe that /pol/ and Reddit follow a steady posting behavior,



(a) all memes (b) racist (c) politics

Figure 4.24: Percentage of posts per day in our dataset for all, racist, and politics-related memes.



(a) Reddit (b) Gab

Figure 4.25: CDF of scores of posts that contain memes on Reddit and Gab.

with a peak in activity around the 2016 US elections. We also find that memes are increasingly more used on Gab (see, e.g., 2016 vs 2017).

As shown in Figure 4.24(b), both /pol/ and Gab include a substantially higher number of posts with racist memes, used over time with a difference in behavior: while /pol/ users share them in a very steady and constant way, Gab exhibits a bursty behavior. A possible explanation is that the former is inherently more racist, with the latter primarily reacting to particular world events. As for political memes (Figure 4.24(c)), we find a lot of activity overall on Twitter, Reddit, and /pol/, but with different spikes in time. On Reddit and /pol/, the peaks coincide with the 2016 US elections. On Twitter, we note a peak that coincides with the 2nd US Presidential Debate on October 2016. For Gab, there is again an increase in posts with political memes after January 2017.

All Memes		Racism-Related Memes		Politics-Related Memes	
Subreddit	**Posts (%)**	**Subreddit**	**Posts (%)**	**Subreddit**	**Posts (%)**
The_Donald	82,698 (12.5%)	The_Donald	359 (9.3%)	The_Donald	24,343 (26.4%)
AdviceAnimals	35,475 (5.3%)	AdviceAnimals	87 (2.2%)	politics	2,751 (3.0%)
me_irl	15,366 (2.3%)	conspiracy	76 (2.0%)	EnoughTrumpSpam	2,679 (2.9%)
politics	8,875 (1,3%)	me_irl	70 (1.8%)	TrumpsTweets	2,363 (2.5%)
funny	8,508 (1.3%)	funny	56 (1.4%)	AdviceAnimals	1,740 (1.9%)
dankmemes	7,744 (1,1%)	CringeAnarchy	43 (1.1%)	USE2016	1,653 (1.8%)
EnoughTrumpSpam	6,973 (1.1%)	EDH	43 (1.1%)	PoliticsAll	1,401(1.5%)
pics	5,945 (0.9%)	magicTCG	42 (1.1%)	dankmemes	881 (0.9%)
AskReddit	5,482 (0.8%)	dankmemes	40 (1.0%)	pics	877 (0.9%)
HOTandTrending	4,674 (0.7%)	ImGoingToHellForThis	39 (1.0%)	me_irl	873 (0.9%)

Table 4.19: Top ten subreddits for all memes, racism-related memes, and politics-related memes.

Score Analysis

As discussed in Chapter 2, Reddit and Gab incorporate a voting system that determines the popularity of content within the Web community and essentially captures the appreciation of other users towards the shared content. To study how users react to racist and politics-related memes, we plot the CDF of the posts' scores that contain such memes in Figure 4.25.

For Reddit (Figure 4.25(a)), we find that posts that contain politics-related memes are rated highly (mean score of 224.7 and a median of 5) than posts that contain non-politics memes (mean 124.9, median 4). On the contrary, posts that contain racist memes are rated lower (average score of 94.8 and a median of 3) than other non-racist memes (average 141.6 and median 4). On Gab (Figure 4.25(b)), posts that contain politics-related memes have a similar score as non-political memes (mean 87.3 vs 82.4). However, this does not apply for racist and non-racist memes, as non-racist memes have over 2 times higher scores than racist memes (means 84.7 vs 35.5).

Overall, this suggests that posts that contain politics-related memes receive high scores by Reddit and Gab users, while for racist memes this applies only on Reddit.

Sub-Communities

Among all the Web communities that we study, only Reddit is divided into multiple sub-communities. We now study which sub-communities share memes with a focus on racist and politics-related content. In Table 4.19, we report the top ten subreddits in terms of the percentage over all posts that contain memes in Reddit for: 1) all memes; 2) racist ones; and

3) politics-related memes.

For all three groups, the most popular subreddit is The_Donald with 12.5%, 9.3%, and 26.4%, respectively. Interestingly, AdviceAnimals, a general-purpose meme subreddit, is among the top-ten sub-communities also for racist and political memes, highlighting their infiltration in otherwise non-hateful communities.

Other popular subreddits for racist memes include conspiracy (2.0%), me_irl (1.8%), and funny (1.4%) subreddits. For politics-related memes, the majority of the subreddits are related to Donald Trump, while there also are general subreddits that talk about politics, e.g., the politics (3.0%) and the PoliticsAll subreddit (1.5%).

Take-Aways

In summary, the main take-aways of our analysis include:

1. Fringe Web communities use many variants of memes related to politics and world events, possibly aiming to share weaponized information about them (Section 4.2.5 include some examples of such memes). For instance, Donald Trump is the KYM entry with the largest number of clusters in /pol/ (2.2%), The_Donald (6.1%), and Gab (2.2%).
2. /pol/ and Gab share hateful and racist memes at a higher rate than mainstream communities, as we find a considerable number of anti-semitic and pro-Nazi clusters (e.g., The Happy Merchant meme [252] appears in 1.3% of all /pol/ annotated clusters and 2.2% of Gab's, while Adolf Hitler in 0.6% of /pol/'s). This trend is steady over time for /pol/ but ramping up for Gab.
3. Seemingly "neutral" memes, like Pepe the Frog (or one of its variants), are used in conjunction with other memes to incite hate or influence public opinion on world events, e.g., with images related to terrorist organizations like ISIS or world events such as Brexit.
4. Our custom distance metric successfully allows us to study the interplay and the overlap of memes, as showcased by the visualizations of the clusters and the dendrogram (see Figs. 4.22 and 4.23).
5. Reddit users are more interested in politics-related memes than other type of memes. That said, when looking at individual subreddits, we find that The_Donald is the most active one when it comes to posting memes in general. It is also the subreddit where most racism and politics-related memes are posted.

/pol/	Twitter	Reddit	T_D	Gab
1,574,045	865,885	581,803	81,924	44,918

Table 4.20: Events per community from the 12.6K clusters.

4.2.5 Influence Estimation

So far we have studied the dissemination of memes by looking at Web communities in isolation. However, in reality, these influence each other: e.g., memes posted on one community are often re-posted to another. Aiming to capture the relationship between them, we use a statistical model known as Hawkes Processes [27, 28], which describes how events occur over time on a collection of processes (for more details regarding Hawkes Processes see Section 2.2).

Influence Results

We fit Hawkes models using Gibbs sampling as described in [28] for the 12.6K annotated clusters; in Table 4.20, we report the total number of meme images posted to each community in these clusters. As seen in Table 4.20, /pol/ has the greatest number of memes posted, followed by Twitter and then Reddit. In terms of total images collected (see Table 4.14), Twitter and Reddit have many more than /pol/. However, many of the images on these communities might not be memes; additionally, because our clusters are created from the memes present on only /pol/, The_Donald, and Gab (as these are the communities primarily of interest in this work), it is possible that there are memes on Twitter and Reddit that are not included in the clusters. This yields an additional interesting question: how *efficient* are different communities at disseminating memes?

First, we report the source of events in terms of the percent of events on the destination community. This describes the results in terms of the data as we have collected it, e.g., it tells us the percentage of memes posted on Twitter that were caused by /pol/. The second way we report influence is by normalizing the values by the total number of events in the source community, which lets us see how much influence each community has, relative to the number of memes they post—in other words, their efficiency.

We first look at the influence of all clusters together. Figure 4.26 shows the percent of events on each *destination* community caused by each *source* community. The values from one community to the same community (for example, from /pol/ to /pol/) include both events caused

by the background rate of that community and events caused by previous events within that community; these values are the largest influence for each community. After this, /pol/ is the strongest source of influence for Reddit, The_Donald, and Gab, but not for Twitter, which is most influenced by Reddit. Interestingly, although Twitter has a greater number of memes posted than Reddit, it causes less influence. Perhaps there is less original content posted directly to Twitter.

Next, we look at the normalized influence of all clusters together. Figure 4.27 shows the influence that a source community has on a destination community, normalized by the total number of memes posted on the *source* community. The values can be understood as an indication of how much influence a community has, relative to the frequency of memes posted. For example, the influence Reddit has on Twitter is equal to 5.71% of the total events on Reddit. If the sum of values for a source is less than 100%, it implies that many of the posts on the source community were caused by other communities, or that posts on the source community do not cause many posts on other communities.

There are several interesting things to note in Figure 4.27. First, The_Donald has by far the greatest influence for the number of memes posted on it. This is particularly apparent when looking at just external influence, where The_Donald has more than 4 times as much influence than the rest of Reddit, the closest other community. Memes from this community spread very well to all of the other communities. While /pol/ has a large total influence on the other communities (as seen in Figure 4.26), when normalized by its size, it has the smallest external influence: just 4.03%. Most of the memes posted on /pol/ do not spread to other communities. Both Gab and Twitter have a total normalized influence of less than 100%; much less in Gab's case, although it has higher external influence.

Using the clusters identified as either racist or non-racist (see the end of Section 4.2.4), we compare how the communities influence the spread of these two types of content. Figure 4.28 shows the percentage of both the destination community's racist and non-racist meme posts caused by the source community. We perform two-sample Kolmogorov-Smirnov tests to compare the distributions of influence from the racist and non-racist clusters; cells with statistically significant differences between influence of racist/non-racist memes (with $p<0.01$) are reported with a * in the figure. /pol/ has the most *total* influence for both racist and non-racist memes, with the notable exception of Twitter, where Reddit has the most the influence. Interestingly, while the percentage of racist meme posts caused by /pol/ is greater than non-racist for Reddit, Twitter, and Gab, this is *not* the case for The_Donald. The only other cases where influence is greater for racist memes are Reddit to The_Donald and Gab to Reddit.

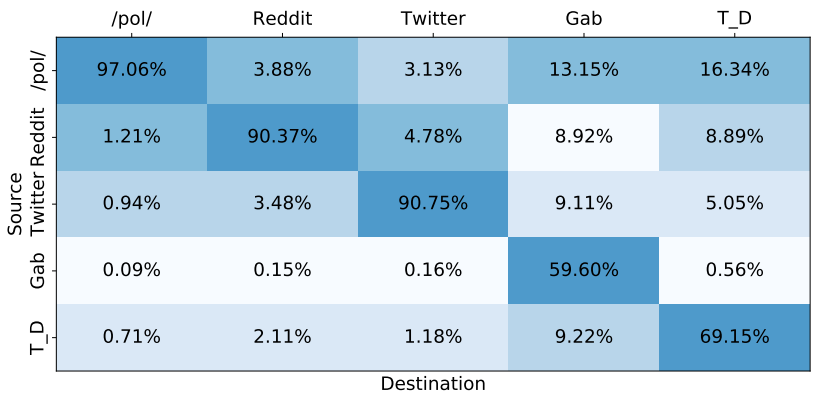


Figure 4.26: Percent of *destination* events caused by the source community on the destination community. Colors indicate the largest-to-smallest influences per destination.

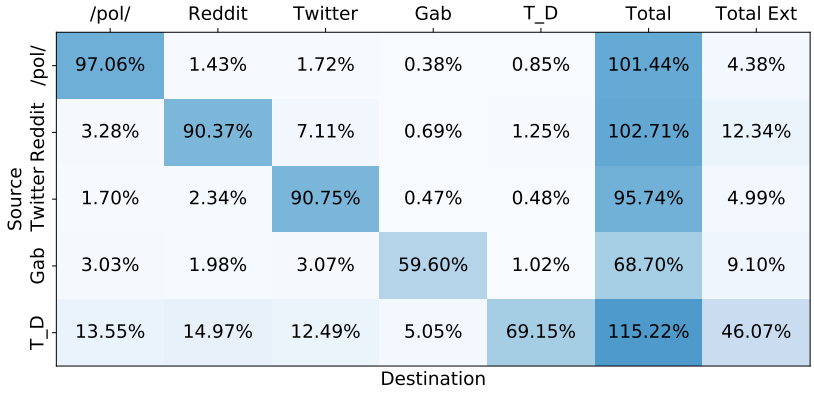


Figure 4.27: Influence from source to destination community, normalized by the number of events in the *source* community. Columns for total influence and total external influence are shown.

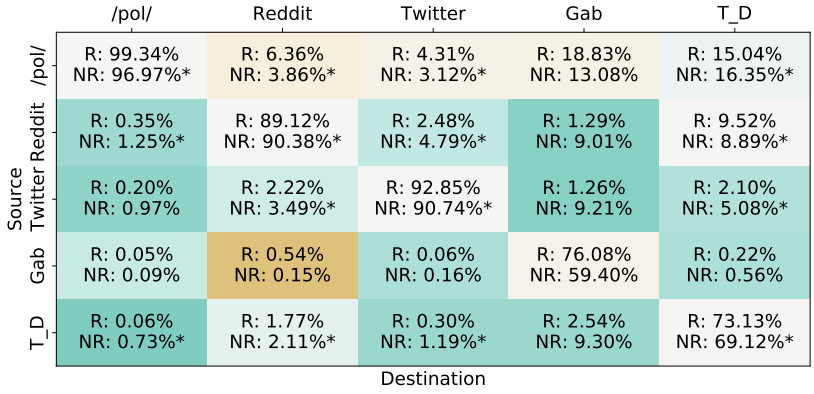


Figure 4.28: Percent of the destination community's racist (R) and non-racist (NR) meme postings caused by the source community. Colors indicate the percent difference between racist and non-racist.

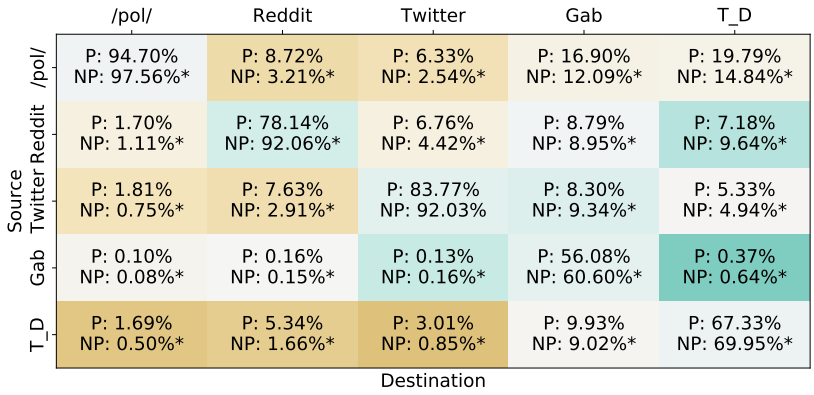


Figure 4.29: Percent of the destination community's political (P) and non-political (NP) meme postings caused by the source community. Colors indicate the percent difference between political and non-political.

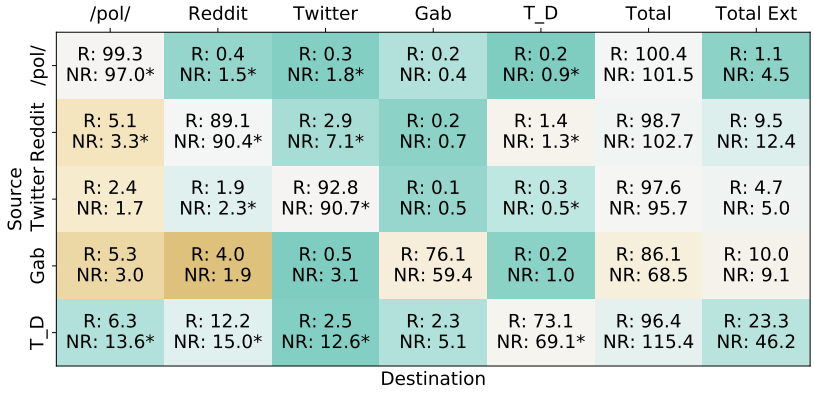


Figure 4.30: Influence from source to destination community of racist and non-racist meme postings, normalized by the number of events in the *source* community.

When looking at political vs non political memes (Figure 4.29), we see a somewhat different story. Here, /pol/ influences The_Donald more in terms of political memes. Further, we see differences in the *percent* increase and decrease of influence between the two figures (as indicated by the cell colors). For example, Twitter has a relatively larger difference in its influence on /pol/ and Reddit for political and non-political memes than for racist and non-racist memes, but a smaller difference in its influence on Gab and The_Donald. This exposes how different communities have varying levels of influence depending on the type of memes they post.

While examining the raw influence provides insights into the meme ecosystem, it obscures notable differences in the meme posting behavior of the different communities. To explore this, we look at the *normalized* influence in Figure 4.30 (racist/non-racist memes) and Figure 4.31

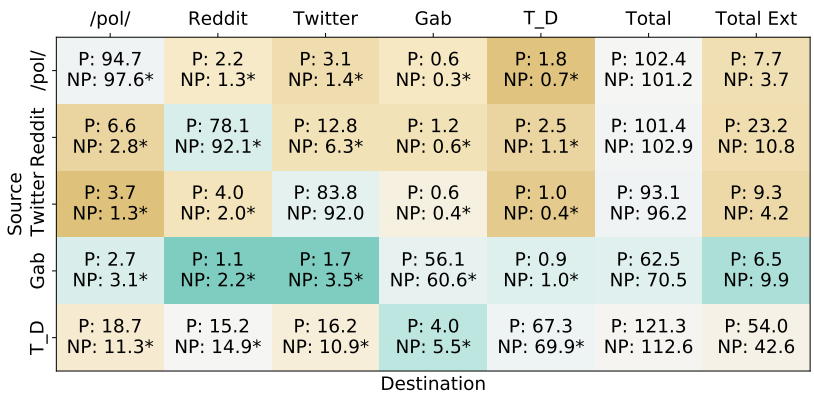


Figure 4.31: Influence from source to destination community of political and non-political meme postings, normalized by the number of events in the *source* community.

(political/non-political memes). As mentioned previously, normalization reveals how *efficient* the communities are in disseminating memes to other communities by revealing the *per meme* influence of meme posts. First, we note that the percent change in influence for the dissemination of racist/non-racist memes is quite a bit larger than that for political/non-political memes (again, indicated by the coloring of the cells). More interestingly, both figures show that, contrary to the *total* influence, /pol/ is the *least* influential when taking into account the number of memes posted. While this might seem surprising, it actually yields a subtle, yet crucial aspect of /pol/'s role in the meme ecosystem: /pol/ (and 4chan in general) acts as an evolutionary microcosm for memes. The constant production of new content [19] results in a "survival of the fittest" [291] scenario. A staggering number of memes are posted on /pol/, but only the *best* actually make it out to other communities. To the best of our knowledge, this is the first result quantifying this analogy to evolutionary pressure.

Take-Aways. There are several take-aways from our measurement of influence. We show that /pol/ is, generally speaking, the most influential disseminator of memes in terms of raw influence. In particular, it is more influential in spreading *racist* memes than non-racist one, and this difference is deeper than in any other community. There is one notable exception: /pol/ is more influential in terms of *non-racist* memes on The_Donald. Relatedly, /pol/ has generally more influence in terms of spreading political memes than other communities. When looking at the normalized influence, however, we surface a more interesting result: /pol/ is the *least* efficient in terms of influence while The_Donald is the *most* efficient. This provides new insight into the meme ecosystem: there are clearly evolutionary effects. Many meme postings do not result in further dissemination, and one of the key components to ensuring they are disseminated is ensuring that new "offspring" are continuously produced. /pol/'s "famed"

Figure 4.32: Image that exists in the clusters that are connected with frogs and Isis Daesh.

Figure 4.33: Image that exists in the clusters that are connected with frogs and Brexit.

meme magic, i.e., the propensity to produce and heavily push memes, is thus the most likely explanation for /pol/'s influence on the Web in general.

Interesting Images

Finally, we report some "interesting" examples of images from our frogs case study (see Section 4.2.4), as well as an example of an image for enhancing/penalizing the public image of specific politicians (as discussed in Section 4.2.4).

Specifically, Figure 4.32 shows an image connecting the Smug Frog [255] and the ISIS memes [283]. Also, Figure 4.33 shows an image connecting the Smug Frog and the Brexit meme [284]. Finally, Figure 4.34 shows a graphic image found in /pol/ that aims to attack the image of Hillary Clinton, while boosting that of Donald Trump. (The image depicts Hillary Clinton as a monster, Medusa, while Donald Trump is presented as Perseus, the hero who beheaded Medusa.)

Figure 4.34: Meme that is used for enhancing/penalizing the public image of specific politicians. Hillary Clinton is represented as Medusa, a monster, while Donald Trump is presented as Perseus (the hero who beheaded Medusa).

4.2.6 Remarks

In this work, we presented a large-scale measurement study of the meme ecosystem. We introduced a novel image processing pipeline and ran it over 160M images collected from four Web communities (4chan's /pol/, Reddit, Twitter, and Gab). We clustered images from fringe communities (/pol/, Gab, and Reddit's The_Donald) based on perceptual hashing and a custom distance metric, annotated the clusters using data gathered from Know Your Meme, and analyzed them along a variety of axes. We then associated images from all the communities to the clusters to characterize them through the lens of memes and the influence they have on each other.

Our analysis highlights that the meme ecosystem is quite complex, with intricate relationships between different memes and their variants. We found important differences between the memes posted on different communities (e.g., Reddit and Twitter tend to post "fun" memes, while Gab and /pol/ racist or political ones). When measuring the influence of each community toward disseminating memes to other Web communities, we found that /pol/ has the largest overall influence for racist and political memes, however, /pol/ was the least *efficient*, i.e., in terms of influence w.r.t. the total number of memes posted, while The_Donald is very

successful in pushing memes to both fringe and mainstream Web communities.

Our work constitutes the first attempt to provide a multi-platform measurement of the meme ecosystem, with a focus on fringe and potentially dangerous communities. Considering the increasing relevance of digital information on world events, our study provides a building block for future cultural anthropology work, as well as for building systems to protect against the dissemination of harmful ideologies. Moreover, our pipeline can already be used by social network providers to assist the identification of hateful content; for instance, Facebook is taking steps to ban Pepe the Frog used in the context of hate [292], and our methodology can help them automatically identify hateful variants. Finally, our pipeline can be used for tracking the propagation of images from any context or other language spheres, provided an appropriate annotation dataset.

Performance. We also measured the time that it takes to associate images posted on Web communities to memes. All other steps in our system are one-time batch tasks, only executed if the annotations dataset is updated. To ease presentation, we only report the time to compare all the 74M images from Twitter (the largest dataset) against the medoids of all 12K annotated clusters: it took about 12 days on our infrastructure, equipped with two NVIDIA Titan Xp GPUs. This corresponds to 14ms per image, or 73 images per second. Note that, if new GPUs are added to our infrastructure, the workload would be divided equally across all GPUs.

Chapter 5

Characterizing the Role of Emerging Web Communities and Services on the Information Ecosystem

In this chapter, we study various Web communities and services, with a particular focus on understanding their role in the spread of information on the Web. Specifically, we study Gab with the goal to understand and characterize the platform with respect to the content and users it attracts. Also, we study Web archiving services (services that archive Web content) and how they are used by users on Twitter, Reddit, 4chan, and Gab.

5.1 What is Gab?

5.1.1 Motivation

The Web's information ecosystem is composed of multiple communities with varying influence [66]. As mainstream online social networks become less novel, users have begun to join smaller, more focused platforms. In particular, as the former have begun to reject fringe communities identified with racist and aggressive behavior, a number of alt-right focused services have been created. Among these emerging communities, the Gab social network has attracted the interest of a large number of users since its creation in 2016 [293], a few months before the US Presidential Election. Gab was created, ostensibly as a censorship-free platform, aiming to protect free speech above anything else. From the very beginning, site

operators have welcomed users banned or suspended from platforms like Twitter for violating terms of service, often for abusive and/or hateful behavior. In fact, there is extensive anecdotal evidence that the platform has become the alt-right's new hub [294] and that it exhibits a high volume of hate speech [26] and racism [8]. As a result, in 2017, both Google and Apple rejected Gab's mobile apps from their stores because of hate speech [26] and non-compliance to pornographic content guidelines [295].

In this work, we provide, to the best of our knowledge, the first characterization of the Gab social network. We crawl the Gab platform and acquire 22M posts by 336K users over a 1.5 year period (August 2016 to January 2018). Overall, the main findings of our analysis include:

1. Gab attracts a wide variety of users, ranging from well-known alt-right personalities like Milo Yiannopoulos to conspiracy theorists like Alex Jones. We also find a number of "troll" accounts that have migrated over from other platforms like 4chan, or that have been heavily inspired by them.
2. Gab is predominantly used for the dissemination and discussion of world events, news, as well as conspiracy theories. Interestingly, we note that Gab reacts strongly to events related to white nationalism and Donald Trump.
3. Hate speech is extensively present on the platform, as we find that 5.4% of the posts include hate words. This is 2.4 times higher than on Twitter, but 2.2 times lower than on 4chan's Politically Incorrect board (/pol/) [19].
4. There are several accounts making coordinated efforts towards recruiting millennials to the alt-right.

In summary, our analysis highlights that Gab appears to be positioned at the border of mainstream social networks like Twitter and "fringe" Web communities like 4chan's /pol/. We find that, while Gab claims to be all about free speech, this seems to be merely a shield behind which its alt-right users hide.

5.1.2 Dataset

Using Gab's API, we crawl the social network using a snowball methodology. Specifically, we obtain data for the most popular users as returned by Gab's API and iteratively collect data from all their followers as well as their followings. We collect three types of information: 1) basic details about Gab accounts, including username, score, date of account creation; 2) all

Followers			Scores			PageRank		
Name	**Username**	**#**	**Name**	**Username**	**#**	**Name**	**Username**	**PR score**
Milo Yiannopoulos	m	45,060	Andrew Torba	a	819,363	Milo Yiannopoulos	m	0.013655
PrisonPlanet	PrisonPlanet	45,059	John Rivers	JohnRivers	606,623	Andrew Torba	a	0.012818
Andrew Torba	a	38,101	Ricky Vaughn	Ricky_Vaughn99	496,962	PrisonPlanet	PrisonPlanet	0.011762
Ricky Vaughn	Ricky_Vaughn99	30,870	Don	Don	368,698	Mike Cernovich	Cernovich	0.006549
Mike Cernovich	Cernovich	29,081	Jared Wyand	JaredWyand	281,798	Ricky Vaughn	Ricky_Vaughn99	0.006143
Stefan Molyneux	stefanmolyneux	26,337	[*omitted*]	TukkRivers	253,781	Sargon of Akkad	Sargonofakkad100	0.005823
Brittany Pettibone	BrittPettibone	24,799	Brittany Pettibone	BrittPettibone	244,025	[*omitted*]	d_seaman	0.005104
Jebs	DeadNotSleeping	22,659	Tony Jackson	USMC-Devildog	228,370	Stefan Molyneux	stefanmolyneux	0.004830
[*omitted*]	TexasYankee4	20,079	[omitted]	causticbob	228,316	Brittany Pettibone	BrittPettibone	0.004218
[*omitted*]	RightSmarts	20,042	Constitutional Drunk	USSANews	224,261	Vox Day	voxday	0.003972
Vox Day	voxday	19,454	Truth Whisper	truthwhisper	206,516	Alex Jones	RealAlexJones	0.003345
[*omitted*]	d_seaman	18,080	Andrew Anglin	AndrewAnglin	203,437	Lauren Southern	LaurenSouthern	0.002984
Alex Jones	RealAlexJones	17,613	Kek_Magician	Kek_Magician	193,819	Donald J Trump	realdonaldtrump	0.002895
Jared Wyand	JaredWyand	16,975	[*omitted*]	shorty	169,167	Dave Cullen	DaveCullen	0.002824
Ann Coulter	AnnCoulter	16,605	[*omitted*]	SergeiDimitrovicIvanov	169,091	[*omitted*]	e	0.002648
Lift	lift	16,544	Kolja Bonke	KoljaBonke	160,246	Chuck C Johnson	Chuckcjohnson	0.002599
Survivor Medic	SurvivorMed	16,382	Party On Weimerica	CuckShamer	155,021	Andrew Anglin	AndrewAnglin	0.002599
[*omitted*]	SalguodNos	16,124	PrisonPlanet	PrisonPlanet	154,829	Jared Wyand	JaredWyand	0.002504
Proud Deplorable	luther	15,036	Vox Day	voxday	150,930	Pax Dickinson	pax	0.002400
Lauren Southern	LaurenSouthern	14,827	W.O. Cassity	wocassity	144,875	Baked Alaska	apple	0.002292

Table 5.1: Top 20 popular users on Gab according to the number of followers, their score, and their ranking based on PageRank in the followers/followings network. We omit the "screen names" of certain accounts for ethical reasons.

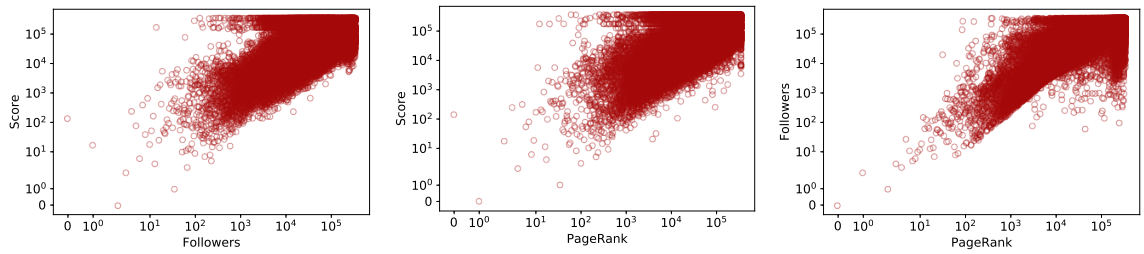
the posts for each Gab user in our dataset; and 3) all the followers and followings of each user that allow us to build the following/followers network. Overall, we collect 22,112,812 posts from 336,752 users, between August 2016 and January 2018.

5.1.3 Analysis

In this section, we provide our analysis on the Gab platform. Specifically, we analyze Gab's user base and posts that get shared across several axes.

Ranking of users

To get a better handle on the interests of Gab users, we first examine the most popular users using three metrics: 1) the number of followers; 2) user account score; and 3) user PageRank. These three metrics provide us a good overview of things in terms of "reach," appreciation of content production, and importance in terms of position within the social network. We report the top 20 users for each metric in Table 5.1. Although we believe that their existence in Table 5.1 is arguably indicative of their public figure status, for ethical reasons, we omit the "screen names" for accounts in cases where a potential link between the screen name and the



(a) (b) (c)

Figure 5.1: Correlation of the rankings for each pair of rankings: (a) Followers - Score; (b) PageRank - Score; and (c) PageRank - Followers.

user's real life names existed *and* it was unclear to us whether or not the user is a public figure. While Twitter has many celebrities in the most popular users [296], Gab seems to have what can at best be described as alt-right celebrities like Milo Yiannopoulos and Mike Cernovich.

Number of followers. The number of followers that each account has can be regarded as a metric of impact on the platform, as a user with many followers can share its posts to a large number of other users. We observe a wide variety of different users; 1) popular alt-right users like Milo Yiannopoulos, Mike Cernovich, Stefan Molyneux, and Brittany Pettibone; 2) Gab's founder Andrew Torba; and 3) popular conspiracy theorists like Alex Jones. Notably lacking are users we might consider as counter-points to the alt-right right, an indication of Gab's heavily right-skewed user-base.

Score. The score of each account is a metric of content popularity, as it determines the number of up-votes and down-votes that they receive from other users. In other words, is the degree of appreciation from other users. By looking at the ranking using the score, we observe two new additional categories of users: 1) users purporting to be news outlets, likely pushing false or controversial information on the network like PrisonPlanet and USSANews; and 2) troll users that seem to have migrated from or been inspired by other platforms (e.g., 4chan) like Kek_Magician and CuckShamer.

PageRank. We also compute PageRank on the followers/followings network and we rank the users according to the obtained score. We use this metric as it quantifies the structural importance of nodes within a network according to its connections. Here, we observe some interesting differences from the other two rankings. For example, the account with username "realdonaldtrump," an account reserved for Donald Trump, appears in the top users mainly because of the extremely high number of users that follow this account, despite the fact that it

Word	(%)	Bigram	(%)
maga	4.35%	free speech	1.24%
twitter	3.62%	trump supporter	0.74%
trump	3.53%	night area	0.49%
conservative	3.47%	area wanna	0.48%
free	3.08%	husband father	0.45%
love	3.03%	check link	0.42%
people	2.76%	freedom speech	0.41%
life	2.70%	hey guys	0.40%
like	2.67%	donald trump	0.40%
man	2.49%	man right	0.39%
truth	2.46%	america great	0.39%
god	2.45%	link contracts	0.35%
world	2.44%	wanna check	0.34%
freedom	2.29%	make america	0.34%
right	2.27%	need man	0.34%
american	2.25%	guys need	0.33%
want	2.23%	president trump	0.32%
one	2.20%	guy sex	0.31%
christian	2.17%	click link	0.30%
time	2.14%	link login	0.30%

Table 5.2: Top 20 words and bigrams found in the descriptions of Gab users.

has no posts or score.

Comparison of rankings. To compare the three aforementioned rankings, we plot the ranking of all the users for each pair of rankings in Fig. 5.1. We observe that the pair with the most agreement is PageRank-Followers (Fig. 5.1(c)), followed by the pair Followers-Score (Fig, 5.1(a)), while the pair with the least agreement is PageRank - Score (Fig 5.1(b). Overall, for all pairs we find a varying degree of rank correlation. Specifically, we calculate the Spearman's correlation coefficient for each pair of rankings; finding 0.53, 0.42, 0.26 for PageRank-Followers, Followers-Score, and PageRank-Score, respectively. While these correlations are not terribly strong, they are significant ($p < 0.01$) for the two general classes of users: those that play an important structural role in the network, perhaps encouraging the diffusion of information, and those that produce content the community finds valuable.

User account analysis

User descriptions. To further assess the type of users that the platform attracts we analyze the description of each created account in our dataset. Note that by default Gab adds a quote from a famous person as the description of each account and a user can later change it. Although not perfect, we look for any user description enclosed in quotes with a "–" followed by a

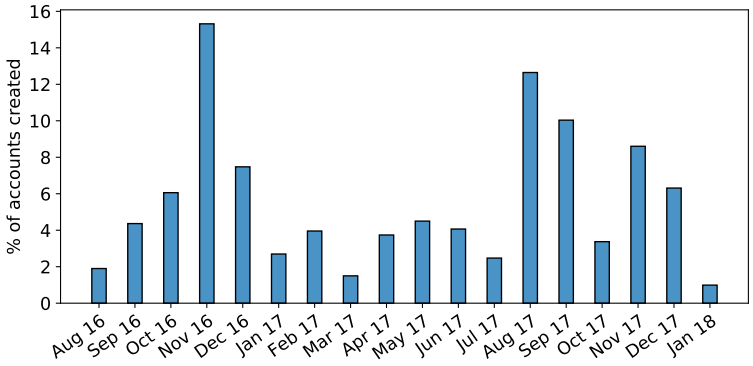
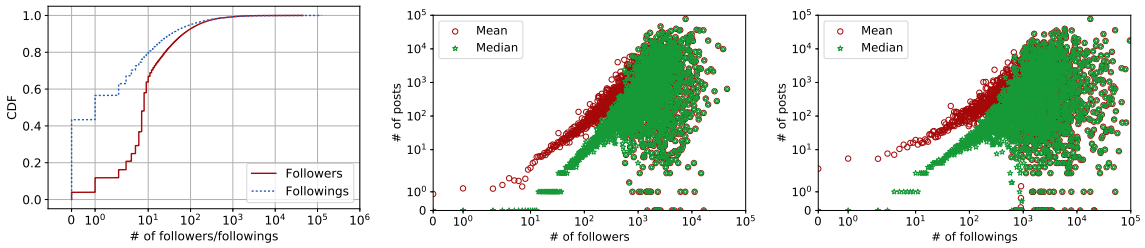


Figure 5.2: Percentage of accounts created per month.

name, and assume it is a default quote. Using this heuristic, we find that only 20% of the users actively change their description from the default. Table 5.2 reports the top words and bigrams found in customized descriptions (we remove stop words for more meaningful results). Examining the list, it is apparent that Gab users are conservative Americans, religious, and supporters of Donald Trump and "free speech." We also find some accounts that are likely bots and trying to deceive users with their descriptions; among the top bigrams there some that nudge users to click on URLs, possibly malicious, with the promise that they will get sex. For example, we find many descriptions similar to the following: "*Do you wanna get sex tonight? One step is left ! Click the link - < url >*." It is also worth noting that our account (created for crawling the platform) was followed by 12 suspected bot accounts between December 2017 and January 2018 without making any interactions with the platform (i.e., our account has never made a post or followed any user).

User account creation. We also look when users joined the Gab platform. Fig. 5.2 reports the percentage of accounts created for each month of our dataset. Interestingly, we observe that we have peaks for account creation on November 2016 and August 2017. These findings highlight the fact that Gab became popular during notable world and politics events like the 2016 US elections as well as the Charlottesville Unite the Right rally [297]. Finally, only a small percentage of Gab's users are either pro or verified, 0.75% and 0.5%, respectively, while 1.7% of the users have a private account (i.e., only their followers can see their gabs).

Followers/Followings. Fig. 5.3 reports our analysis based on the number of followers and followings for each user. From Fig. 5.3(a) we observe that in general Gab users have a larger number of followers when compared with following users. Interestingly, 43% of users are following zero other users, while only 4% of users have zero followers. I.e., although counter-intuitive, most users have more followers than users they follow. Figs. 5.3(b) and



(a) (b) (c)

Figure 5.3: Followers and Following analysis (a) CDF of number of followers and following (b) number of followers and number of posts and (c) number of following and number of posts.

5.3(c) show the number of followers and following in conjunction with the number of posts for each Gab user. We bin the data in log-scale bins and we report the mean and median value for each bin. We observe that in both cases, that there is a near linear relationship with the number of posts and followers/followings up until around 10 followers/followings. After this point, we see this relationship diverge, with a substantial number of users with huge numbers of posts, some over 77K. This demonstrates the extremely heavy tail in terms of content production on Gab, as is typical of most social medial platforms.

Reciprocity. From the followers/followings network we find a low level of reciprocity: specifically, only 29.5% of the node pairs in the network are connected both ways, while the remaining 71.5% are connected one way. When compared with the corresponding metric on Twitter [296], these results highlight that Gab has a larger degree of network reciprocity indicating that the community is more tightly-knit, which is expected when considering that Gab mostly attracts users from the same ideology (i.e., alt-right community).

Posts Analysis

Basic Statistics. First, we note that 63% of the posts in our dataset are original posts while 37% are reposts. Interestingly, only 0.14% of the posts are marked as NSFW. This is surprising given the fact that one of the reasons that Apple rejected Gab's mobile app is due to the share of NSFW content [295]. From browsing the Gab platform, we also can anecdotally confirm the existence of NSFW posts that are not marked as such, raising questions about how Gab moderates and enforces the use of NSFW tags by users. When looking a bit closer at their policies, Gab notes that they use a 1964 United States Supreme Court Ruling [298] on pornography that provides the famous "I'll known it when I see it" test. In any case, it would

Domain	(%)	Domain	(%)
youtube.com	4.22%	zerohedge.com	0.53%
youtu.be	2.67%	twimg.com	0.53%
twitter.com	1.96%	dailycaller.com	0.49%
breitbart.com	1.44%	t.co	0.47%
bit.ly	0.82%	ussanews.com	0.46%
thegatewaypundit.com	0.74%	dailymail.co.uk	0.46%
kek.gg	0.69%	tinyurl.com	0.44%
imgur.com	0.68%	wordpress.com	0.43%
sli.mg	0.61%	foxnews.com	0.41%
infowars.com	0.56%	blogspot.com	0.32%

Table 5.3: Top 20 domains in posts and their respective percentage over all posts.

seem that Gab's social norms are relatively lenient with respect to what is considered NSFW.

We also look into the languages of the posts, as returned by Gab's API. We find that Gab's API does not return a language code for 56% of posts. By looking at the dataset, we find that all posts before June 2016 do not have an associated language; possibly indicating that Gab added the language field afterwards. Nevertheless, we find that the most popular languages are English (40%), Deutsch (3.3%), and French (0.14%); possibly shedding light to Gab's users locations which are mainly the US, the UK, and Germany.

URLs. Next, we assess the use of URLs in Gab; overall we find 3.5M unique URLs from 81K domains. Table 6.14 reports the top 20 domains according to their percentage of inclusion in all posts. We observe that the most popular domain is YouTube with almost 7% of all posts, followed by Twitter with 2%. Interestingly, we note the extensive use of alternative news sources like Breitbart (1.4%), The Gateway Pundit (0.7%), and Infowars (0.5%), while mainstream news outlets like Fox News (0.4%) and Daily Mail (0.4%) are further below. Also, we note the use of image hosting services like Imgur (0.6%), `sli.mg` (0.6%), and kek.gg (0.7%) and URL shorteners like `bit.ly` (0.8%) and `tinyurl.com` (0.4%). Finally, it is worth mentioning that The Daily Stormer, a well known neo-Nazi web community is five ranks ahead of the most popular mainstream news source, The Hill.

Hashtags & Mentions As discussed in Chapter 2, Gab supports the use of hashtags and mentions similar to Twitter. Table 5.4 reports the top 20 hashtags/mentions that we find in our dataset. We observe that the majority of the hashtags are used in posts about Trump, news, and

Hashtag	**(%)**	**Mention**	**(%)**
MAGA	6.06%	a	0.69%
GabFam	4.22%	TexasYankee4	0.31%
Trump	3.01%	Stargirlx	0.26%
SpeakFreely	2.28%	YouTube	0.24%
News	2.00%	support	0.23%
Gab	0.88%	Amy	0.22%
DrainTheSwamp	0.71%	RaviCrux	0.20%
AltRight	0.61%	u	0.19%
Pizzagate	0.57%	BlueGood	0.18%
Politics	0.53%	HorrorQueen	0.17%
PresidentTrump	0.47%	Sockalexis	0.17%
FakeNews	0.41%	Don	0.17%
BritFam	0.37%	BrittPettibone	0.16%
2A	0.35%	TukkRivers	0.15%
maga	0.32%	CurryPanda	0.15%
NewGabber	0.28%	Gee	0.15%
CanFam	0.27%	e	0.14%
BanIslam	0.25%	careyetta	0.14%
MSM	0.22%	PrisonPlanet	0.14%
1A	0.21%	JoshC	0.12%

Table 5.4: Top 20 hashtags and mentions found in Gab. We report their percentage over all posts.

politics. We note that among the top hashtags are "AltRight", indicating that Gab users are followers of the alt-right movement or they discuss topics related to the alt-right; "Pizzagate", which denotes discussions around the notorious conspiracy theory [20]; and "BanIslam", which indicate that Gab users are sharing their islamophobic views. It is also worth noting the use of hashtags for the dissemination of popular memes, like the Drain the Swamp meme that is popular among Trump's supporters [299]. When looking at the most popular users that get mentioned, we find popular users related to the Gab platform like Andrew Torba (Gab's CEO with username @a).

We also note users that are popular with respect to mentions, but do *not* appear in Table 5.1's lists of popular users. For example, Amy is an account purporting to be Andrew Torba's mother. The user Stargirlx, who we note changed usernames three times during our collection period, appears to be an account presenting itself as a millennial "GenZ" young woman.

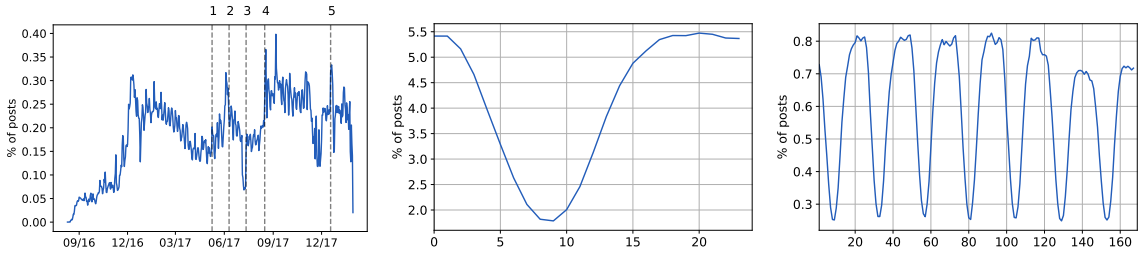
Topic	(%)	Category	(%)
Deutsch	2.29%	News	15.91%
BritFam	0.73%	Politics	10.30%
Introduce Yourself	0.59%	AMA	4.46%
International News	0.19%	Humor	3.50%
DACA	0.17%	Technology	1.44%
Las Vegas Terror Attack	0.16%	Philosophy	1.06%
Hurricane Harvey	0.16%	Entertainment	1.01%
Gab Polls	0.13%	Art	0.72%
London	0.12%	Faith	0.69%
2017 Meme Year in Review	0.12%	Science	0.56%
Twitter Purge	0.12%	Music	0.52%
Seth Rich	0.11%	Sports	0.39%
Memes	0.11%	Photography	0.37%
Vegas Shooting	0.11%	Finance	0.31%
Judge Roy Moore	0.09%	Cuisine	0.16%

Table 5.5: Top 15 categories and topics found in the Gab dataset

Interestingly, it seems that Amy and Stargirlx have been organizing Gab "chats," which are private groups of users, for 18 to 29 year olds to discuss politics; possibly indicating efforts to recruit millennials to the alt-right community.

Categories & Topics. As discussed in Chapter 2, gabs may be part of a topic or category. By analyzing the data, we find that this happens for 12% and 42% of the posts for topics and categories, respectively. Table 5.5 reports the percentage of posts for each category as well as for the top 15 topics. For topics, we observe that the most popular are general "Ask Me Anything" (AMA) topics like Deutsch (2.29%, for German users), BritFam (0.73%, for British users), and Introduce Yourself (0.59%). Furthermore, other popular topics include world events and news like International News (0.59%), Las Vegas shooting (0.27%), and conspiracy theories like Seth Rich's Murder (0.11%). When looking at the top categories we find that by far the most popular categories are News (15.91%) and Politics (10.30%). Other popular categories include AMA 4.46%), Humor (3.50%), and Technology (1.44%).

These findings highlight that Gab is heavily used for the dissemination and discussion of world events and news. Therefore, its role and influence on the Web's information ecosystem should be assessed in the near future. Also, this categorization of posts can be of great importance



(a) Date (b) Hour of Day (c) Hour of Week

Figure 5.4: Temporal analysis of the Gab posts (a) each day; (b) based on hour of day and (c) based on hour of week.

for the research community as it provides labeled ground truth about discussions around a particular topic and category.

Hate speech assessment. As previously discussed, Gab was openly accused of allowing the dissemination of hate speech. In fact, Google removed Gab's mobile app from its Play Store because it violates their hate speech policy [26]. Due to this, we aim to assess the extent of hate speech in our dataset. Using the modified Hatebase [300] dictionary used by the authors of [19], we find that 5.4% of all Gab posts include a hate word. In comparison, Gab has 2.4 times the rate of hate words when compared to Twitter, but less than halve the rate of hate words compared to 4chan's Politically Incorrect board (/pol/) [19]. These findings indicate that Gab resides on the border of mainstream social networks like Twitter and fringe Web communities like 4chan's Politically Incorrect (/pol/) board.

Temporal Analysis. Finally, we study the posting behavior of Gab users from a temporal point of view. Fig. 6.13 shows the distribution of the Gab posts in our dataset according to each day of our dataset, as well as per hour of day and week (in UTC). We observe that the general trend is that the number of Gab's posts increase over time (Fig. 5.8(a)); this indicates an increase in Gab's popularity. Furthermore, we note that Gab users posts most of their gabs during the afternoon and late night (after 3 PM UTC) while they rarely post during the morning hours (Fig. 5.8(b)). Also, the aforementioned posting behavior follow a diurnal weekly pattern as we show in Fig. 5.4(c).

To isolate significant days in the time series in Fig. 5.8(a), we perform a *changepoint analysis* using the Pruned Exact Linear Time (PELT) method [11]. First, we use our knowledge of the weekly variation in average post numbers from Fig. 5.4(c) to subtract from our timeseries the mean number of posts for each day. This leaves us with a mean-zero timeseries of the deviation

of the number of posts per day from the daily average. We assume that this timeseries is drawn from a normal distribution, with mean and variance that can change at a discrete number of changepoints. We then use the PELT algorithm to maximize the log-likelihood function for the mean(s) and variance(s) of this distribution, with a penalty for the number of changepoints. By ramping down the penalty function, we produce a ranking of the changepoints.

Examining current events around these changepoints provides insight into they dynamics that drive Gab behavior. First, we note that there is a general increase in activity up to the Trump inauguration, at which point activity begins to decline. When looking later down the timeline, we see an increase in activity after the changepoint marked **1** in Fig. 5.8(a). Changepoint **1** coincides with James Comey's firing from the FBI, and the relative acceleration of the Trump-Russian collusion probe [301].

The next changepoint (**2**) coincides with the so-called "March Against Sharia" [302] organized by the alt-right, with the event marked **4** corresponding to Trump's "blame on both sides" response to violence at the Unite the Right Rally in Charlottesville [303]. Similarly, we see a meaningful response to Twitter's banning of abusive users [304] marked as changepoint **5**.

Changepoint **3**, occurring on July 12, 2017 is of particular interest, since it is the most extreme *reduction* in activity recognized as a changepoint. From what we can tell, this is a reaction to Donald Trump Jr. releasing emails that seemingly evidenced his meeting with a Russian lawyer to receive compromising intelligence on Hillary Clinton's campaign [305]. I.e., the disclosure of evidence of collusion with Russia corresponded to the single largest drop in posting activity on Gab.

5.1.4 Remarks

In this work, we have provided the first characterization of a new social network called Gab. We analyzed 22M posts from 336K users, finding that Gab attracts the interest of users ranging from alt-right supporters and conspiracy theorists to trolls. We showed that Gab is extensively used for the discussion of news, world events, and politics-related topics, further motivating the need take it into account when studying information cascades on the Web. By looking at the posts for hate words, we also found that 5.4% of the posts include hate words. Finally, using changepoint analysis, we highlighted how Gab reacts very strongly to real-world events focused around white nationalism and support of Donald Trump.

5.2 Understanding Web Archiving Services and their Use on Multiple Web Communities

5.2.1 Motivation

In today's digital society, the availability and persistence of Web resources are very relevant issues. A substantial number of URLs shared on the Web becomes unavailable after some time as websites are shutdown or redesigned in a way that does not preserve old URLs – a phenomenon known as *"link rot"* [9]. Moreover, content might be taken down by authorities on a legal basis, deleted by users who have shared it on social media, removed as per the "right to be forgotten" [306], etc. Overall, the ephemerality of Web content often prompts debate with respect to its impact on the availability of information, accountability, or even censorship.

In this context, an important role is played by services like the Wayback Machine (`archive.org`), which *proactively* archives large portions of the Web, allowing users to search and retrieve the history of more than 300 billion pages. At the same time, *on-demand* archiving services like `archive.is` have also become popular: users can take a snapshot of a Web page by entering its URL, which the system crawls and archives, returning a permanent short URL serving as a time capsule that can be shared across the Web.

Archiving services serve a variety of purposes beyond addressing link rot. Platforms like `archive.is` are reportedly used to preserve controversial blogs and tweets that the author may later opt to delete [307]. Moreover, they also reduce Web traffic toward "source URLs" when the original content is still accessible, thus depriving them of potential ad revenue streams (users do not visit the original site, but just the archived copy). In fact, anecdotal evidence has emerged that alt-right communities target outlets they disagree with by nudging their users to share archive URLs instead [308], or discrediting them by pointing at earlier versions of articles [309].

Given the role in helping content persist, their use on social networks, as well as anecdotal evidence of their misuse in contexts where information could be weaponized [310], archiving services are arguably impactful actors that should be thoroughly analyzed. To this end, we aim to shed light on the Web archiving ecosystem, aiming to answer the following research questions: How are archive URLs disseminated across popular social networks? What kind of content gets archived, by whom and why? Are archiving services misused in any way?

To answer these questions, we perform a large-scale quantitative analysis of Web archives, based on two data sources: 1) 21M URLs collected from the `archive.is` live feed, and 2) 356K `archive.is` plus 391K Wayback Machine URLs that were shared on four social networks: Reddit, Twitter, Gab, and 4chan's Politically Incorrect board (/pol/).

Our main findings include:

1. News and social media posts are the most common types of content archived, likely due to their (perceived) ephemeral and/or controversial nature.

2. URLs of archiving services are extensively shared on "fringe" communities within Reddit and 4chan to preserve possibly contentious content, or to refer to it without increasing the Web traffic to the source. We also find that /pol/ and Gab users favor archive.is over Wayback Machine (respectively, 15x and 16x), highlighting a particular use case in "controversial" online communities.

3. Web archives are exploited by users to bypass censorship policies in some communities: for instance, /pol/ users post `archive.is` URLs to share content from 8chan and Facebook, which are banned on the platform, or to circumvent accidental censorship of some news sources because of substitution filters (e.g., 'smh' becomes 'baka', so links to `smh.com.au` are unusable).

4. Reddit bots are responsible for posting a very large portion of archive URLs in Reddit (respectively, 44% and 85% of `archive.is` and Wayback Machine URLs). This is due to moderators aiming to alleviate the effects of link rot on the platform; however, this pro-active archival of content also impact traffic to archived sites originating from Reddit.

5. The_Donald subreddit systematically targets ad revenue of news sources with conflicting ideologies: moderation bots block URLs from those sites and prompt users to post archive URLs instead (some domains, e.g. `nydailynews.com`, have up to 46% of their content censored). According to our conservative estimates, popular news sources like the Washinghton Post lose yearly approximately $70K from their ad revenue because of the use of archiving services on Reddit.

5.2.2 Background

Our analysis focuses on two popular archiving services: `archive.is` and the Wayback Machine (`archive.org`). The former stores *snapshots* of Web pages upon request, while the latter is run by a non-profit organization (the Internet Archive) aiming to archive pages mainly through a constant crawling process.

Archive.is offers a free, on-demand archival service of Web pages: a user visits the service and enters a URL to be archived. It also acts as a link shortener which obfuscates the source URL, by generating a 5-character URL. For instance, `http://archive.is/HVbU` shows the snapshot of Google's homepage, archived on July, 03, 2012 at 07:03:24.

Wayback Machine. Launched in 2001, the Wayback Machine archives a large portion of Web content, storing periodic snapshots of various pages. It mainly works through a proactive crawler[1], which visits various sites and captures a snapshot of the content. However, users can also trigger information archival on demand. When a page is archived, an archive URL is created in the following format: https://web.archive.org/web/[*time of archival*]/[*source URL*]. For example, the archive URL `https://web.archive.org/web/20100205062719/http://www.google.com/` returns the version of Google's home page on February 5, 2010, at 06:27:19 (UTC). In the rest of the thesis, we refer to the URLs generated by archiving services as ***archive URLs***, and to the archived URL as ***source URLs***.

We opt to study the Wayback Machine and `archive.is` for a few reasons. First of all, they are popular services: as of Jan 2018, their Alexa Global Rank is, resp., 300 and 2,920. The Wayback Machine is actually one of the oldest initiatives, with about 300 billion pages archived as of 2017. We also choose these two because of some important differences between them. The Wayback Machine is run by a 501(c)(3) non-profit organization, while `archive.is` is hosted by Russian provider Hostkey (interestingly, it is only accessible via HTTP in Russia). Moreover, the former respects robots exclusion standards (even retroactively) and generally gives website owners the right to request removal of pages from the archive, while the latter only complies (albeit inconsistently) with DMCA take-down requests. Finally, `archive.is` is reportedly used in "fringe" Web communities within 4chan and Reddit, which are known for generating [20] and incubating [231] fake news stories, and for their influence on the information ecosystem [66].

[1] `http://crawler.archive.org/index.html`

Platform	Archive	#Posts with Archive URLs (%all posts)	Archive URLs	Source URLs	Source Domains	Filtered
Live Feed	archive.is		21,537,554	20,608,834	5,388,112	-
Reddit	archive.is	327,050 ($2.9{\cdot}10^{-4}$%)	310,392	291,382	15,994	35.70%
	Wayback	320,379 ($2.8{\cdot}10^{-4}$%)	387,081	343,851	21,124	17.20%
/pol/	archive.is	46,912 ($1.1{\cdot}10^{-3}$%)	36,277	33,824	3,970	4.67%
	Wayback	3,848 ($9.7{\cdot}10^{-5}$%)	2,325	2,207	976	83.12%
Gab	archive.is	6,602 ($3.4{\cdot}10^{-4}$%)	5,943	5,773	1,300	5.54%
	Wayback	478 ($5.1{\cdot}10^{-5}$%)	361	349	240	61.18%
Twitter	archive.is	6,750 ($3.1{\cdot}10^{-6}$%)	3,772	3,669	845	8.23%
	Wayback	1,905 ($9.0{\cdot}10^{-7}$%)	1,290	1,257	846	7.49%

Table 5.6: Overview of our datasets: number and percentage of posts that include archive URLs, unique number of archive URLs, source URLs, and source domains. We also filter URLs that are malformed, unreachable, or point to resources other than Web pages.

5.2.3 Datasets

We now present the datasets studied in our work as well as our data collection methodology. We perform two crawls: 1) `archive.is` URLs obtained from the live feed page and 2) Wayback Machine and `archive.is` URLs posted on four social networks, namely, Twitter, Reddit, Gab, and 4chan's /pol/. The resulting datasets are summarized in Table 5.6.

`Archive.is` live feed. To gather a large dataset of `archive.is` generated URLs, we use the live feed page (`http://archive.is/livefeed/`), which provides a view of the archive based on archival time (e.g., the first page lists URLs archived in the previous 10 minutes). In August 2017, we crawl the first 100K pages of the live feed, acquiring 45.2M URLs, archived between October 7, 2015 and August 26, 2017.

Next, we visit the `archive.is` URLs, and scrape the content to get the archival time and the source URL. To avoid issues for the site operators, we throttle our crawler and do not visit all 45.2M URLs. Instead, we randomly sample them while ensuring temporal coverage, visiting 21.5M (48%) archive URLs, corresponding to 20.6M unique source URLs from 5.3M unique domains. Note that given the substantial size of our sample, which guarantees temporal coverage over almost two years, the resulting dataset is representative of the archive. In other words, our sampling strategy does not likely introduce substantial biases affecting our results.

Archive URLs posted on social networks. We search for archive.is and Wayback Machine URLs on Twitter, Reddit, and /pol/, between Jul 1, 2016 and Aug 31, 2017, and on Gab between Aug 1, 2016–Aug 31, 2017. We obtain the 4chan dataset from the authors of [19], the Reddit one from pushshift.io, while, for Twitter, we rely on the 1% Streaming API.[2] For Gab, we use a snowball sampling by collecting popular users returned by Gab's API, and iteratively collecting posts for all their followers and users they follow.

Overall, the resulting dataset includes 50K posts from /pol/, 528K posts from Reddit, 7K posts from Gab, and about 9K tweets. Note that we have some gaps due to failure of our data collection infrastructure, specifically, there are 70 and 13 days missing for Twitter and /pol/, respectively.

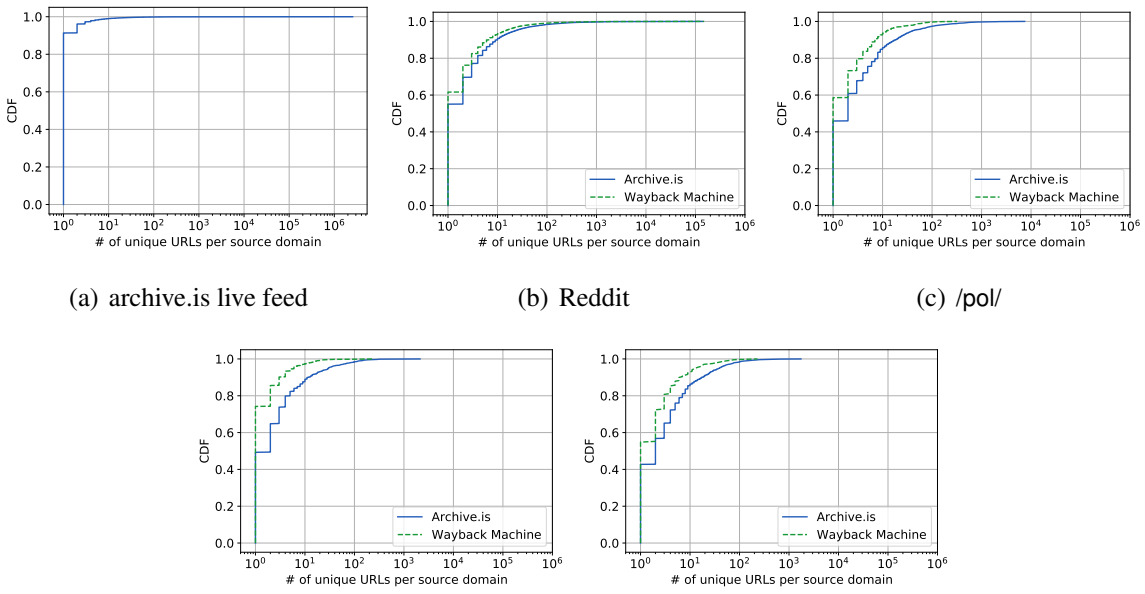
Basic Statistics. In Table 5.6, we report statistics from our archive.is live-feed crawl as well as the crawl of archive.is and Wayback Machine URLs shared on Twitter, Reddit, /pol/, and Gab. We report the number of posts with archive URLs, along with the percentage over the total number of posts, as well as the number of unique archive URLs, unique source URLs, unique source domains, and the percentage of URLs that are filtered out. Specifically, besides malformed URLs, we exclude, for archive.is, URLs unreachable between Aug 29 and Oct 7, 2017, while for Wayback Machine those pointing to types of information other than Web pages (e.g., images, videos, software, etc.).

Overall, /pol/ and Gab users often share Wayback Machine URLs that point to non-Web pages: around 83% and 61% of the total, respectively, suggesting that archive.is is used mostly for the dissemination of Web pages, while Wayback Machine is preferred for other content. Also, a high percentage of malformed archive.is URLs are shared on Reddit (35%), due to bots trying to pro-actively archive resources but failing. From the normalized percentages, we observe that Twitter users rarely share URLs from archiving services, while Reddit users do so from both archiving services. On /pol/ and Gab, we find 15 and 16 times, respectively, more archive.is URLs than Wayback Machine ones.

5.2.4 Cross-Platform Analysis

In this section, we present a cross-platform analysis of archive URLs collected from the archive.is live feed, as well as Wayback Machine and archive.is URLs shared on Twitter, Reddit, Gab, and /pol/. We focus on understanding what kind of content gets archived,

[2] https://dev.twitter.com/streaming/overview



(a) archive.is live feed (b) Reddit (c) /pol/

(d) Twitter (e) Gab

Figure 5.5: CDF of the number of distinct URLs per source domain.

as well as the related temporal characteristics, and on assessing whether archived content is still available from the source.

Source Domains

Live Feed. In Fig. 5.5(a), we plot the CDF of the number of distinct URLs per domain in our `archive.is` live feed dataset. The vast majority (90%) of domains only appear once, while a few domains yield a large numbers of archive URLs – e.g., there are 1.2M distinct `archive.is` URLs for which `twitter.com` is the source domain. In Table 5.7, we report the top 20 source domains as well as the top 20 domain suffixes (Sx). Surprisingly, the top domain (11.8%) is actually the Wayback Machine's `archive.org`. Mainstream social networks like Twitter and Facebook are also included, likely due to their (perceived) ephemerality, i.e., users want to preserve social network posts before they are removed or deleted. As for the suffixes, we observe that common ones, such as .com and .org, are the majority, followed by domains from Germany (.de) and Japan (.jp) with 7% and 5.6% of the URLs, respectively. This suggests that a substantial portion of `archive.is`'s user base might be in Germany and Japan.

Social Networks. In Figs 5.5(b)–5.5(e), we plot the CDF of the number of URLs for each

Domain	(%)	Sx	(%)	Domain	(%)	Sx	(%)
archive.org	11.82%	.com	38.29%	ru-board.com	0.50%	.pl	1.24%
twitter.com	5.73%	.org	17.64%	asstr.org	0.49%	.ch	1.23%
quora.com	3.18%	.de	7.02%	ruliweb.com	0.43%	.eu	1.01%
livejournal.com	2.17%	.jp	5.61%	4chan.org	0.40%	.se	0.80%
reddit.com	1.81%	.net	3.19%	googleusercontent.com	0.40%	.cz	0.69%
facebook.com	1.31%	.ru	3.10%	ameblo.jp	0.39%	.br	0.66%
nhk.or.jp	0.78%	.nl	2.56%	wordpress.com	0.38%	.at	0.63%
youtube.com	0.65%	.uk	1.51%	yahoo.co.jp	0.38%	.es	0.57%
wikipedia.org	0.52%	.it	1.39%	aaaaarg.fail	0.37%	.be	0.55%
tumblr.com	0.51%	.fr	1.39%	blogspot.nl	0.36%	.ca	0.51%

Table 5.7: Top 20 domains and suffixes of the source URLs in the archive.is live feed dataset.

source domain in each dataset, finding that over 40% of the source domains only appear once. Wayback Machine generally archives more URLs per source domain than archive.is, although for Reddit the distributions are quite similar. Then, in Tables 5.8–5.11, we report the top 20 source domains observed on each platform, along with their *archival fraction* (AF), i.e., the number of times a source domain appears in an archive over the total number of times it appears in the dataset (either archived or not).

On all platforms except for Gab, the most popular domain archived through archive.is is the platform itself; e.g., archives of tweets are the most shared ones on Twitter. This also happens for Wayback Machine URLs, but only on Reddit. On Reddit, this may be due to meta-subreddits focused on the preservation and discussion of dramatic happenings, e.g., flame wars and intra-Reddit conflict, that would otherwise be lost when deleted by moderators after some time. These meta-subreddits tend to make use of bots that automatically archive drama submitted by their members.

Overall, we notice a strong presence of both mainstream (e.g., Washington Post) and alternative (e.g., Breitbart) news sources archived and shared on Reddit, /pol/, and Gab. Moreover, on /pol/, archive.is is often used for links to hypothes.is, a service that lets users annotate news articles, possibly due to the fact that /pol/ users often "unravel" conspiracy theories by researching and commenting on news articles. On Twitter, where the footprint of archive URLs is relatively low, we find a relatively large number of Japanese domains, which might possibly indicate a stronger presence of Japanese Twitter users relying on archives.

The AFs are quite low overall, implying that archiving services disseminate a small fraction of most domains. However, on /pol/, specific domains have extremely high AFs. For instance, we

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
reddit.com	31.21%	<0.01	reddit.com	36.88%	<0.01
pastebin.com	6.80%	0.08	imgur.com	7.05%	<0.01
twitter.com	5.89%	<0.01	twitter.com	5.19%	<0.01
imgur.com	3.02%	<0.01	redd.it	4.79%	<0.01
washingtonpost.com	2.46%	0.02	youtube.com	3.90%	<0.01
youtube.com	2.33%	<0.01	washingtonpost.com	1.54%	0.01
redd.it	2.14%	<0.01	youtu.be	1.19%	<0.01
nytimes.com	1.76%	0.01	nytimes.com	0.98%	<0.01
cnn.com	1.64%	0.02	cnn.com	0.90%	<0.01
wikipedia.org	1.37%	<0.01	reddituploads.com	0.89%	0.06
huffingtonpost.com	0.93%	0.02	archive.is	0.61%	<0.01
theguardian.com	0.78%	<0.01	streamable.com	0.61%	<0.01
googleusercontent.com	0.65%	0.08	thehill.com	0.54%	0.01
politico.com	0.64%	0.02	wikipedia.org	0.52%	<0.01
wsj.com	0.61%	0.03	politico.com	0.49%	0.02
dailymail.co.uk	0.54%	0.01	theguardian.com	0.46%	<0.01
4chan.org	0.53%	0.16	rawstory.com	0.45%	0.06
facebook.com	0.52%	<0.01	huffingtonpost.com	0.44%	<0.01
thehill.com	0.43%	0.01	bbc.com	0.44%	0.01
breitbart.com	0.40%	0.01	kickstarter.com	0.37%	0.02

Table 5.8: Top 20 source domains of `archive.is` and Wayback Machine URLs, and archival fraction (AF), in the Reddit dataset.

find that `facebook.com` (AF = 0.96) and `8ch.net` (AF = 1.0) are marked as spam from /pol/, and posts including links to them are rejected, a phenomenon we refer to as *platform-specific censorship*. We manually analyze other domains with high AF values, specifically, `hypothes.is`, `chetlyzarko.com`, `tdbming.com`, `justice4germans.com`, and `jeffreyepsteinscience.com`, without finding evidence of censorship on /pol/. There is also "accidental" censorship on /pol/: for instance, the Australian newspaper `smh.com.au`, is affected because of a substitution filter (used for fun), which replace one word with another, as the word "smh" is automatically replaced on /pol/ with "baka."

URL Characterization

We now proceed to characterize the type of content archived. To this end, we extract the domain categories of source URLs using the free Virus Total API (`virustotal.com`), which we choose since it consolidates categories from multiple services including Bit Defender, TrendMicro, Alexa, etc. Although categorization is done at domain-level, results are presented at a per-URL level (a URL is assigned the same category as its domain) in order to capture the popularity of each domain in our datasets.

Live Feed. Due to throttling enforced by the API, we are not able to categorize all the 20.6M

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
4chan.org	9.35%	0.54	justice4germans.com	7.50%	**0.94**
theguardian.com	3.78%	0.13	chetlyzarko.com	3.90%	**1.00**
washingtonpost.com	3.70%	0.20	twitter.com	2.82%	<0.01
nytimes.com	3.46%	0.16	dailymail.co.uk	2.47%	<0.01
cnn.com	2.78%	0.14	revcom.us	2.16%	0.66
twitter.com	2.75%	0.01	reddit.com	1.98%	<0.01
independent.co.uk	2.37%	0.13	tumblr.com	1.85%	0.02
breitbart.com	1.96%	0.08	thebilzerianreport.com	1.57%	0.72
reddit.com	1.85%	0.09	jeffreyepsteinscience.com	1.55%	**1.00**
dailymail.co.uk	1.72%	0.05	cnn.com	1.51%	<0.01
facebook.com	1.69%	**0.96**	tdbimg.com	1.43%	**1.00**
huffingtonpost.com	1.37%	0.20	huffingtonpost.com	1.43%	0.01
thehill.com	1.21%	0.16	metapedia.org	1.22%	0.04
politico.com	1.04%	0.13	nytimes.com	1.15%	<0.01
bbc.com	1.01%	0.08	washingtonpost.com	1.11%	<0.01
8ch.net	0.98%	**1.00**	theguardian.com	1.08%	<0.01
googleusercontent.com	0.91%	0.59	independent.co.uk	1.08%	<0.01
hypothes.is	0.87%	**0.98**	wordpress.com	1.06%	<0.01
telegraph.co.uk	0.85%	0.03	idrsolutions.com	1.01%	0.86
theatlantic.com	0.81%	0.24	wikileaks.com	1.01%	<0.01

Table 5.9: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the /pol/ dataset.

source URLs in our archive.is live feed dataset. Therefore, we first aggregate URLs into their domain, then, we follow a sampling approach using: 1) the top 100K most popular domains in our dataset, which correspond to 15M (73%) source URLs, and 2) a sample of 121K domains drawn according to their empirical distribution in our archive datasets, resulting in 1.4M (7%) source URLs.

In Fig. 5.6, we report the top 15 categories obtained from Virus Total for both samples. Note that Virus Total is unable to provide a category for 1% and 7% of the URLs for the two sets of domains that we checked, respectively. From Fig. 5.6(a), we observe that the most popular category is Reference Materials (23%), which is due to the fact that, as discussed earlier, many archive.is URLs archive Wayback Machine URLs. Other popular categories include Social Networks (15%), News Sources (14%), Education (13%), and Business (12%). Adult Content accounts for 4% of source URLs. Fig. 5.6(b) shows that, for the empirically distributed sample, the top 15 categories are slightly different, including Business (21%), News (13%), and Adult Content (12%).

Social Networks. Unlike the live feed dataset, we perform URL characterization for *all* source URLs (aggregated by domain) found on Reddit, /pol/, Gab, and Twitter, again using the Virus Total API. In Fig. 5.7, we report the top categories and their corresponding percentages for both archiving services (specifically, the union of categories that appear in the top 10 categories

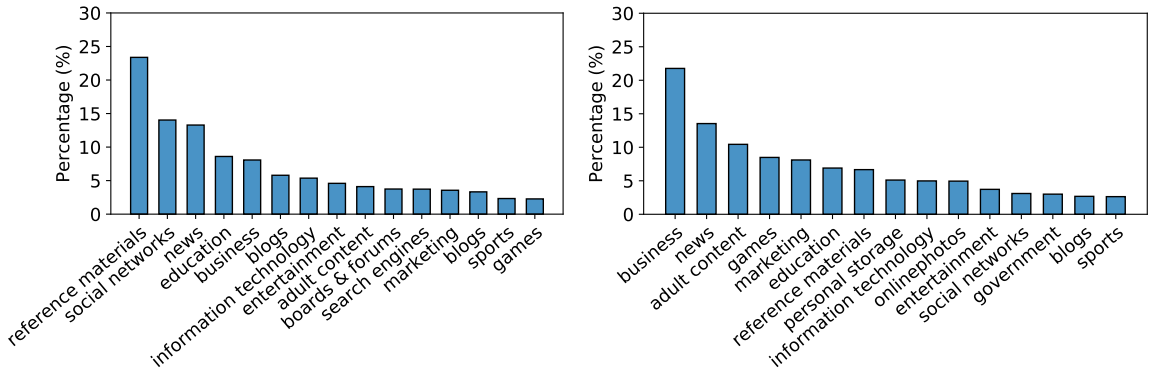
Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
twitter.com	25.02 %	<0.01	justpaste.it	11.90 %	0.02
facebook.com	3.65 %	<0.01	twitter.com	6.90 %	0.01
togetter.com	3.58 %	<0.01	dailymail.co.uk	1.95 %	0.13
seesaa.net	2.97 %	**0.91**	nikkansports.com	1.50 %	0.18
justpaste.it	2.19 %	0.01	mikelofgren.net	1.20%	**1.00**
yahoo.co.jp	2.03 %	0.21	blogspot.com	1.10%	0.09
googleusercontent.com	1.77 %	**0.98**	whitehouse.gov	1.05%	0.02
time.com	1.75 %	0.01	journalists-in-russia.org	1.00%	**1.00**
monjiro.net	1.66 %	0.51	pcdepot.co.jp	0.90%	**0.90**
pastebin.com	1.45 %	0.04	rydon.co.uk	0.85%	**1.00**
google.com	1.39 %	0.01	yeniakit.com.tr	0.85%	0.16
jimin.jp	1.35 %	**0.95**	cdse.edu	0.75%	**0.93**
notepad.cc	1.33 %	0.47	tetsureki.com	0.75%	**1.00**
ameblo.jp	1.16 %	<0.01	donaldjtrump.com	0.75%	0.04
nhk.or.jp	1.16 %	0.33	reidreport.com	0.75%	**1.00**
magi.md	1.16 %	0.49	ameblo.cjp	0.70%	<0.01
opensecrets.org	1.05 %	0.67	jreast.co.jp	0.70%	**0.93**
fc2.com	0.99 %	0.27	eastandard.net	0.65%	**1.00**
dailyshincho.jp	0.93 %	**0.94**	yahoo.co.jp	0.60%	0.01
reddit.com	0.89 %	0.03	livedoor.jp	0.60%	0.07

Table 5.10: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the Twitter dataset.

Domain (archive.is)	(%)	AF	Domain (Wayback)	(%)	AF
twitter.com	12.28%	<0.01	dailymail.co.uk	20.98%	< 0.01
nytimes.com	4.71%	0.03	washingtonpost.com	7.08%	0.01
washingtonpost.com	4.17%	0.03	infowars.com	5.54%	<0.01
reddit.com	3.10%	0.03	brandenburg.de	4.35%	0.10
googleusercontent.com	2.43%	0.18	twitter.com	3.63%	< 0.01
breitbart.com	1.82%	< 0.01	huffingtonpost.com	3.08%	<0.01
cnn.com	1.63%	0.01	abcnews.go.com	2.54%	< 0.01
4chan.org	1.59%	0.07	salon.com	1.72%	0.01
dailymail.co.uk	1.44%	<0.01	alexa.com	1.63%	0.03
theguardian.com	1.29%	< 0.01	news.com.au	1.54%	<0.01
wsj.com	1.22%	0.01	tu-dortmunt.de	1.45%	0.80
bbc.com	1.15%	0.01	causes.com	1.27%	0.50
huffingtonpost.com	1.14%	0.03	vigilantcitizen.com	1.18%	0.02
google.com	1.01%	< 0.01	reddit.com	1.08%	<0.01
facebook.com	0.92%	< 0.01	sahra-wagenknecht.de	0.99%	0.78
latimes.com	0.85%	0.01	quillette.com	0.99%	0.02
yahoo.com	0.81%	< 0.01	derwesten.de	0.99%	<0.01
dailycaller.com	0.77%	< 0.01	politico.com	0.91%	<0.01
thehill.com	0.74%	< 0.01	mikelofgren.net	0.81%	**0.90**
wikileaks.org	0.73%	0.01	alexanderhiggins.com	0.81%	0.02

Table 5.11: Top 20 source domains of archive.is and Wayback Machine URLs, and archival fraction (AF), in the Gab dataset.

for each service). The Virus Total API is unable to provide a category for, on average, 1.5% and 9% of the archive.is and Wayback Machine URLs found on Reddit, /pol/, Gab, and Twitter, respectively. Overall, both archiving services are often used to disseminate URLs from



(a) Top 100K Domains

(b) Sample of 121K Domains

Figure 5.6: Top 15 domain categories for the `archive.is` live feed.

news sources, social networks, and marketing sites on all social networks. However, there are interesting differences for the two archiving services: Education and Government URLs appear as top categories for the Wayback Machine (see Fig. 5.7(b), 5.7(c), and 5.7(d)), while sites that contain obscene language only for `archive.is` (see Fig. 5.7(c)). This suggests that the latter is used more extensively for "questionable" content.

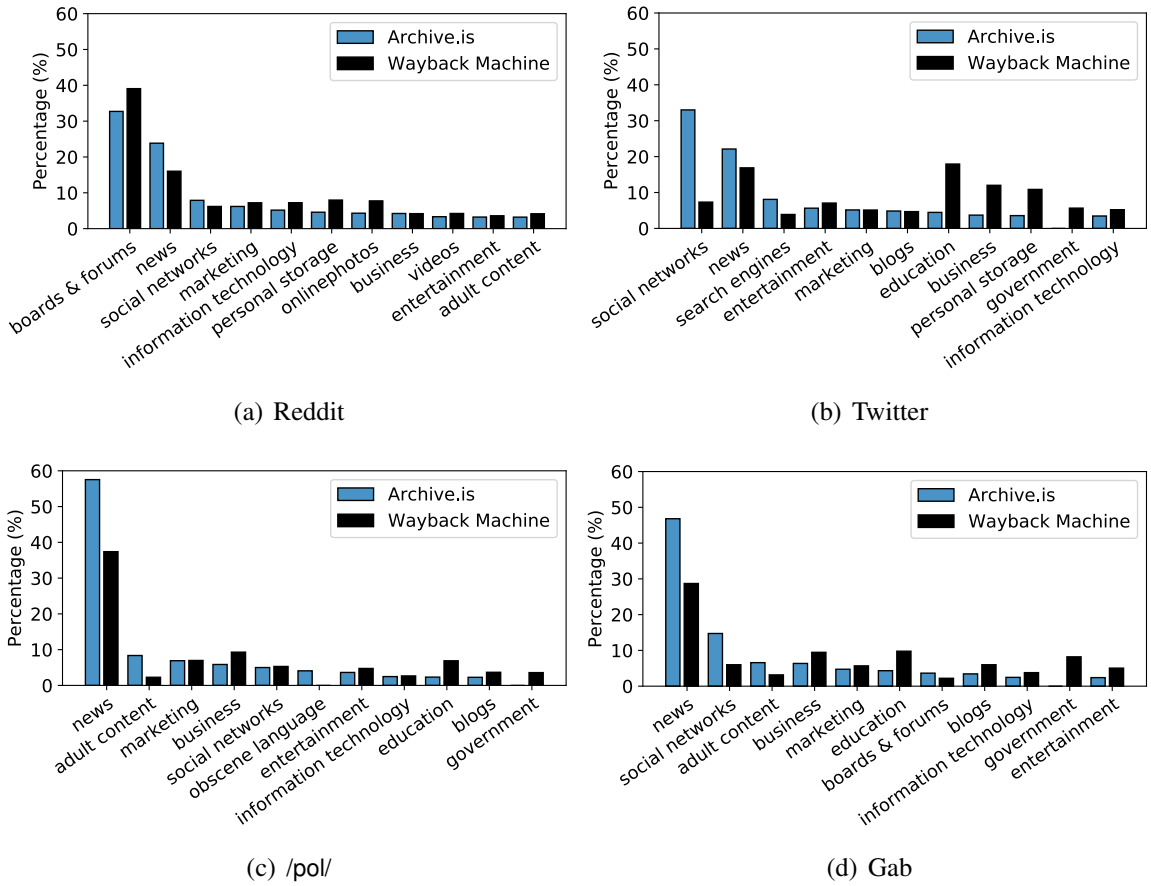
Moreover, we observe that Adult Content is among the top categories for all social networks except Twitter, while Gab and Reddit users often share archive URLs for domains related to Boards and Forums. Also, on /pol/, `archive.is` is used to archive and disseminate pages with obscene language, which is somewhat in line with previous observations [19] showing that /pol/ conversations often include hate speech and aggressive behavior, and so `archive.is` URLs likely point to similar content.

Temporal Dynamics

Next, we study, from a temporal point of view, how archive URLs are created and shared on social networks.

Live Feed. In Fig. 5.8, we plot the day and hour of day of the creation of the `archive.is` URLs. Each day, between 1K and 10K URLs are archived (Fig. 5.8(a)), mostly between 11AM and 4PM UTC time, with a peak at 2PM (Fig. 5.8(b)), which seems to suggest that a great number of users may be located in Europe and the US. According to Alexa, the top country for `archive.is` is the US, with 37% of the visitors.

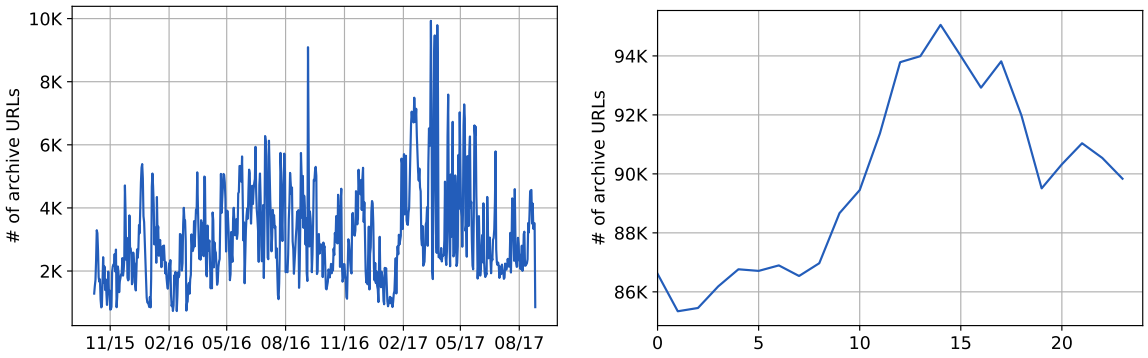
Social Networks. Next, we measure the time interval between the archiving of a URL and its appearance on one of the four social networks. In Fig. 5.9, we plot the CDF of these time



(a) Reddit (b) Twitter

(c) /pol/ (d) Gab

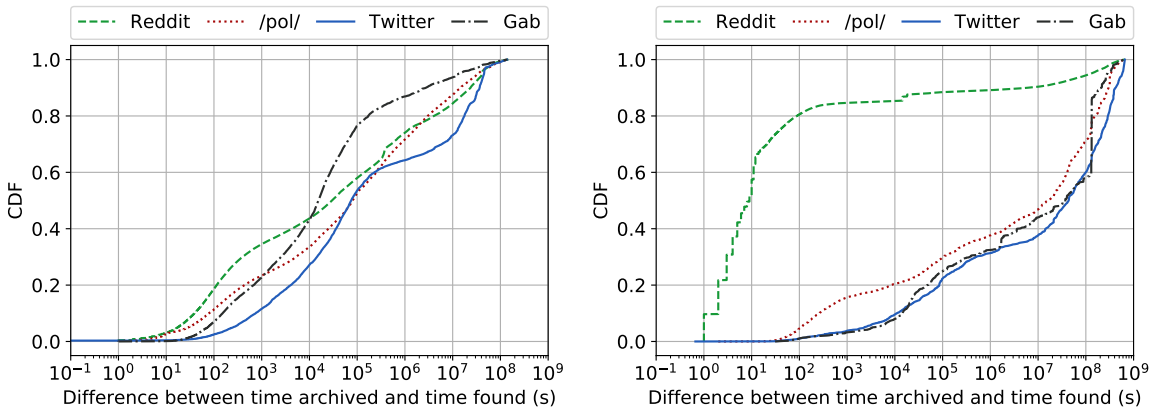
Figure 5.7: Top domain categories for archive URLs appearing on the four social networks.



(a) Date (b) Hour of Day

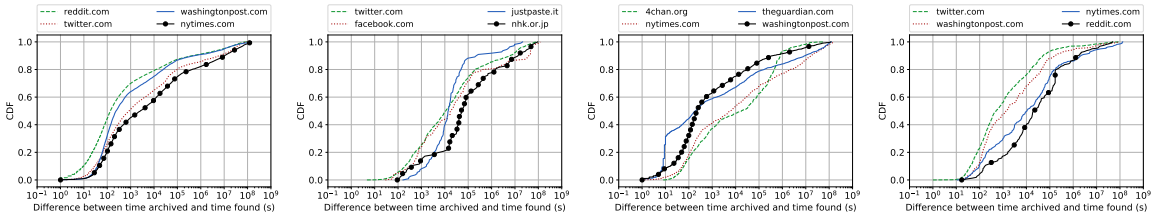
Figure 5.8: Temporal analysis of the `archive.is` live feed dataset, reporting the number of URLs that are archived (a) each day and (b) based on hour of day.

intervals, finding that the interval between archiving and sharing times of a URL ranges from a few seconds (in which case, Reddit/4chan/Twitter/Gab users themselves might be creating the



(a) archive.is (b) Wayback Machine

Figure 5.9: CDF of the time difference between the archival time and the time appeared on each of the four platforms. (Note log scale on x-axis).

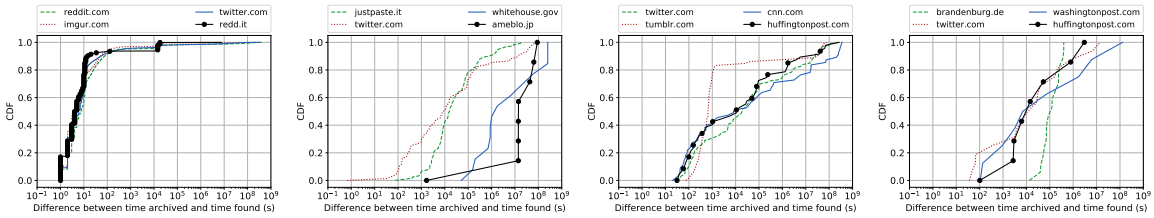


(a) Reddit (b) Twitter (c) /pol/ (d) Gab

Figure 5.10: CDF of the time difference between archival time on `archive.is` and appearance on social networks for the top four source domains.

archive) to years. Reddit is the "fastest" platform for Wayback Machine URLs, mainly because of bots that actively archive URLs (as we show later in this work), while for `archive.is` it is Gab.

We also focus on the top source domains shared via archive URLs: Figs 5.10–5.11 plot the CDF of the slack time of the top four domains for `archive.is` and Wayback Machine URLs, respectively. On Reddit, the top domains archived via Wayback Machine follow very similar distributions, likely due to bots, while for `archive.is` URLs, distributions vary, with the fastest domain being `reddit.com` itself. On Twitter, slack times vary for URLs archived via `archive.is`, with the fastest domain being Twitter and the slowest `nhk.org.jp`. The same applies for the Wayback Machine, with the fastest domain being Twitter and the slowest `ameblo.jp`. We also find that, on /pol/, `archive.is` URLs pointing to 4chan are considerably slower, suggesting that users are more interested in archiving the URL for



(a) Reddit (b) Twitter (c) /pol/ (d) Gab

Figure 5.11: CDF of the time difference between archival time on Wayback Machine and appearance on social networks for top four source domains.

persistence rather than sharing the content within /pol/. Based on anecdotal observations, we believe users might be archiving threads with "evidence" for conspiracy theories/false narratives, and using them in the future to perpetuate mis/disinformation. This is not the case for news sources like the Washington Post or Guardian, as /pol/ users might be more focused on reducing Web traffic to the source domain instead (indeed we find users explicitly mentioning this when manually examining posts). Finally, on Gab, the faster domain is Twitter, and Reddit the slowest.

Original Content Availability

We then assess the availability of the original content that gets archived; this allow us to determine whether users are archiving URLs that are subsequently deleted. To this end, we make an HTTP request for each source URL in our datasets, on October 14–21, 2017 for the live feed dataset, on October 4–5, 2017 for Reddit, Twitter, /pol/ datasets and on January 3, 2018 for Gab dataset. We treat each URL as unavailable if we receive HTTP codes 404/410/451/5xx, or if the request times out.

Live Feed. We find that 12% of the source URLs corresponding to archive URLs on archive.is live feed are no longer available. Domains with most unavailable content include `twitter.com` (6%), `nhk.or.jp` (6%), googleusercontent.com (3%), aaaaarg.fail (3%), 4chan.org (3%), and `8ch.net` (2%).

Social Networks. In Reddit, source URLs corresponding to both `archive.is` and Wayback Machine are still available to a large degree (93% and 89% of them, respectively). This can be explained by the fact that Reddit bots archive URLs without considering the content. In /pol/, the original content is available 82% and 66% of the times, while on Gab 87% and 48% for `archive.is` and Wayback Machine URLs, respectively. Percentages decrease further for

Twitter; 76% and 49% for archive.is and Wayback Machine URLs, respectively.

We also find that the top domains for which content is no longer available differ across platforms. Except for Gab, the top unavailable domain are the social networks themselves: 10%, 54%, and 28%, for Reddit, /pol/, and Twitter, respectively. URLs from cache servers (i.e., googleusercontent.com) and Twitter are also frequently unavailable; 9% and 10% in Reddit, 5% and 4% in /pol/, 8% and 28% in Twitter, and 12% and 19% in Gab, for googleusercontent.com and Twitter, respectively. We also note the presence of unavailable 8ch.net URLs (another ephemeral imageboard) with 5% and 4% on /pol/ and Gab, respectively.

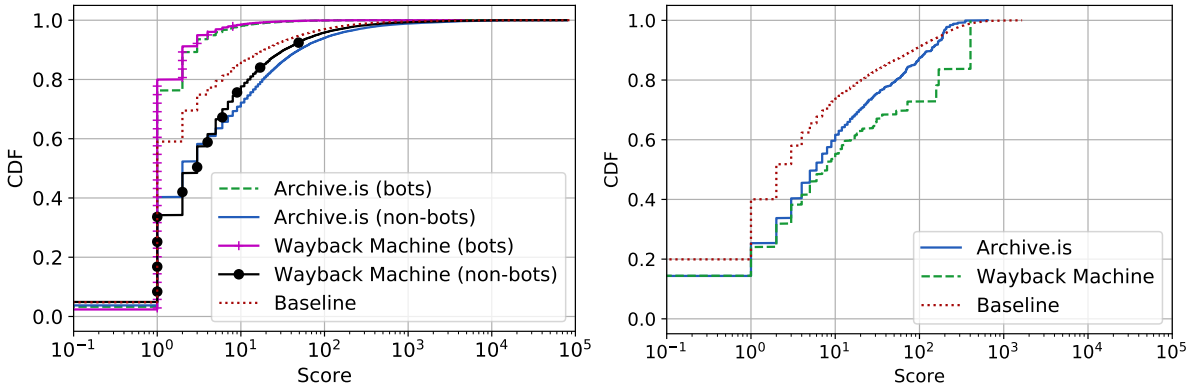
Take-Aways

Overall, we find that archiving services play an important role in the information ecosystem, as they are used to preserve news sources as well as ephemeral or controversial content. Also, users on fringe communities such as /pol/ and Gab favor less popular Web archiving services like archive.is to archive and disseminate Web pages. This prompts questions as to *why* less popular, and seemingly less durable, archiving services are favored by more controversial communities like /pol/ and Gab. Although this would be out of the scope of this work, we do find one potential answer in that these communities also use archiving services to bypass platform-specific censorship policies.

We also observe that temporal dynamics of how archive URLs are shared on social networks differ according to their content: for instance, on /pol/, content from news sources has a considerably larger time lag between first appearing on the platform and archival compared to 4chan threads. Lastly, a non-negligible percentage of archived content is no longer available at the source; in particular, a substantial percentage of posts from social networks like Twitter are eventually deleted from the platform, yet remain stored in the archives.

5.2.5 Social-Network-based Analysis

In this section, we present a social-network-specific analysis by taking into account the fundamental differences of each platform. We analyze the users involved in the dissemination of archive URLs as well as the content that is shared along with those URLs. Lastly, we discuss a case study of ad revenue deprivation on Reddit.



(a) Reddit (b) Gab

Figure 5.12: CDF of the scores of posts that include `archive.is` and Wayback Machine URLs.

User Base

Reddit. Our analysis shows that archiving services are extensively used by Reddit bots. In fact, 31% of all `archive.is` URLs and 82% of Wayback Machine URLs in our Reddit dataset are posted by a specific bot, namely, SnapshillBot (which is used by subreddit moderators to preserve "drama-related" happenings discussed earlier or just as a subreddit specific policy to preserve *every* submission). Other bots include AutoModerator, 2016VoteBot, yankbot, and autotldr. We also attempt to quantify the percentage of archive URLs posted from bots, assuming that, if a username includes "bot" or "auto", it is likely a bot. This is a reasonable strategy since Reddit bots are extensively used for moderation purposes, and do not usually try to obfuscate the fact that they are bots.[3] Using this heuristic, we find that bots are responsible for disseminating 44% of all the `archive.is` and 85% of all the Wayback Machine URLs that appear on Reddit between July 1, 2016 and August 31, 2017.

We also use the score of each Reddit post to get an intuition of users' appreciation for posts that include archive URLs. In Fig. 5.12(a), we plot the CDF of the scores of posts with `archive.is` and Wayback Machine URLs, as well as all posts that contain URLs as a baseline, differentiating between bots and non-bots. For both archiving services, posts by bots have a substantially smaller score: 80% of them have score of at most one, as opposed to 37% for non-bots and 59% of the baseline.

Reddit Sub-Communities. We then study how specific subreddits share URLs from archiv-

[3]This is somewhat evident from the list of Reddit bots available at `https://www.reddit.com/r/autowikibot/wiki/redditbots`

Subreddit (archive.is)	(%)	Subreddit (Wayback)	(%)
The_Donald	24.48%	EnoughTrumpSpam	31.82%
KotakuInAction	15.83%	MGTOW	7.38%
EnoughTrumpSpam	12.06%	SnapshillBotEx	7.19%
MGTOW	3.48%	undelete	5.90%
undelete	2.74%	SubredditDrama	5.50%
SubredditDrama	2.61%	Drama	5.03%
Drama	2.33%	Gamingcirclejerk	3.47%
Gamingcirclejerk	1.57%	ShitAmericansSay	1.63%
conspiracy	1.44%	TopMindsOfReddit	1.51%
MensRights	1.12%	TheBluePill	1.25%
savedyouaclick	1.00%	Buttcoin_1000	1.15%
politics	0.98%	AgainstHateSubreddits	1.06%
DerekSmart	0.76%	subredditcancer	0.99%
ShitAmericansSay	0.75%	The_Donald	0.95%
PoliticsAll	0.72%	badeconomics	0.75%
TopMindsOfReddit	0.71%	ShitWehraboosSay	0.74%
4chan4trump	0.62%	shittykickstarters	0.71%
SnapshillBotEx	0.59%	jesuschristreddit	0.68%
Buttcoin	0.56%	badhistory	0.66%
AgainstHateSubreddits	0.55%	politics	0.59%

Table 5.12: Top 20 subreddits sharing archive.is and Wayback Machine URLs.

ing services. In Table 5.12, we report the top subreddits that share the most archive URLs from archive.is and the Wayback Machine. Among these, we find a variety of subreddits ranging from politics (e.g., EnoughTrumpSpam, The_Donald, politics) to gaming (e.g., Gamingcirclejerk) and "drama-related" communities (e.g., SubredditDrama and Drama). Several subreddits prefer to use archive.is rather than the Wayback Machine, e.g., KotakuInAction, which historically covers the GamerGate controversy [311], The_Donald, which discusses politics with a focus on Donald Trump, and Conspiracy, which focuses on various conspiracy theories.

Gab. On Gab, each post has a score that determines the popularity of the content. In Fig. 5.12(b), we report the CDF of the scores in posts that contain archive.is and Wayback Machine URLs, between August 2016 and August 2017. Once again, we also include a baseline, which is the scores for all the posts with URLs. We find that posts with Wayback Machine URLs have higher scores than those with archive.is URLs, and the baseline. Specifically, the mean score for Wayback Machine is 90, while for archive.is and the baseline the mean score is 35 and 30, respectively. This trend mirrors the one observed on Reddit for posts not authored by bots.

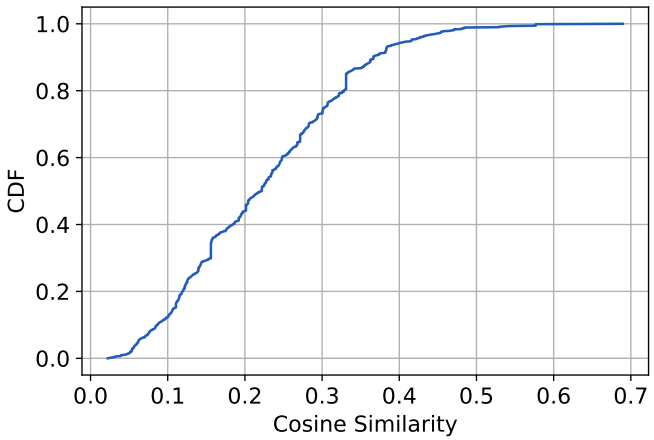


Figure 5.13: CDF of cosine similarity of words obtained from LDA topics on Reddit and dspol threads.

/pol/. As mentioned earlier, 4chan is an anonymous imageboard, which prevents us from performing user-level analysis. However, we can use the flag attribute to provide a country-level estimation. The top country sharing archive URLs is the USA, which is in line with previous characterizations of the board [19]. We also find a substantial percentage of "troll" flags: 9% and 5% for `archive.is` and Wayback Machine, respectively. This is somewhat surprising, since troll flags were re-introduced to /pol/ on June 13, 2017, thus they were only available for about 3 months of our 14-month dataset.

Content Analysis

Next, we focus on the content that gets shared along with archive URLs on social platforms. We aim to evaluate if users share the same information for a given archive URL on multiple platforms. We do so using Latent Dirichlet Allocation (LDA) [312]. Before running LDA, we exclude /pol/ and Reddit threads that contain less than 100 posts, so that the LDA can extract topics from a reasonable amount of documents. We then select only threads that have archive URLs appearing in *both* Reddit and /pol/ datasets; there are 425 such threads on /pol/ and 299 on Reddit. Next, we run LDA on all the posts within these threads and extract terms for 10 topics per thread.

In Fig. 5.13, we plot the CDF of the cosine similarities on the terms extracted from LDA topics on the two platforms when sharing the same archive URLs. We observe that 80% of the terms have similarity under 0.3, which is expected given the fact that the two communities discuss topics in a different way. By manually observing terms with high similarity scores, we find that a number of them relate to well-known conspiracy theories, like the Seth Rich

News Source	Count	(%)	News Source	Count	(%)
washingtonpost.com	3,814	44.13%	change.org	96	7.52%
cnn.com	3,354	39.39%	huffpost.com	62	13.39%
nydailynews.com	1,070	46.32%	fusion.net	60	44.77%
huffingtonpost.com	978	43.77%	cnn.it	58	44.61%
nationalreview.com	774	45.58%	alternet.org	26	20.01%
theblaze.com	704	46.74%	infostormer.com	16	27.11%
buzzfeed.com	588	45.97%	dailynewsbin.com	4	26.67%
salon.com	373	44.88%	todayvibes.com	4	7.27%
vice.com	372	45.14%	usanewsbets.ga	4	10.52%
vox.com	323	45.23%	fullycucked.com	1	1.78%
weeklystandard.com	253	46.25%	northcrane.com	1	0.13%
politifact.com	185	33.09%			

Table 5.13: Number and percentage of submissions deleted from The_Donald with links to different news sources.

murder [37] or Pizzagate [36], as well as general discussions around politics (e.g., tensions between North Korea and the USA). Once again, this highlights that archiving services are used to preserve content related to controversial stories and conspiracy theories.

Ad Revenue Deprivation

During our experiments, we find evidence that at least one Reddit bot, AutoModerator[4], is used to remove links to unwanted domains and nudge users to share `archive.is` instead. In particular, it posts:

> *Your submission was removed because it is from `cnn.com`, which has been identified as a severely anti-Trump domain. Please submit a cached link or screenshot when submitting content from this domain. We recommend using `www.archive.is` for this purpose.*

This kind of notification appears in five different subreddits that discuss mainly politics and news, specifically, The_Donald, Mr_Trump, TheNewRight, Vote_Trump, and Republicans. In particular, in The_Donald, there are 13K such comments. AutoModerator blocks URLs from 23 news sources likely to be considered as anti-Trump by that community. In Table 5.13 we

[4] `https://www.reddit.com/r/AutoModerator/`

Domain	Visits	Loss ($)	Domain	Visits	Loss ($)
washingtonpost.com	79,880	5,928	wsj.com	11,389	845
cnn.com	70,483	5,231	breitbart.com	11,357	842
nytimes.com	46,442	3,446	bbc.com	10,708	794
huffingtonpost.com	27,125	2,013	salon.com	10,364	769
thehill.com	18,643	1,383	buzzfeed.com	10,359	768
theguardian.com	16,376	1,215	foxnews.com	9,638	715
politico.com	15,774	1,170	yahoo.com	9,497	704
dailymail.co.uk	14,442	1,071	latimes.com	9,277	688
dailycaller.com	12,735	945	vox.com	8,976	667
google.com	11,576	859	washingtontimes.com	8,862	657

Table 5.14: Top 20 domains with the largest ad revenue losses because of the use of archiving services on Reddit. We report an estimate of the average monthly visits from Reddit as well as the average monthly ad revenue loss.

report the number of submissions deleted for each of the sources, along with the percentage over *all* submissions that include that source. Mainstream news outlets like Washington Post and CNN are the top domains that get removed from The_Donald (3.8K and 3.3K submissions, respectively), and this happens slightly less than half the times (44% and 39% of the submissions, respectively). Interestingly, only URLs posted via the URL submission field are censored by AutoModerator, but not URLs that are inserted as part of the title field.

We attempt to estimate possible ad revenue deprivation due to the practice of nudging users to share archive URLs instead of source URLs on Reddit. We do so by providing a conservative approximation of the ad revenue loss. Since we do not have knowledge of how many times a particular URL is clicked, we use the up- and down-votes of a post. That is, we assume that when a user up-votes or down-votes a post, he also clicks on the URL included on the post. This constitutes a best-effort technique as prior work shows that a substantial portion of users on Reddit do not vote [313], while, at the same time, users that do vote do not necessarily read or click on the articles [314]. That said, this approach is reasonably conservative considering the complex influence that Reddit has with respect to news dissemination [66].

We then calculate the potential revenue loss using only ad impressions, i.e., we conservatively estimate the revenue generated when a user visits the website without taking into account any potential further action (e.g., clicking on the actual ad). To this end, we use an average Cost per 1,000 impressions (CPM) of $24.74, as reported by Statista[5], while we assume an average of 3 ads per page [315]. In other words, we calculate the monthly revenue loss, for

[5] https://www.statista.com/statistics/308015/online-display-cpm-usa/

each domain, based on the average CPM value as well as the conservative estimate of the visits using the up- and down-votes. Overall, replacing URLs with archive URLs, as done, e.g., by the AutoModerator bot, yields an estimate of $30K per month in revenue loss (for the top 20 domains in terms of views). This is detailed in Table 5.14, where we break down the estimate for each of the top 20 revenue-deprived domains.

On a purely pragmatic level, consider that our estimate of ad revenue deprivation is around $70K per year for the Washington Post alone. Although a more detailed impact analysis is out of the scope of this work, we suspect that even $70K could have a real world effect, e.g., on intern budgets or even early career hires. In light of recent criticism of their credibility by President Donald Trump [316], Trump-supporting communities' deliberate use of `archive.is`, and the conservative nature of our revenue loss estimate, we believe this attack on the Fourth Estate is particularly worrying and in need of future exploration.

Take-Aways

In summary, our social-network-specific analysis shows, among other things, that moderation bots on Reddit proactively leverage Web archiving services to ensure that content shared on their community persists. In particular, we find that 44% and 85% of `archive.is` and Wayback Machine URLs are shared by Reddit moderation bots.

Also, Web archiving services are extensively used for the archival and dissemination of content related to conspiracy theories (e.g., Pizzagate) as well as other world events related to politics (e.g., tensions between North Korea and the USA), thus suggesting that these services play an important role in the (false) information ecosystem and need to be taken into account when designing systems to detect and contain the cascade of mis/disinformation on the Web.

Finally, we find evidence that moderators from specific Reddit sub-communities force users to misuse Web archiving services so as to ideologically target certain news sources by depriving them of traffic and potential ad revenue. We also provide a best-effort conservative estimate of ad revenue loss of popular news sources showing that they can lose up to $70K per year.

5.2.6 Remarks

This work presented a large-scale analysis of how popular Web archiving services such as `archive.is` and the Wayback Machine are used on social media. Our study is based two data crawls: 1) 21M URLs, spanning almost two years, obtained from the `archive.is` live

feed; and 2) 356K `archive.is` plus 391K Wayback Machine URLs that were shared on four social networks: Reddit, Twitter, Gab, and 4chan's Politically Incorrect board (/pol/) over 14 months. Among other things, we showed that these services are extensively used to archive and disseminate news, social network posts, and controversial content—in particular by users of fringe Web communities within Reddit and 4chan. We also found that users not only use them to ensure persistence of Web content, but also to bypass censorship policies enforced on some social networks.

We uncovered evidence that certain subreddits, as well as 4chan's Politically Incorrect board (/pol/), actually nudge users to share archive URLs instead of links to news sources they perceive as having contrasting ideologies, taking away potentially hundreds of thousands of dollars in ad revenue. Overall, our measurements illustrate the importance of archiving services in the Web's information and ad ecosystems, and the need to carefully consider them when studying such ecosystems.

Chapter 6

Towards Understanding State-Sponsored Actors

In this chapter, we study the behavior of state-sponsored actors on the Web. To do this, we leverage ground truth datasets released by Twitter and Reddit pertaining to Russian and Iranian trolls. By analyzing the dataset across several axes we provide a comprehensive analysis on these actors.

Note that the methodology for detecting state-sponsored trolls employed by Twitter and Reddit is not publicly available, therefore, it is unclear on whether there are false positive or how comprehensive these datasets are (i.e., if there are still a lot of unidentified troll accounts). Despite this fact, in this Chapter, we assume that the released datasets are high-quality ground truth datasets with negligible percentage of false positives and adequate coverage of state-sponsored trolls accounts. Therefore, the reader should take into account that all claims and analysis made throughout this Chapter are based on these datasets and it is not clear how these claims and results will change with larger datasets or with datasets from other state-sponsored accounts (e.g., originating from other countries other than Russia and Iran).

6.1 How State-Sponsored Trolls Compare to Random Users and How do Their Accounts Evolve?

6.1.1 Motivation

Recent political events and elections have been increasingly accompanied by reports of disinformation campaigns attributed to state-sponsored actors [317]. In particular, "troll farms," allegedly employed by Russian state agencies, have been actively commenting and posting content on social media to further the Kremlin's political agenda [318]. In late 2017, the US Congress started an investigation on Russian interference in the 2016 US Presidential Election, releasing the IDs of 2.7K Twitter accounts identified as Russian trolls.

Despite the growing relevance of state-sponsored disinformation, the activity of accounts linked to such efforts has not been thoroughly studied. Previous work has mostly looked at campaigns run by bots [317, 319, 183]; however, automated content diffusion is only a part of the issue, and in fact recent research has shown that human actors are actually key in spreading false information on Twitter [207]. Overall, many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? And, perhaps more importantly, how do they compare to a set of random users?

In this work, we aim to address these questions, by relying on the set of 2.7K accounts released by the US Congress as ground truth for Russian state-sponsored trolls. From a dataset containing all tweets released by the 1% Twitter Streaming API, we search and retrieve 27K tweets posted by 1K Russian trolls between January 2016 and September 2017. We characterize their activity by comparing to a random sample of Twitter users.

Main findings. Our study leads to several key observations:

1. The main topics discussed by Russian trolls target very specific world events (e.g., Charlottesville protests) and organizations (such as ISIS), and political threads related to Donald Trump and Hillary Clinton.

2. Trolls adopt different identities over time, i.e., they "reset" their profile by deleting their previous tweets and changing their screen name/information.

3. Trolls exhibit significantly different behaviors compared to other (random) Twitter accounts. For instance, the locations they report concentrate in a few countries like the

USA, Germany, and Russia, perhaps in an attempt to appear "local" and more effectively manipulate opinions of users from those countries. Also, while random Twitter users mainly tweet from mobile versions of the platform, the majority of the Russian trolls do so via the Web Client.

6.1.2 Datasets

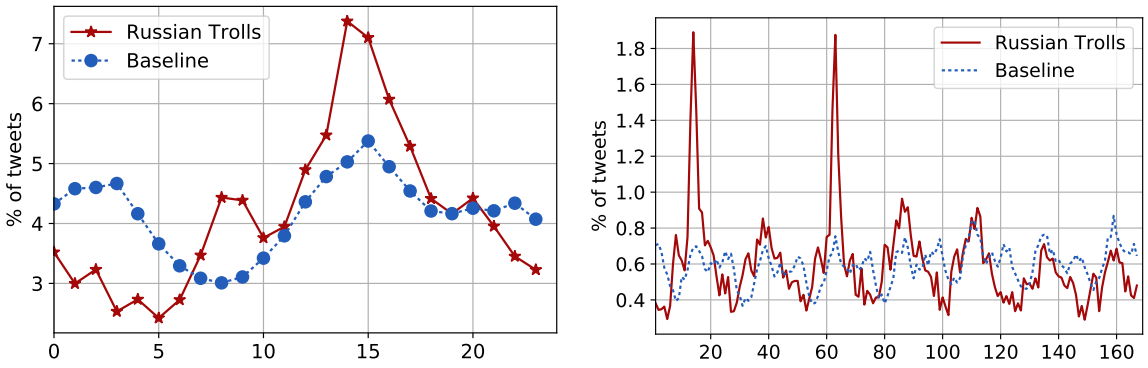
Russian trolls. We start from the 2.7K Twitter accounts suspended by Twitter because of connections to Russia's Internet Research Agency. The list of these accounts was released by the US Congress as part of their investigation of the alleged Russian interference in the 2016 US presidential election, and includes both Twitter's *user ID* (which is a numeric unique identifier associated to the account) and the *screen name*.[1] From a dataset storing all tweets released by the 1% Twitter Streaming API, we search for tweets posted between January 2016 and September 2017 by the user IDs of the trolls. Overall, we obtain 27K tweets from 1K out of the 2.7K Russian trolls.

Baseline dataset. We also compile a list of random Twitter users, while ensuring that the distribution of the average number of tweets per day posted by the random users is similar to the one by trolls. To calculate the average number of tweets posted by an account, we find the first tweet posted after January 1, 2016 and retrieve the overall tweet count. This number is then divided by the number of days since account creation. Having selected a set of 1K random users, we then collect all their tweets between January 2016 and September 2017, obtaining a total of 96K tweets. We follow this approach as it gives a good approximation of posting behavior, even though it might not be perfect, since (1) Twitter accounts can become more or less active over time, and (2) our datasets are based on the 1% Streaming API, thus, we are unable to control the number of tweets we obtain for each account.

6.1.3 Analysis

In this section, we present an in-depth analysis of the activities and the behavior of Russian trolls. First, we provide a general characterization of the accounts and a geographical analysis of the locations they report. Then, we analyze the content they share and how they evolved until their suspension by Twitter. Finally, we present a case study of one specific account.

[1] See `https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf`



(a) Hour of Day (b) Hour of Week

Figure 6.1: Temporal characteristics of tweets from Russian trolls and random Twitter users.

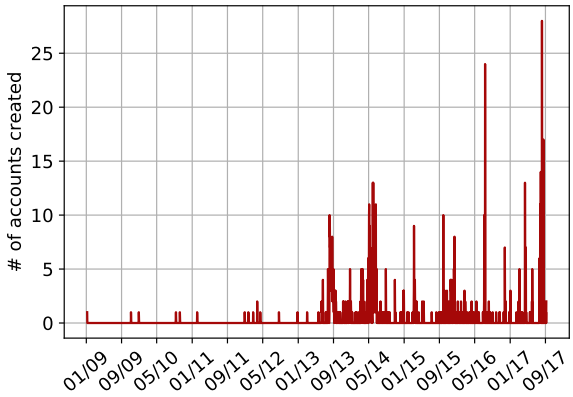


Figure 6.2: Number of Russian troll accounts created per day.

General Characterization

Temporal analysis. We observe that Russian trolls are continuously active on Twitter between January, 2016 and September, 2017, with a peak of activity just before the second US presidential debate (October 9, 2016). Fig. 6.1(a) shows that most tweets from the trolls are posted between 14:00 and 15:00 UTC. In Fig. 6.1(b), we also report temporal characteristics based on hour of the week, finding that both datasets follow a diurnal pattern, while trolls' activity peaks around 14:00 and 15:00 UTC on Mondays and Wednesdays. Considering that Moscow is three hours ahead UTC, this distribution does not rule out that tweets might actually be posted from Russia.

Account creation. Next, we examine the dates when the trolls infiltrated Twitter, by looking at the account creation dates. From Fig. 6.2, we observe that 71% of them are actually

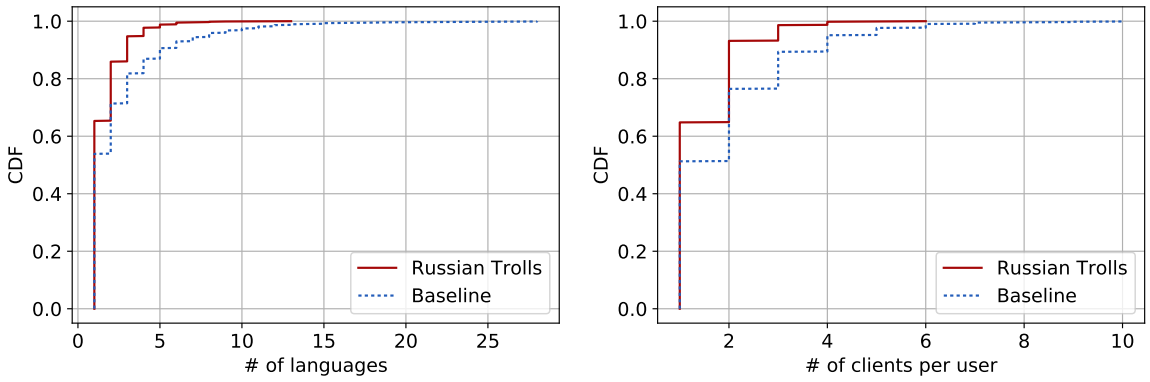
Screen Name				Description			
Word	**(%)**	**4-gram**	**(%)**	**Word**	**(%)**	**Word bigram**	**(%)**
news	1.3%	news	1.5%	news	10.7%	follow me	7.8%
bote	1.2%	line	1.5%	follow	10.7%	breaking news	2.6%
online	1.1%	blac	1.3%	conservative	8.1%	news aus	2.1%
daily	0.8%	bote	1.3%	trump	7.8%	uns in	2.1%
today	0.6%	rist	1.1%	und	6.2%	deiner stdt	2.1%
ezekiel2517	0.6%	nlin	1.1%	maga	5.9%	die news	2.1%
maria	0.5%	onli	1.0%	love	5.8%	wichtige und	2.1%
black	0.5%	lack	1.0%	us	5.3%	nachrichten aus	2.1%
voice	0.4%	bert	1.0%	die	5.0%	aus deiner	2.1%
martin	0.4%	poli	1.0%	nachrichten	4.3%	die dn	2.1%

Table 6.1: Top 10 words found in Russian troll screen names and account descriptions. We also report character 4-grams for the screen names and word bigrams for the description.

created before 2016. There are some interesting peaks, during 2016 and 2017: for instance, 24 accounts are created on July 12, 2016, approx. a week before the Republican National Convention (when Donald Trump received the nomination), while 28 appear on August 8, 2017, a few days before the infamous Unite the Right rally in Charlottesville. Taken together, this might be evidence of coordinated activities aimed at manipulating users' opinions with respect to specific events.

Account characteristics. We also shed light on the troll account profile information. In Table 6.1, we report the top ten words appearing in the screen names and the descriptions of Russian trolls, as well as character 4-grams for screen names and word bigrams for profile descriptions. Interestingly, a substantial number of Russian trolls pose as news outlets, evident from the use of the term "news" in both the screen name (1.3%) and the description (10.7%). Also, it seems they attempt to increase the number of their followers, thus their reach of Twitter users, by nudging users to follow them (see, e.g., "follow me" appearing in almost 8% of the accounts). Finally, 10.3% of the Russian trolls describe themselves as Trump supporters: "trump" and "maga" (Make America Great Again, one of Trump campaign's main slogans).

Language. Looking at the language (as provided via the Twitter API) of the tweets posted by Russian trolls, we find that most of them (61%) are in English, although a substantial portion are in Russian (27%), and to a lesser extent in German (3.5%). In Fig. 6.3(a), we plot the cumulative distribution function (CDF) of the number of different languages for each user:



(a) (b)

Figure 6.3: CDF of number of (a) languages used (b) clients used per user.

64% of the Russian trolls post all their tweets in only one language, compared to only 54% for random users. Overall, by comparing the two distributions, we observe that random Twitter users tend to use more languages in their tweets compared to the trolls.

Client. Finally, we analyze the clients used to post tweets. We do so since previous work [320] shows that the client used by official or professional accounts are quite different that the ones used by regular users. Table 6.2 reports the top 10 clients for both Russian trolls and baseline users. We find the latter prefer to use Twitter clients for mobile devices (48%) and the TweetDeck dashboard (32%), whereas, the former mainly use the Web client (50%). We also assess how many different clients Russian trolls use throughout our dataset: in Fig. 6.3(b), we plot the CDF of the number of clients used per user. We find that 65% of the Russian trolls use only one client, 28% of them two different clients, and the rest more than three, which is overall less than the random baseline users.

Geographical Analysis

Location. We then study users' location, relying on the self-reported location field in their profiles. Note that users not only may leave it empty, but also change it any time they like, so we look at locations for each tweet. We retrieve it for 75% of the tweets by Russian trolls, gathering 261 different entries, which we convert to a physical location using the Google Maps Geocoding API. In the end, we obtain 178 unique locations for the trolls, as depicted in Fig. 6.4 (red circles). The size of the circles on the map indicates the number of tweets that appear at each location. We do the same for the baseline, getting 2,037 different entries, converted

Client (Trolls)	(%)	Client (Baseline)	(%)
Twitter Web Client	50.1%	TweetDeck	32.6%
twitterfeed	13.4%	Twitter for iPhone	26.2%
Twibble.io	9.0%	Twitter for Android	22.6%
IFTTT	8.6%	Twitter Web Client	6.1%
TweetDeck	8.3%	GrabInbox	2.0%
NovaPress	4.6%	Twitter for iPad	1.4%
dlvr.it	2.3%	IFTTT	1.0%
Twitter for iPhone	0.8%	twittbot.net	0.9%
Zapier.com	0.6%	Twitter for BlackBerry	0.6%
Twitter for Android	0.6%	Mobile Web (M2)	0.4%

Table 6.2: Top 10 Twitter clients (as % of tweets).

by the API to 894 unique locations. We observe that most of the tweets from Russian trolls come from locations within the USA and Russia, and some from European countries, like Germany, Belgium, and Italy. On the other hand, tweets in our baseline are more uniformly distributed across the globe, with many tweets from North and South America, Europe, and Asia. This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or Germany, aiming to pose as locals and better manipulate opinions. This explanation becomes more plausible when we consider that a plurality of trolls' tweets have their location set as a generic form of "US," as opposed to a specific city, state, or even region. Interestingly, the 2nd, 3rd, and 4th most popular location for trolls to tweet from are Moscow, St. Petersburg, and a generic form of "Russia." We also assess whether users change their country of origin based on the self-reported location: only a negligible percentage (1%) of trolls change their country, while for the baseline the percentage is 16%.

Timezone. We then study the timezone chosen by the users in their account setting. In Table 6.3, we report the top 10 timezones for each dataset, in terms of the corresponding tweet volumes. Two thirds of the tweets by trolls appear to be from US timezones, while a substantial percentage (18%) from Russian ones. Whereas, the baseline has a more diverse set of timezones, which seems to mirror findings from our location analysis.

We also check whether users change their timezone settings, finding that 7% of the Russian trolls do so two to three times. The most popular changes are Berlin to Bern (18 times), Nairobi to Moscow (10), and Nairobi to Volgograd (10).By contrast, this almost never happens

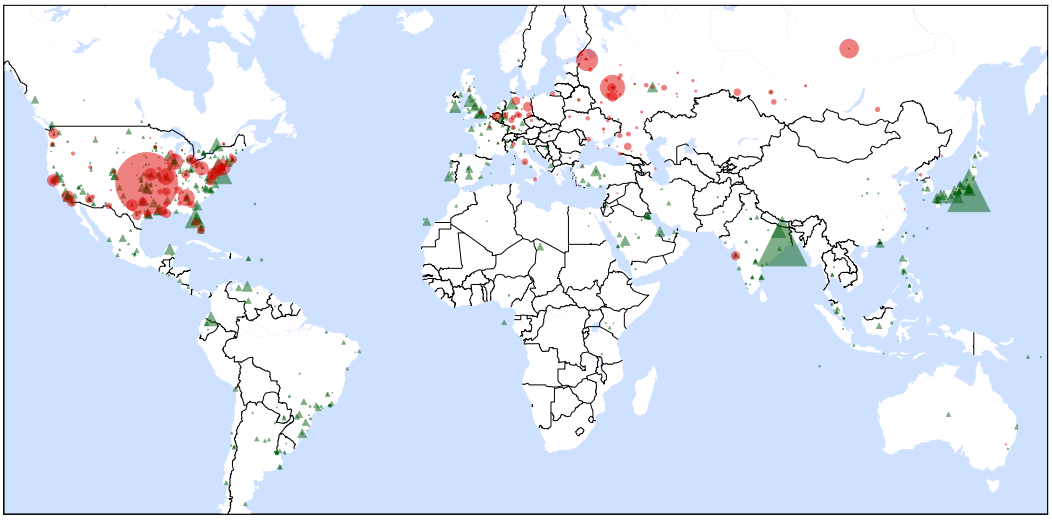


Figure 6.4: Distribution of reported locations for tweets by Russian trolls (red circles) and baseline (green triangles).

for baseline accounts.

Content Analysis

Text. Next, we quantify the number of characters and words contained in each tweet, and plot the corresponding CDF in Fig. 6.5, finding that Russian trolls tend to post longer tweets.

Media. We then assess whether Russian trolls use images and videos in a different way than random baseline users. For Russian trolls (resp., baseline accounts), 66% (resp., 73%) of the tweets include no images, 32% (resp., 18%) exactly one image, and 2% (resp., 9%) more than one. This suggests that Russian trolls disseminate a considerable amount of information via single-image tweets. As for videos, only 1.5% of the tweets from Russian trolls includes a video, as opposed to 6.4% for baseline accounts.

Hashtags. Our next step is to study the use of hashtags in tweets. Russian trolls use at least one hashtag in 32% of their tweets, compared to 10% for the baseline. Overall, we find 4.3K and 7.1K unique hashtags for trolls and random users, respectively, with 74% and 78% of them only appearing once. In Table 6.4, we report the top 20 hashtags for both datasets. Trolls appear to use hashtags to disseminate news (7.2%) and politics (2.6%) related content, but also use several that might be indicators of propaganda and/or controversial topics, e.g., #ISIS, #IslamKills, and #BlackLivesMatter. For instance, we find some notable examples including: "We just have to close the borders, 'refugees' are simple terrorists #IslamKills" on March 22,

Timezone (Trolls)	(%)	Timezone (Baseline)	(%)
Eastern Time	38.87%	Athens	24.41%
Pacific Time	18.37%	Pacific Time	21.41%
Volgograd	10.03%	London	21.27%
Central Time	9.43%	Tokyo	3.83%
Moscow	8.18%	Central Time	3.75%
Bern	2.56%	Eastern Time	2.10%
Minsk	2.06%	Seoul	1.97%
Yerevan	1.96%	Brasilia	1.97%
Nairobi	1.52%	Buenos Aires	1.92%
Baku	1.29%	Urumqi	1.50%

Table 6.3: Top 10 timezones (as % of tweets).

2016, "#SyrianRefugees ARE TERRORISTS from #ISIS #IslamKills" on March 22, 2016, and "WATCH: Here is a typical #BlackLivesMatter protester: 'I hope I kill all white babes!' #BatonRouge <url>" on July 17, 2016.

We also study when these hashtags are used by the trolls, finding that most of them are well distributed over time. However, there are some interesting exceptions, e.g., with #Merkelmussbleiben (a hashtag seemingly supporting Angela Merkel) and #IslamKills. Specifically, tweets with the former appear exclusively on July 21, 2016, while the latter on March 22, 2016, when a terrorist attack took place at Brussels airport. These two examples illustrate how the trolls may be coordinating to push specific narratives on Twitter.

Mentions. We find that 46% of trolls' tweets include *mentions* to 8.5K unique Twitter users. This percentage is much higher for the random baseline users (80%, to 41K users). Table 6.5 reports the 20 top mentions we find in tweets from Russian trolls and baseline users. We find several Russian accounts, like 'leprasorium' (a popular Russian account that mainly posts memes) in 2% of the mentions, as well as popular politicians like 'realDonaldTrump' (0.6%). The practice of mentioning politicians on Twitter may reflect an underlying strategy to mutate users' opinions regarding a particular political topic, which has been also studied in previous work [184].

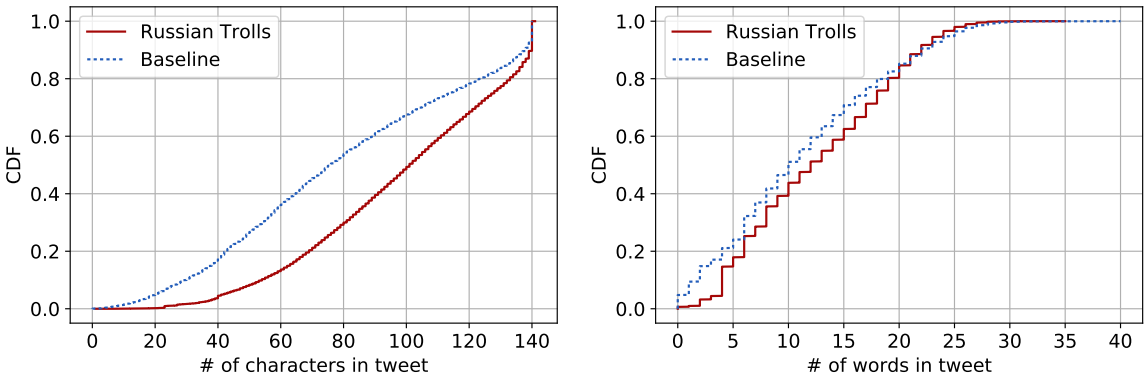
URLs. We then analyze the URLs included in the tweets. First of all, we note that 53% of the trolls' tweets include at least a URL, compared to only 27% for the random baseline. There is an extensive presence of URL shorteners for both datasets, e.g., `bit.ly` (12% for trolls

Trolls				Baseline			
Hashtag	**(%)**	**Hashtag**	**(%)**	**Hashtag**	**(%)**	**Hashtag**	**(%)**
news	7.2%	US	0.7%	iHeartAwards	1.8%	UrbanAttires	0.6%
politics	2.6%	tcot	0.6%	BestFanArmy	1.6%	Vacature	0.6%
sports	2.1%	PJNET	0.6%	Harmonizers	1.0%	mPlusPlaces	0.6%
business	1.4%	entertainment	0.5%	iOSApp	0.9%	job	0.5%
money	1.3%	top	0.5%	JouwBaan	0.9%	Directioners	0.5%
world	1.2%	topNews	0.5%	vacature	0.9%	JIMIN	0.5%
MAGA	0.8%	ISIS	0.4%	KCA	0.9%	PRODUCE101	0.5%
health	0.8%	Merkelmussbleiben	0.4%	Psychic	0.8%	VoteMainFPP	0.5%
local	0.7%	IslamKills	0.4%	RT	0.8%	Werk	0.4%
BlackLivesMatter	0.7%	breaking	0.4%	Libertad2016	0.6%	dts	0.4%

Table 6.4: Top 20 hashtags in tweets from Russian trolls and baseline users.

Trolls				Baseline			
Mention	**(%)**	**Mention**	**(%)**	**Mention**	**(%)**	**Mention**	**(%)**
leprasorium	2.1%	postsovet	0.4%	TasbihIstighfar	0.3%	RasSpotlights	0.1%
zubovnik	0.8%	DLGreez	0.4%	raspotlights	0.2%	GenderReveals	0.1%
realDonaldTrump	0.6%	DanaGeezus	0.4%	FunnyBrawls	0.2%	TattedChanel	0.1%
midnight	0.6%	ruopentwit	0.3%	YouTube	0.2%	gemvius	0.1%
blicqer	0.6%	Spoontamer	0.3%	Harry_Styles	0.2%	DrizzyNYC__	0.1%
gloed_up	0.6%	YouTube	0.3%	shortdancevids	0.2%	August_Alsina_	0.1%
wylsacom	0.5%	ChrixMorgan	0.3%	UrbanAttires	0.2%	RihannaBITCH_	0.1%
TalibKweli	0.4%	sergeylazarev	0.3%	BTS_twt	0.2%	sexualfeed	0.1%
zvezdanews	0.4%	RT_com	0.3%	KylieJenner_NYC	0.2%	PetsEvery30	0.1%
GiselleEvns	0.4%	kozheed	0.3%	BaddiessNation	0.2%	IGGYAZALEAoO	0.1%

Table 6.5: Top 20 mentions in tweets from trolls and baseline.



(a) (b)

Figure 6.5: CDF of the number of (a) characters and (b) words in each tweet.

and 26% for the baseline) and `ift.tt` (10% for trolls and 2% for the baseline), therefore, in November 2017, we visit each URL to obtain the final URL after all redirections. In Fig. 6.6, we plot the CDF of the number of URLs per unique domain. We observe that Russian trolls disseminate more URLs in their tweets compared to the baseline. This might indicate that Russian trolls include URLs to increase their credibility and positive user perception; indeed, [120] show that adding a URL in a tweet correlates with higher credibility scores. Also, in Table 6.6, we report the top 20 domains for both Russian trolls and the baseline. Most URLs point to content within Twitter itself; 13% and 35%, respectively. Links to a number of popular social networks like YouTube (1.8% and 4.2%, respectively) and Instagram (1.5% and 1.9%) appear in both datasets. We also note that among the top 20 domains, there are also a number of news outlets linked from trolls' tweets, e.g., Washington Post (0.7%), Seattle Times (0.7%), and state-sponsored news outlets like RT (0.8%) in trolls' tweets, but much less so from the baseline.

Sentiment analysis. Next, we assess the sentiment and subjectivity of each tweet for both datasets using the Pattern library [321]. Fig. 6.7(a) plots the CDF of the sentiment scores of tweets posted by Russian trolls and our baseline users. We observe that 30% of the tweets from Russian trolls have a positive sentiment, and 18% negative. These scores are not too distant from those of random users where 36% are positive and 16% negative, however, Russian trolls exhibit a unique behavior in terms of sentiment, as a two-sample Kolmogorov-Smirnov test unveils significant differences between the distributions ($p < 0.01$). Overall, we observe that Russian trolls tend to be more negative/neutral, while our baseline is more positive. We also compare subjectivity scores (Fig. 6.7(b)), finding that tweets from trolls tend to be

Domain (Trolls)	(%)	Domain (Baseline)	(%)
twitter.com	12.81%	twitter.com	35.51%
reportsecret.com	7.02%	youtube.com	4.21%
riafan.ru	3.42%	vine.co	3.94%
politexpert.net	2.10%	factissues.com	3.24%
youtube.com	1.88%	blogspot.com.cy	1.92%
vk.com	1.58%	instagram.com	1.90%
instagram.com	1.53%	facebook.com	1.68%
yandex.ru	1.50%	worldstarr.info	1.47%
infreactor.org	1.36%	trendytopic.info	1.39%
cbslocal.com	1.35%	minibird.jp	1.25%
livejournal	1.35%	yaadlinksradio.com	1.24%
nevnov.ru	1.07%	soundcloud.com	1.24%
ksnt.com	1.01%	linklist.me	1.15%
kron4.com	0.93%	twimg.com	1.09%
viid.me	0.93%	appparse.com	1.08%
newinform.com	0.89%	cargobayy.net	0.88%
inforeactor.ru	0.84%	virralclub.com	0.84%
rt.com	0.81%	tistory.com	0.50%
washigntonpost.com	0.75%	twitcasting.tv	0.49%
seattletimes.com	0.73%	nytimes.com	0.48%

Table 6.6: Top 20 domains included in tweets from Russian trolls and baselines users.

more subjective; again, we perform significance tests revealing differences between the two distributions ($p < 0.01$).

LDA analysis. We also use the Latent Dirichlet Allocation (LDA) model to analyze tweets' semantics. We train an LDA model for each of the datasets and extract 10 distinct topics with 10 words, as reported in Table 6.7. Overall, topics from Russian trolls refer to specific world events (e.g., Charlottesville) as well as specific news related to politics (e.g., North Korea and Donald Trump). By contrast, topics extracted from the random sample are more general (e.g., tweets regarding birthdays).

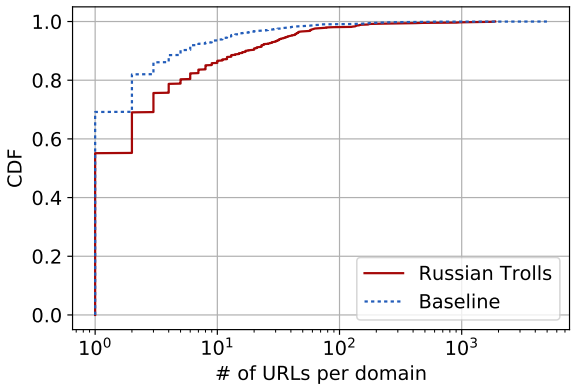
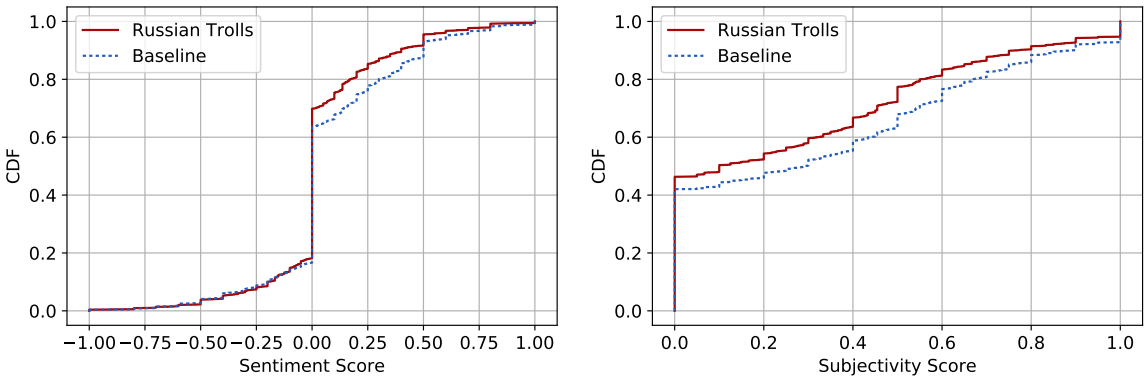


Figure 6.6: CDF of number of URLs per domain.



(a) (b)

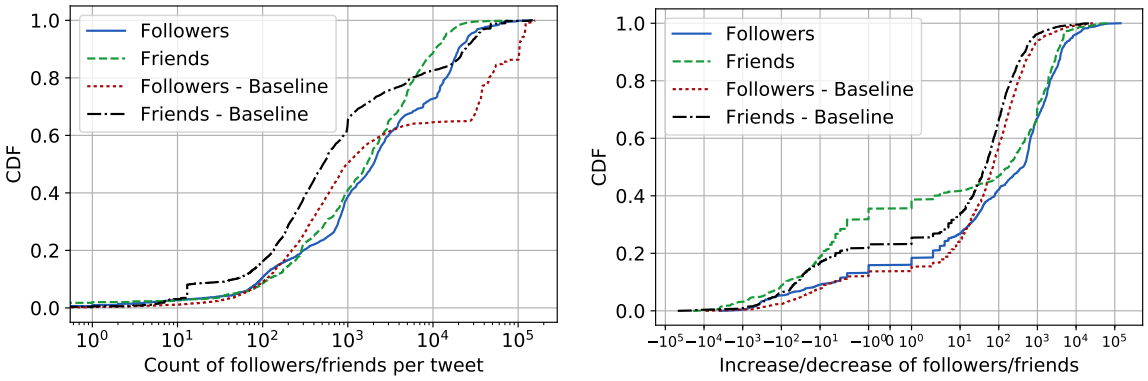
Figure 6.7: CDF of sentiment and subjectivity scores for tweets of Russian trolls and random users.

Account Evolution

Screen name changes. Previous work [322] has shown that malicious accounts often change their screen name in order to assume different identifies. Therefore, we investigate whether trolls show a similar behavior, as they might change the narrative with which they are attempting to influence public opinion. Indeed, we find that 9% of the accounts operated by trolls change their screen name, up to 4 times during the course of our dataset. Some examples include changing screen names from "OnlineHouston" to "HoustonTopNews," or "Jesus Quintin Perez" to "WorldNewsPolitics," in a clear attempt to pose as news-related accounts. In our baseline, we find that 19% of the accounts changed their Twitter screen names, up to 11 times during our dataset; highlighting that changing screen names is a common behavior of Twitter users in general.

Topic	Terms (Trolls)	Topic	Terms (Baseline)
1	trump, black, people, really, one, enlist, truth, work, can, get	1	want, can, just, follow, now, get, see, don, love, will
2	trump, year, old, just, run, obama, breaking, will, news, police	2	2016, july, come, https, trump, social, just, media, jabberduck, get
3	new, trump, just, breaking, obamacare, one, sessions, senate, politics, york	3	happy, best, make, birthday, video, days, come, back, still, little
4	man, police, news, killed, shot, shooting, woman, dead, breaking, death	4	know, never, get, love, just, night, one, give, time, can
5	trump, media, tcot, just, pjnet, war, like, video, post, hillary	5	just, can, everyone, think, get, white, fifth, veranomtv2016, harmony, friends
6	sports, video, game, music, isis, charlottesville, will, new, health, amb	6	good, like, people, lol, don, just, look, today, said, keep
7	can, don, people, want, know, see, black, get, just, like	7	summer, seconds, team, people, miss, don, will, photo, veranomtv2016, new
8	trump, clinton, politics, hillary, video, white, donald, president, house, calls	8	like, twitter, https, first, can, get, music, better, wait, really
9	news, world, money, business, new, one, says, state, 2016, peace	9	dallas, right, fuck, vote, police, via, just, killed, teenchoice, aldubmainecelebration
10	now, trump, north, korea, people, right, will, check, just, playing	10	day, black, love, thank, great, new, now, matter, can, much

Table 6.7: Terms extracted from LDA topics of tweets from Russian trolls and baseline users.



(a) (b)

Figure 6.8: CDF of the number of (a) followers/friends for each tweet and (b) increase in followers/friends for each user from the first to the last tweet.

Followers/Friends. Next, we look at the number of followers and friends (i.e., the accounts one follows) of the Russian trolls, as this is an indication of the overall impact of a tweet. In Fig. 6.8(a), we plot the CDF of the number of followers per tweet measured at the time of that tweet. On average, Russian trolls have 7K followers and 3K friends, while our baseline has 25K followers and 6K friends. We also note that in both samples, tweets reached a large number of Twitter users; at least 1K followers, with peaks up to 145K followers. These results highlight that Russian trolls have a non-negligible number of followers, which can assist in pushing specific narratives to a much greater number of Twitter users. We also assess the evolution of the Russian trolls in terms of the number of their followers and friends. To this end, we get the follower and friend count for each user on their first and last tweet and calculate the difference. Fig. 6.8(b) plots the CDF of the increase/decrease of the followers and friends for each troll as well as random user in our baseline. We observe that, on average, Russian trolls increase their number of followers and friends by 2,065 and 1,225, respectively, whereas

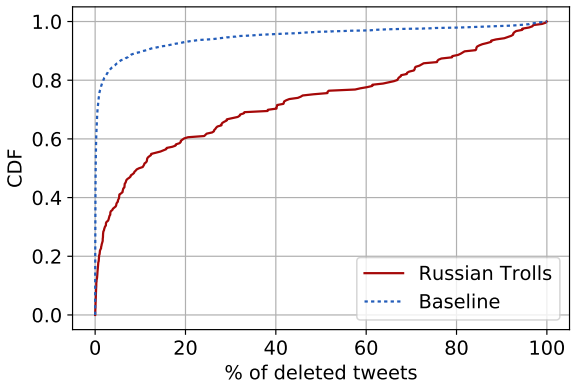


Figure 6.9: CDF of the number of deleted tweets per observe deletion.

for the baseline we observe an increase of 425 and 133 for followers and friends, respectively. This suggests that Russian trolls work hard to increase their reachability within Twitter.

Tweet Deletion. Arguably, a reasonable strategy to avoid detection after posting tweets that aim to manipulate other users might be to delete them. This is particularly useful when troll accounts change their identity and need to modify the narrative that they use to influence public opinion. With each tweet, the Streaming API returns the total number of available tweets a user has up to that time. Retrieving this count allows us to observe if a user has deleted a tweet, and around what period; we call this an "observed deletion." Recall that our dataset is based on the 1% sample of Twitter, thus, we can only estimate, in a conservative way, how many tweets are deleted; specifically, in between subsequent tweets, a user may have deleted and posted tweets that we do not observe. In Fig. 6.9, we plot the CDF of the number of deleted tweets per observed deletion. We observe that 13% of the Russian trolls delete some of their tweets, with a median percentage of tweet deletion equal to 9.7%. Whereas, for the baseline set, 27% of the accounts delete at least one tweet, but the median percentage is 0.1%. This means that the trolls delete their tweets in batches, possibly trying to cover their tracks or get a clean slate, while random users make a larger number of deletions but only a small percentage of their overall tweets, possibly because of typos. We also report the distribution, over each month, of tweet deletions in Fig. 6.10. Specifically, we report the mean of the percentages for all observed deletions in our datasets. Most of the tweets from Russian trolls are deleted in October 2016, suggesting that these accounts attempted to get a clean slate a few months before the 2016 US elections.

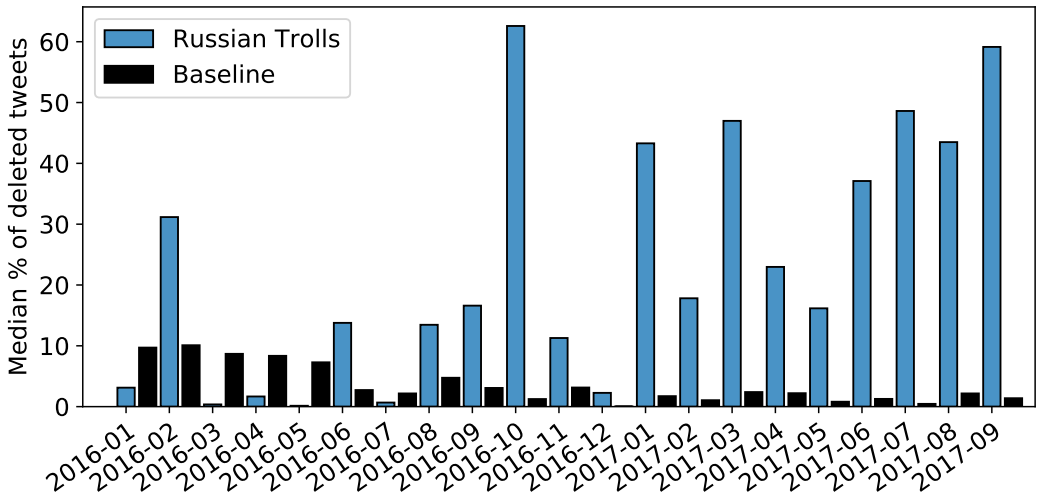


Figure 6.10: Average percentage of observed deletions per month.

Case Study

While the previous results provide a quantitative characterization of Russian trolls behavior, we believe there is value showing a concrete example of the behavior exhibited and how techniques played out. We start on May 15, 2016, where the troll with screen name 'Pen_Air', was posing as a news account via its profile description: "National American news." On September 8, 2016 as the US presidential elections approached, 'Pen_Air' became a Trump supporter, changing its screen name to 'Blacks4DTrump' with a profile description of "African-Americans stand with Trump to make America Great Again!" Over the next 11 months, the account's tweet count grew from 49 to 642 while its follower count rose from 1.2K to *nearly* 9K. Then, around August 18, 2017, the account was seemingly repurposed. Almost all of its previous tweets were deleted (the account's tweet count dropped to 35), it gained a new screen name ('southlonestar2'), and was now posing as a "Proud American and TEXAN patriot! Stop ISLAM and PC. Don't mess with Texas" according to its profile description. When examining the accounts tweets, we see that most are clearly related to politics, featuring blunt right-wing attacks and "talking points." For example, "Mr. Obama! Maybe you bring your girls and leave them in the bathroom with a grown man! #bathroombill #NObama <url>" on May 15, 2016, "#HiLIARy has only two faces! And I hate both! #NeverHillary #Hillaryliesmatter <url>" on May 19, 2016, and "RT @TEN_GOP: WikiLeaks #DNCLeaks confirms something we all know: system is totally rigged! #NeverHillary <url>." on July 22, 2016.

Take-aways

In summary, our analysis leads to the following observations. First, we find evidence that trolls were actively involved in the dissemination of content related to world news and politics, as well as propaganda content regarding various topics such as ISIS and Islam. Moreover, several Russian trolls were created or repurposed a few weeks before notable world events, including the Republican National Convention meeting or the Charlottesville rally. We also find that the trolls deleted a substantial amount of tweets in batches and overall made substantial changes to their accounts during the course of their lifespan. Specifically, they changed their screen names aiming to pose as news outlets, experienced significant rises in the numbers of followers and friends, etc. Furthermore, our location analysis shows that Russian trolls might have tried to manipulate users located in the USA, Germany, and possibly in their own country (i.e., Russia), by appearing to be located in those countries. Finally, the fact that these accounts were active up until their recent suspension also highlights the need to develop more effective tools to detect such actors.

6.1.4 Remarks

In this work, we analyzed the behavior and use of the Twitter platform by Russian trolls during the course of 21 months. We showed that Russian trolls exhibited interesting differences when compared with a set of random users, actively disseminated politics-related content, adopted multiple identities during their account's lifespan, and that they aimed to increase their impact on Twitter by increasing their followers.

6.2 A comprehensive analysis of Russian and Iranian trolls on Twitter and Reddit and their influence on the Web

6.2.1 Motivation

In this work, we are motivated by the fact that many aspects of state-sponsored disinformation remain unclear, e.g., how do state-sponsored trolls operate? What kind of content do they disseminate? How does their behavior change over time? And, more importantly, is it possible to quantify the influence they have on the overall information ecosystem on the Web?

Here, we aim to address these questions, by relying on two different sources of ground truth

data about state-sponsored actors. First, we use 10M tweets posted by Russian and Iranian trolls between 2012 and 2018 [323]. Second, we use a list of 944 Russian trolls, identified by Reddit, and find all their posts between 2015 and 2018 [324]. We analyze the two datasets across several axes in order to understand their behavior and how it changes over time, their targets, and the content they shared. For the latter, we leverage word embeddings to understand in what context specific words/hashtags are used and shed light to the ideology of the trolls. Also, we use Hawkes Processes [27] to model the influence that the Russian and Iranian trolls had over multiple Web communities; namely, Twitter, Reddit, 4chan's Politically Incorrect board (/pol/) [19], and Gab [43].

Main findings. Our study leads to several key observations:

1. Our influence estimation results reveal that Russian trolls were extremely influential and efficient in spreading URLs on Twitter. Also, by comparing Russian to Iranian trolls, we find that Russian trolls were more efficient and influential in spreading URLs on Twitter, Reddit, Gab, but not on /pol/.

2. By leveraging word embeddings, we find ideological differences between Russian and Iranian trolls. For instance, we find that Russian trolls were pro-Trump, while Iranian trolls were anti-Trump.

3. We find evidence that the Iranian campaigns were motivated by real-world events. Specifically, campaigns against France and Saudi Arabia coincided with real-world events that affect the relations between these countries and Iran.

4. We observe that the behavior of trolls varies over time. We find substantial changes in the use of language and Twitter clients over time for both Russian and Iranian trolls. These insights allow us to understand the targets of the orchestrated campaigns for each type of trolls over time.

5. We find that the topics of discussion vary across Web communities. For example, we find that Russian trolls on Reddit were extensively discussing about cryptocurrencies, while this does not apply in great extent for the Russian trolls on Twitter.

Finally, we make our source code publicly available [325] for reproducibility purposes and to encourage researchers to further work on understanding other types of state-sponsored trolls on Twitter (i.e., on January 31, 2019, Twitter released data related to trolls originating from Venezuela and Bangladesh [326]).

Platform	Origin of trolls	# trolls	# trolls with tweets/posts	# of tweets/posts
Twitter	**Russia**	3,836	3,667	9,041,308
	Iran	770	660	1,122,936
Reddit	**Russia**	944	335	21,321

Table 6.8: Overview of Russian and Iranian trolls on Twitter and Reddit. We report the overall number of identified trolls, the trolls that had at least one tweet/post, and the overall number of tweets/posts.

6.2.2 Troll Datasets

In this section, we describe our dataset of Russian and Iranian trolls on Twitter and Reddit.

Twitter. On October 17, 2018, Twitter released a large dataset of Russian and Iranian troll accounts [323]. Although the exact methodology used to determine that these accounts were state-sponsored trolls is unknown, based on the most recent Department of Justice indictment [327], the dataset appears to have been constructed in a manner that we can assume essentially no false positives, while we cannot make any postulation about false negatives. Table 6.8 summarizes the troll dataset.

Reddit. On April 10, 2018, Reddit released a list of 944 accounts which they determined were Russian state-sponsored trolls [324]. We recover the submissions, comments, and account details for these accounts using two mechanisms: 1) Reddit dumps provided by Pushshift [236]; and 2) crawling the user pages of those accounts. Although omitted for lack of space, we note that the union of these two datasets reveals some gaps in both, likely due to a combination of subreddit moderators removing posts or the troll users themselves deleting them, which would affect the two datasets in different ways. In any case, we merge the two datasets, with Table 6.8 describing the final dataset. Note that only about one third (335) of the accounts released by Reddit had at least one submission or comment in our dataset. We suspect the rest were either completely missed by our data sources, or, more likely, were used as dedicated voting accounts used in an effort to push (or bury) specific content.

6.2.3 Analysis

In this section, we present an in-depth analysis of the activities and the behavior of Russian and Iranian trolls on Twitter and Reddit.

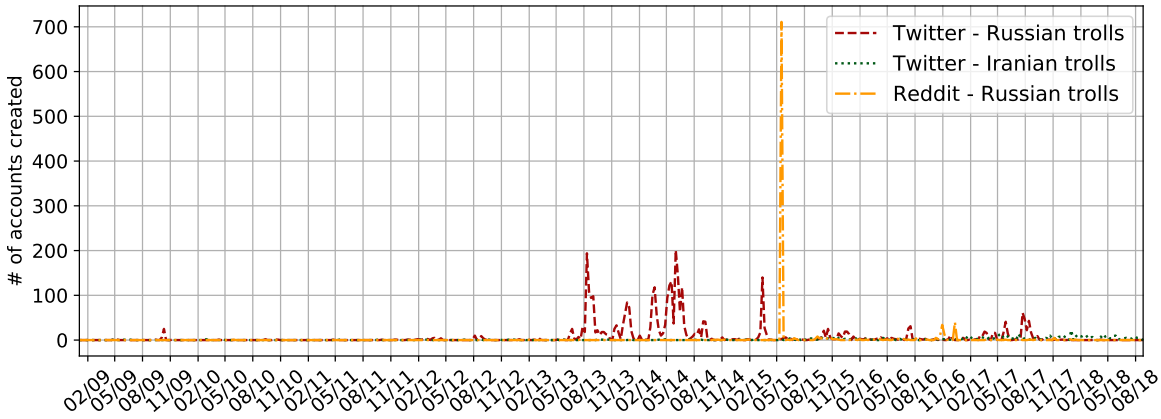


Figure 6.11: Number of Russian and Iranian troll accounts created per week.

Russian troll on Twitter				Iranian trolls on Twitter			
Word	**(%)**	**Bigrams**	**(%)**	**Word**	**(%)**	**Bigrams**	**(%)**
follow	7.7%	follow me	6.4%	journalist	3.6%	human rights	1.6%
love	4.8%	breaking news	0.8%	news	3.2%	independent news	1.4%
life	4.5%	donald trump	0.7%	independent	2.8%	news media	1.4%
trump	4.4%	lokale nachrichten	0.6%	lover (in Farsi)	2.6%	media organization	1.4%
conservative	4.3%	nachrichten aus	0.6%	social	2.6%	organization aim	1.4%
news	3.4%	hier kannst	0.6%	politics	2.6%	aim inspire	1.4%
maga	3.4%	kannst du	0.6%	media	2.4%	inspire action	1.4%
люблю	2.4%	du wichtige	0.6%	love	2.2%	action likes	1.4%
will	2.4%	wichtige und	0.6%	justice	2.0%	likes social	1.4%
proud	2.2%	und aktuelle	0.6%	low (in Farsi)	2.0%	social justice	1.4%

Table 6.9: Top 10 words and bigrams found in the descriptions of Russian and Iranian trolls on Twitter.

Accounts Characteristics

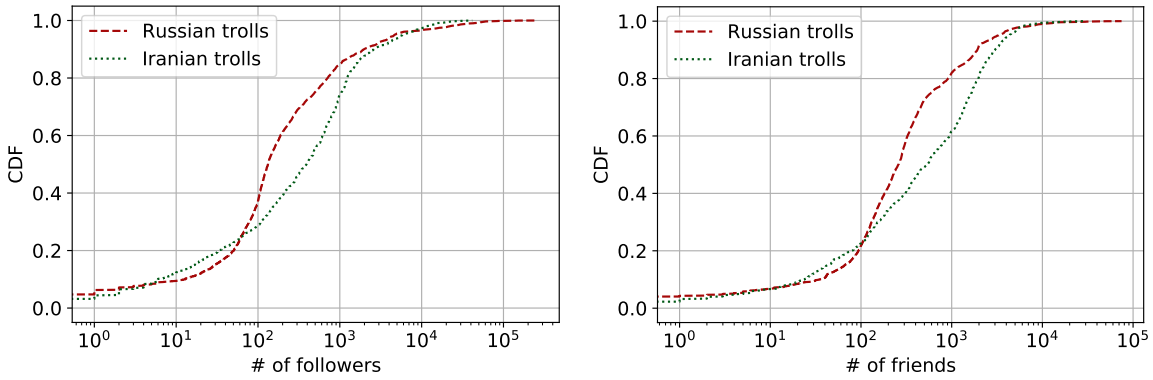
First we explore when the accounts appeared, what they posed as, and how many followers/friends they had on Twitter.

Account Creation. Fig. 6.11 plots the Russian and Iranian troll accounts creation dates on Twitter and Reddit. We observe that the majority of Russian troll accounts were created around the time of the Ukrainian conflict: 80% of have an account creation date earlier than 2016. That said, there are some meaningful peaks in account creation during 2016 and 2017. 57 accounts were created between July 3-17, 2016, which was right before the start of the Republican National Convention (July 18-21) where Donald Trump was named the Republican nominee

for President [328] . Later, 190 accounts were created between July, 2017 and August, 2017, during the run up to the infamous Unite the Right rally in Charlottesville [297]. Taken together, this might be evidence of coordinated activities aimed at manipulating users' opinions on Twitter with respect to specific events. This is further evidenced when examining the Russian trolls on Reddit: 75% of Russian troll accounts on Reddit were created in a single massive burst in the first half of 2015. Also, there are a few smaller spikes occurring just prior to the 2016 US Presidential election. For the Iranian trolls on Twitter we observe that they are much "younger," with the larger bursts of account creation *after* the 2016 US Presidential election.

Account Information. To avoid being obvious, state sponsored trolls might attempt to present a persona that masks their true nature or otherwise ingratiates themselves to their target audience. By examining the profile description of trolls we can get a feeling for how they might have cultivated this persona. In Table 6.9, we report the top ten words and bigrams that appear in profile descriptions of trolls on Twitter. Note that we do this only for Twitter trolls as we do not have descriptions for Reddit accounts. From the table we see that a relatively large number of Russian trolls pose as news outlets, with "news" (1.3%) and "breaking news" (0.8%) appearing in their description. Further, they seem to use their profile description to more explicitly increase their reach on Twitter, by nudging users to follow them (e.g., "follow me" appearing in almost 6.4% of profile descriptions). Finally, 3.4% of the Russian trolls describe themselves as Trump supporters: see "trump" (4.4%) and "maga" (3.4%) terms. Iranian trolls are even more likely to pose as news outlets or journalists: 3.6% have "journalist" and 3.2% have "news" in their profile descriptions. This highlights that accounts that pose as news outlets may in fact be accounts controlled by state-sponsored actors, hence regular users should critically think in order to assess whether the account is credible or not.

Followers/Friends. Fig. 6.12 plots the CDF of the number of followers and friends for both Russian and Iranian trolls. 25% of Iranian trolls had more than 1k followers, while the same applies for only 15% of the Russian trolls. In general, Iranian trolls tend to have more followers than Russian trolls (median of 392 and 132, respectively). Both Russian and Iranian trolls tend to follow a large number of users, probably in an attempt to increase their follower count via reciprocal follows. Iranian trolls have a median followers to friends ratio of 0.51, while Russian trolls have a ratio of 0.74. This might indicate that Iranian trolls were more effective in acquiring followers without resorting in massive followings of other users, or perhaps that they took advantages of services that offer followers for sale [329].



(a) (b)

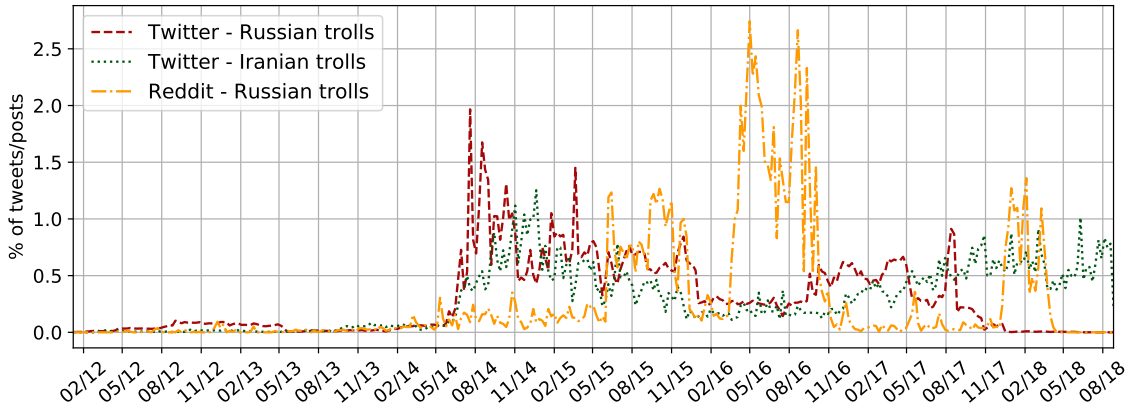
Figure 6.12: CDF of the number of a) followers and b) friends for the Russian and Iranian trolls on Twitter.

Temporal Analysis

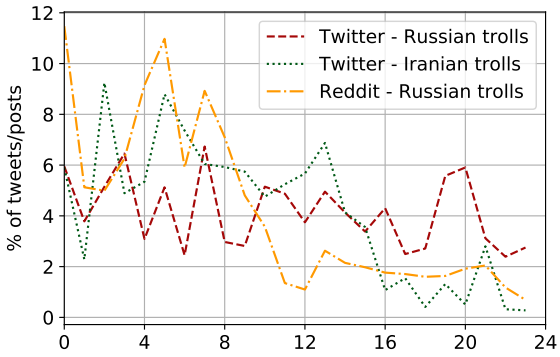
We next explore aggregate troll activity over time, looking for behavioral patterns. Fig. 6.13(a) plots the (normalized) volume of tweets/posts shared per week in our dataset. We observe that both Russian and Iranian trolls on Twitter became active during the Ukrainian conflict. Although lower in overall volume, there an increasing trend starts around August 2016 and continues through summer of 2017.

We also see three major spikes in activity by Russian trolls on Reddit. The first is during the latter half of 2015, approximately around the time that Donald Trump announced his candidacy for President. Next, we see solid activity through the middle of 2016, trailing off shortly before the election. Finally, we see another burst of activity in late 2017 through early 2018, at which point the trolls were detected and had their accounts locked by Reddit.

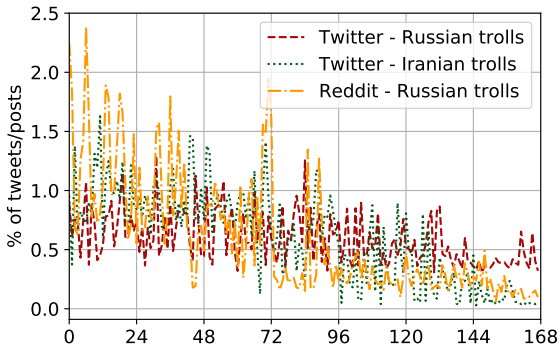
Next, we examine the hour of day and week that the trolls post. Fig. 6.13(b) shows that Russian trolls on Twitter are active throughout the day, while on Reddit they are particularly active during the first hours of the day. Similarly, Iranian trolls on Twitter tend to be active from early morning until 13:00 UTC. In Fig. 6.13(c), we report temporal characteristics based on hour of the week, finding that Russian trolls on Twitter follow a diurnal pattern with slightly less activity during Sunday. In contrast, Russian trolls on Reddit and Iranian trolls on Twitter are particularly active during the first days of the week, while their activity decreases during the weekend. For Iranians this is likely due to the Iranian work week being from Sunday to Wednesday with a half day on Thursday.



(a) Date



(b) Hour of Day



(c) Hour of Week

Figure 6.13: Temporal characteristics of tweets from Russian and Iranian trolls.

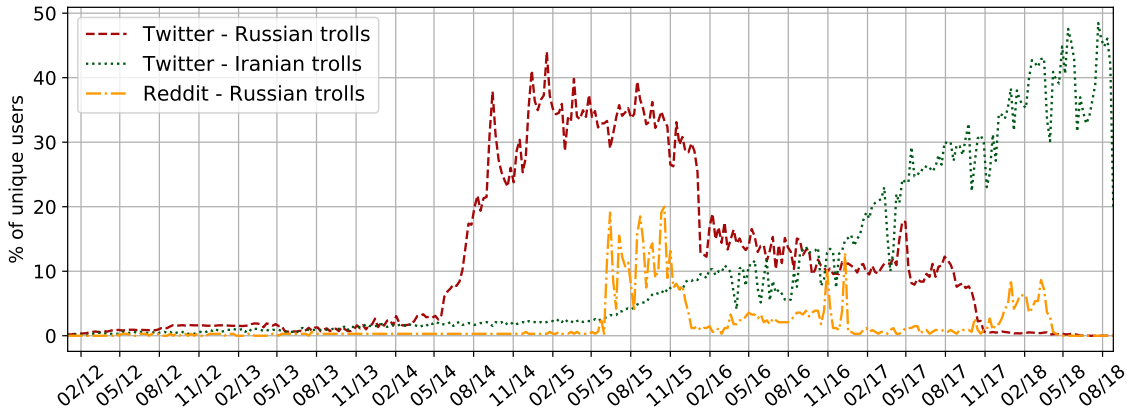


Figure 6.14: Percentage of unique trolls that were active per week.

But are *all* trolls in our dataset active throughout the span of our datasets? To answer this question, we plot the percentage of unique troll accounts that are active per week in Fig. 6.14

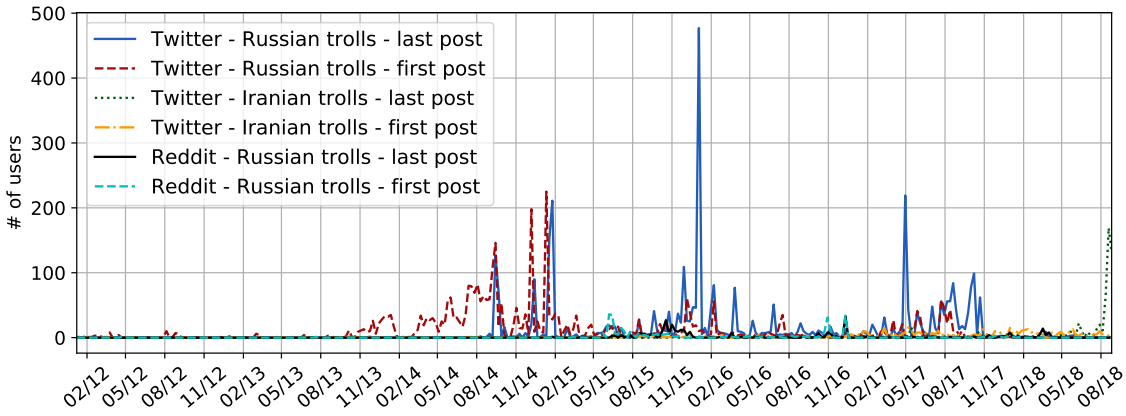


Figure 6.15: Number of trolls that posted their first/last tweet/post for each week in our dataset.

from which we draw the following observations. First, the Russian troll campaign on Twitter targeting Ukraine was much more diverse in terms of accounts when compared to later campaigns. There are several possible explanations for this. One explanation is that trolls learned from their Ukrainian campaign and became more efficient in later campaigns, perhaps relying on large networks of bots in their earlier campaigns which were later abandoned in favor of more focused campaigns like project Lakhta [330]. Another explanation could be that attacks on the US election might have required "better trained" trolls, perhaps those that could speak English more convincingly. The Iranians, on the other hand, seem to be slowly building their troll army over time. There is a steadily increasing number of active trolls posting per week over time. We speculate that this is due to their troll program coming online in a slow-but-steady manner, perhaps due to more effective training. Finally, on Reddit we see most Russian trolls posted irregularly, possibly performing other operations on the platform like manipulating votes on other posts.

Next, we investigate the point in time when each troll in our dataset made his first and last tweet. Fig. 6.15 shows the number of users that made their first/last post for each week in our dataset, which highlights when trolls became active as well as when they "retired." We see that Russian trolls on Twitter made their first posts during early 2014, almost certainly in response to the Ukrainian conflict. When looking at the last tweets of Russian trolls on Twitter we see that a substantial portion of the trolls "retired" by the end of 2015. In all likelihood this is because the Ukrainian conflict was over and Russia turned their information warfare arsenal to other targets (e.g., the USA, this is also aligned with the increase in the use of English language, see Section 6.2.3). When looking at Russian trolls on Reddit, we do not see a substantial spike in first posts close to the time that the majority of the accounts were created

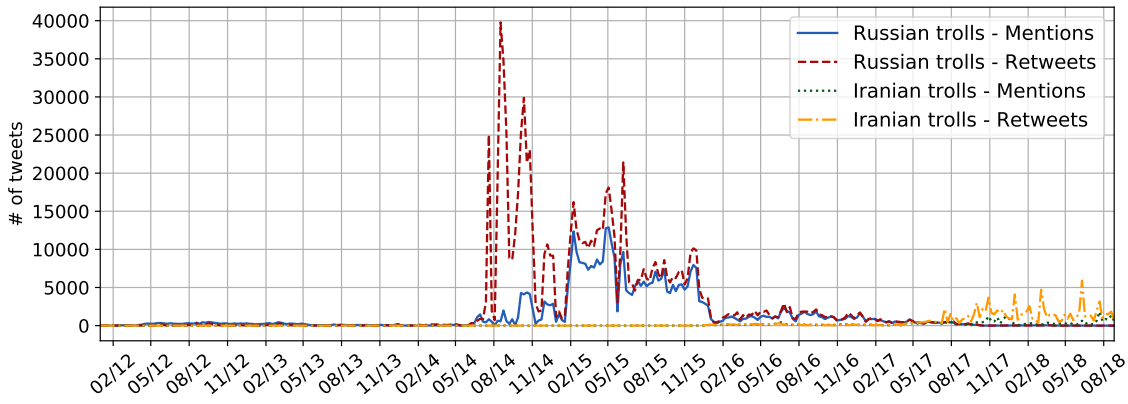


Figure 6.16: Number of tweets that contain mentions among Russian trolls and among Iranian trolls on Twitter.

(see Fig. 6.11). This indicates that the newly created Russian trolls on Reddit became active gradually (in terms of posting behavior).

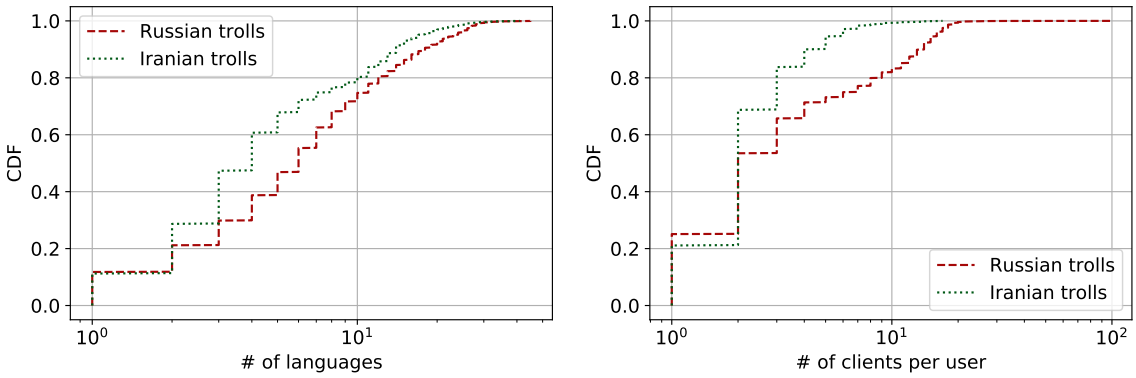
Finally, we assess whether Russian and Iranian trolls mention or retweet each other, and how this behavior occurs over time. Fig. 6.16 shows the number of tweets that were mentioning/retweeting other trolls' tweets over the course of our datasets. Russian trolls were particularly fond of this strategy during 2014 and 2015, while Iranian trolls started using this strategy after August, 2017. This again highlights how the strategies employed by trolls adapts and evolves to new campaigns.

Languages and Clients

In this section, we study the languages that Russian and Iranian Twitter trolls posted in, as well as their Twitter clients they used to make tweets (this information is not available for Reddit).

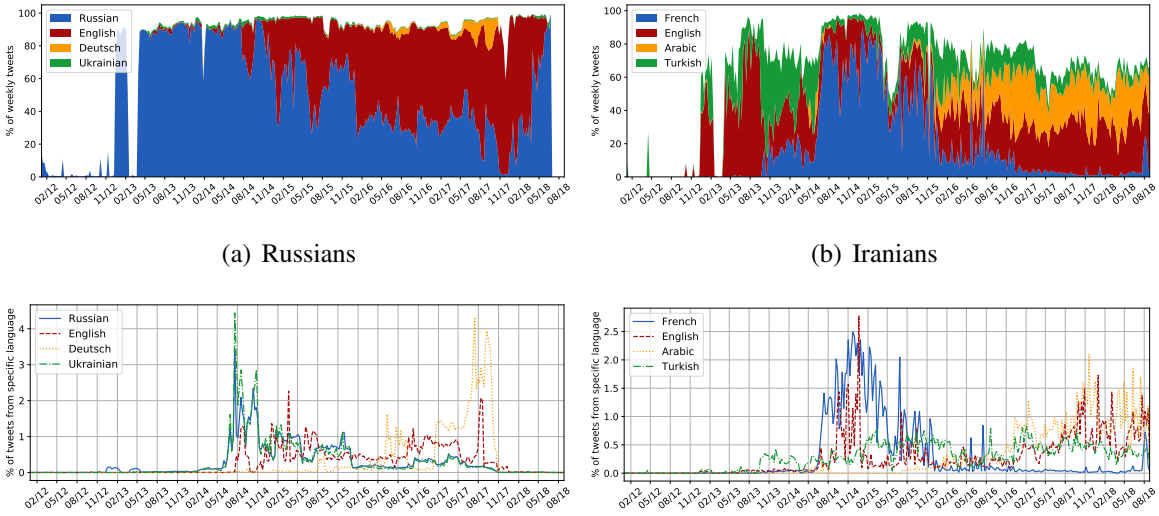
Languages. First we study the languages used by trolls as it provides an indication of their targets. The language information is included in the datasets released by Twitter. Fig. 6.17(a) plots the CDF of the number of languages used by troll accounts. We find that 80% and 75% of the Russian and Iranian trolls, respectively, use more than 2 languages. Next, we note that in general, Iranian trolls tend to use fewer languages than Russian trolls. The most popular language for Russian trolls is Russian (53% of all tweets), followed by English (36%), Deutsch (1%), and Ukrainian (0.9%). For Iranian trolls we find that French is the most popular language (28% of tweets), followed by English (24%), Arabic (13%), and Turkish (8%).

Fig. 6.18 plots the use of different languages over time. Fig. 6.18(a) and Fig. 6.18(b) plot the



(a) (b)

Figure 6.17: CDF of number of (a) languages used (b) clients used for Russian and Iranian trolls on Twitter.



(a) Russians (b) Iranians

(c) Russians (d) Iranians

Figure 6.18: Use of the four most popular languages by Russian and Iranian trolls over time on Twitter. (a) and (b) show the percentage of weekly tweets in each language. (c) and (d) show the percentage of total tweets per language that occurred in a given week.

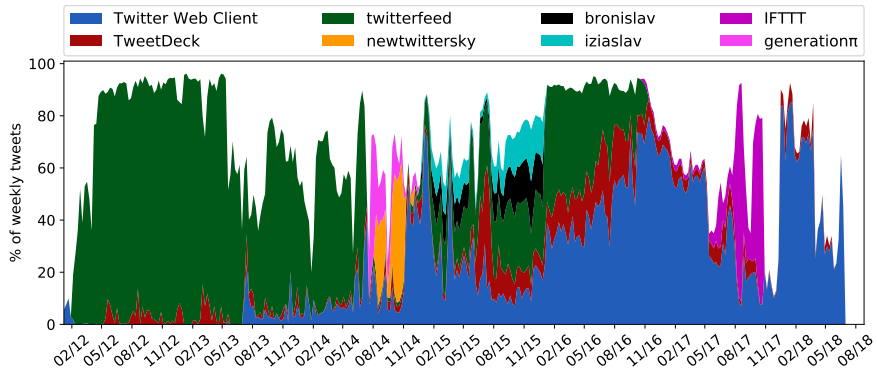
percentage of tweets that were in a given language on a given week for Russian and Iranian trolls, respectively, in a stacked fashion, which lets us see how the usage of different languages changed over time relative to each other. Fig. 6.18(c) and Fig. 6.18(d) plot the language use from a different perspective: normalized to the overall number of tweets in a given language. This view gives us a better idea of how the use of each particular language changed over time. From the plots we make the following observations. First, there is a clear shift in targets

based on the campaign. For example, Fig. 6.18(a) shows that the overwhelming majority of early tweets by Russian trolls were in Russian, with English only reaching the volume of Russian language tweets in 2016. This coincides with the "retirement" of several Russian trolls on Twitter (see Fig 6.15). Next, we see evidence of other campaigns, for example German language tweets begin showing up in early to mid 2016, and reach their highest volume in the latter half of 2017, in close proximity with the 2017 German Federal elections. Additionally, we note that Russian language tweets have a huge drop off in activity the last two months of 2017.

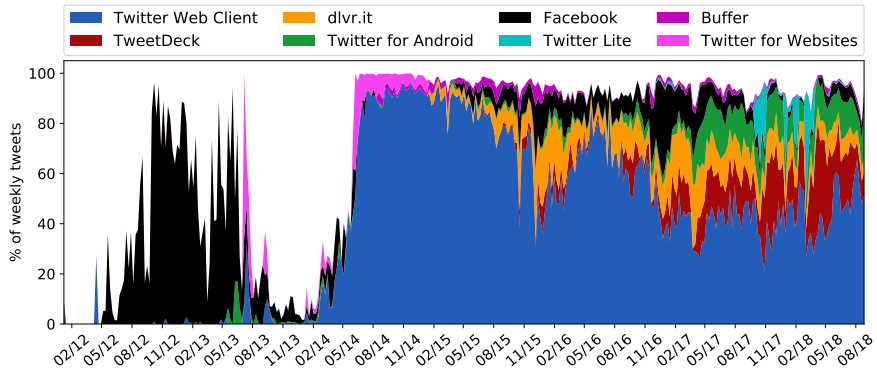
For the Iranians, we see more obvious evidence of multiple campaigns. For example, although Turkish and English are present for most of the timeline, French quickly becomes a commonly used language in the latter half of 2013, becoming the dominant language used from around May 2014 until the end of 2015. This is likely due to political events that happened during this time period. E.g., in November, 2013 France blocked a stopgap deal related to Iran's uranium enrichment program [331], leading to some fiery rhetoric from Iran's government (and apparently the launch of a troll campaign targeting French speakers). As tweets in French fall off, we also observe a dramatic increase in the use of Arabic in early 2016. This coincides with an attack on the Saudi embassy in Tehran [332], the primary reason the two countries ended diplomatic relations.

When looking at the language usage normalized by the total number of tweets in that language, we can get a more focused perspective. In particular, from Fig. 6.18(c) it becomes strikingly clear that the initial burst of Russian troll activity was targeted at Ukraine, with the majority of Ukrainian language tweets coinciding directly with the Crimean conflict [333]. From Fig. 6.18(d) we observe that English language tweets from Iranian trolls, while consistently present over time, have a relative peak corresponding with French language tweets, likely indicating an attempt to influence non-French speakers with respect to the campaign against French speakers.

Client usage. Finally, we analyze the clients used to post tweets. When looking at the most popular clients, we find that Russian and Iranian trolls use the main Twitter Web Client (28.5% for Russian trolls, and 62.2% for Iranian trolls). This is in contrast with what normal users use: using a random set of Twitter users, we find that mobile clients make up a large chunk of tweets (48%), followed by the TweetDeck dashboard (32%). We next look at how many different clients trolls use throughout our dataset: in Fig. 6.17(b), we plot the CDF of the number of clients used per user. 25% and 21% of the Russian and Iranian trolls, respectively,



(a) Russians



(b) Iranians

Figure 6.19: Use of the eight most popular clients by Russian and Iranian trolls over time on Twitter.

use only one client, while in general Russian trolls tend to use more clients than Iranians.

Fig. 6.19 plots the usage of clients over time in terms of weekly tweets by Russian and Iranian trolls. We observe that the Russians (Fig. 6.19(a)) started off with almost exclusive use of the "twitterfeed" client. Usage of this client drops off when it was shutdown in October, 2016. During the Ukrainian crisis, however, we see several new clients come into the mix. Iranians (Fig. 6.19(b)) started off almost exclusively using the "facebook" Twitter client. To the best of our knowledge, this is a client that automatically Tweets any posts you make on Facebook, indicating that Iranians likely started with a campaign on Facebook. At the beginning of 2014, we see a shift to using the Twitter Web Client, which only begins to decrease towards the end of 2015. Of particular note in Fig. 6.19(b) is the appearance of "dlvr.it," an automated social media manager, in the beginning of 2015. This corresponds with the creation of IUVM [334], which is a fabricated ecosystem of (fake) news outlets and social media accounts created by the Iranians, and might indicate that Iranian trolls stepped up their game around that time, starting

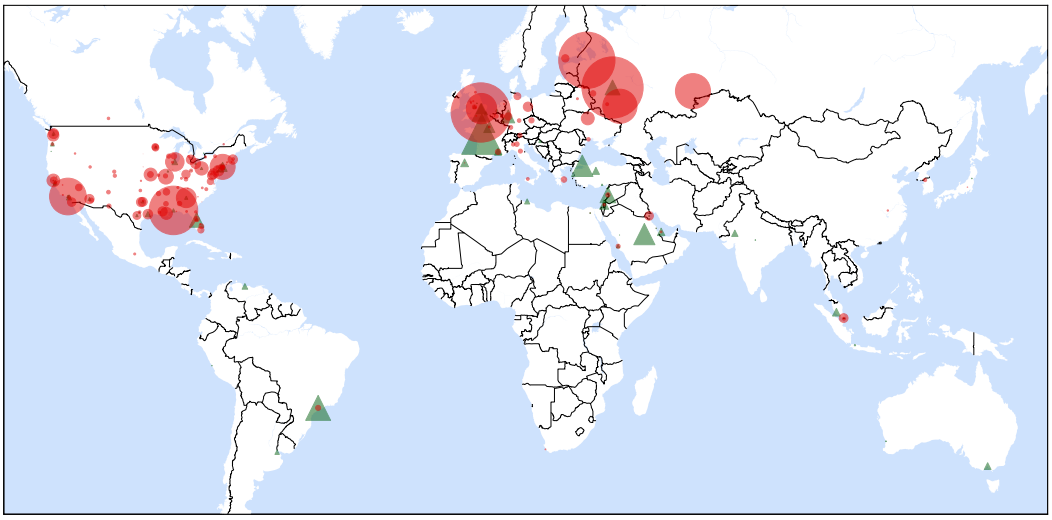


Figure 6.20: Distribution of reported locations for tweets by Russian trolls (100%) (red circles) and Iranian trolls (green triangles).

using services that allowed them for better account orchestration to run their campaigns more effectively.

Geographical Analysis

We then study users' location, relying on the self-reported location field in their profiles. Note that this field is not required, and users are also able to change it whenever they like, so we look at locations for each tweet. Note that 16.8% and 20.9% of the tweets from Russian and Iranians trolls, respectively, do not include a self-reported location. To infer the geographical location from the self-reported text, we use pigeo [335], which provides geographical information (e.g., latitude, longitude, country, etc.) given the text that corresponds to a location. Specifically, we extract 626 self-reported locations for the Russian trolls and 201 locations for the Iranian trolls. Then, we use pigeo to systematically obtain a geographical location (and its associated coordinates) for each text that corresponds to a location. Fig. 6.20 shows the locations inferred for Russian trolls (red circles) and Iranian trolls (green triangles). The size of the shapes on the map indicates the number of tweets that appear on each location. We observe that most of the tweets from Russian trolls come from locations within Russia (34%), the USA (29%), and some from European countries, like United Kingdom (16%), Germany (0.8%), and Ukraine (0.6%). This suggests that Russian trolls may be pretending to be from certain countries, e.g., USA or United Kingdom, aiming to pose as locals and effectively manipulate opinions. A similar pattern exists with Iranian trolls, which were particularly active in France (26%),

Russian trolls on Twitter		Iranian trolls on Twitter	
Word	**Cosine Similarity**	**Word**	**Cosine Similarity**
trumparmi	0.68	impeachtrump	0.81
trumptrain	0.67	stoptrump	0.80
votetrump	0.65	fucktrump	0.79
makeamericagreatagain	0.65	trumpisamoron	0.79
draintheswamp	0.62	dumptrump	0.79
trumppenc	0.61	ivankatrump	0.77
@realdonaldtrump	0.59	theresist	0.76
wakeupamerica	0.58	trumpresign	0.76
thursdaythought	0.57	notmypresid	0.76
realdonaldtrump	0.57	worstpresidentev	0.75
presidenttrump	0.57	antitrump	0.74

Table 6.10: Top 10 similar words to "maga" and their respective cosine similarities (obtained from the word2vec models).

Brazil (9%), the USA (8%), Turkey (7%), and Saudi Arabia (7%). It is also worth noting that Iranians trolls, unlike Russian trolls, did not report locations from their country, indicating that these trolls were primarily used for campaigns targeting foreign countries. Finally, we note that the location-based findings are in-line with the findings on the languages analysis (see Section 6.2.3), further evidencing that both Russian and Iranian trolls were specifically targeting different countries over time.

Content Analysis

Word Embeddings Recent indictments by the US Department of Justice have indicated that troll messaging was crafted, with certain phrases and terminology designated for use in certain contexts. To get a better handle on how this was expressed, we build two word2vec models on the corpus of tweets: one for the Russian trolls and one for the Iranian trolls. To train the models, we first extract the tweets posted in English, according to the data provided by Twitter. Then, we remove stop words, perform stemming, tokenize the tweets, and keep only words that appear at least 500 and 100 times for the Russian and Iranian trolls, respectively.

Table 6.10 shows the top 10 most similar terms to "maga" for each model. We see a marked difference between its usage by Russian and Iranian trolls. Russian trolls are clearly pushing heavily in favor of Donald Trump, while it is the exact opposite with Iranians.

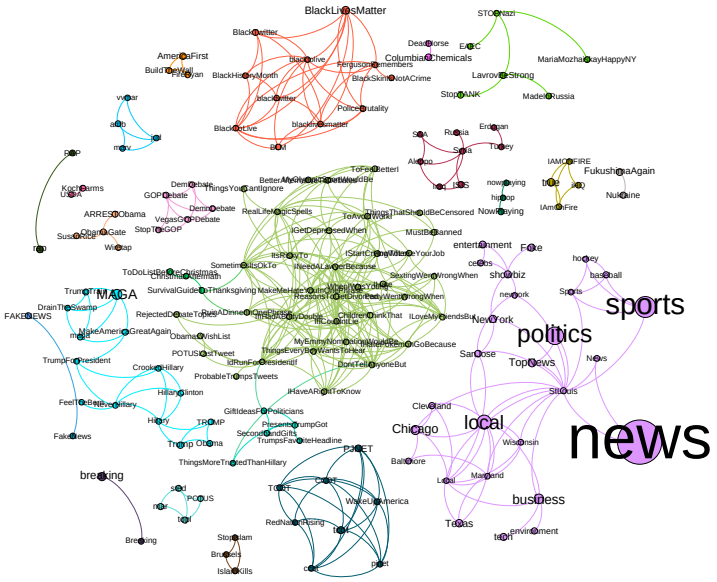
Russian trolls on Twitter				Iranian trolls on Twitter			
Hashtag	**(%)**	**Hashtag**	**(%)**	**Hashtag**	**(%)**	**Hashtag**	**(%)**
news	9.5%	USA	0.7%	Iran	1.8%	Palestine	0.6%
sports	3.8%	breaking	0.7%	Trump	1.4%	Syria	0.5%
politics	3.0%	TopNews	0.6%	Israel	1.1%	Saudi	0.5%
local	2.1%	BlackLivesMatter	0.6%	Yemen	0.9%	EEUU	0.5%
world	1.1%	true	0.5%	FreePalestine	0.8%	Gaza	0.5%
MAGA	1.1%	Texas	0.5%	QudsDay4Return	0.8%	SaudiArabia	0.4%
business	1.0%	NewYork	0.4%	US	0.7%	Iuvm	0.4%
Chicago	0.9%	Fukushima2015	0.4%	realiran	0.6%	InternationalQudsDay2018	0.4%
health	0.8%	quote	0.4%	ISIS	0.6%	Realiran	0.4%
love	0.7%	Foke	0.4%	DeleteIsrael	0.6%	News	0.4%

Table 6.11: Top 20 (English) hashtags in tweets from Russian and Iranian trolls on Twitter.

Hashtags. Next, we aim to understand the use of hashtags with a focus on the ones written in English. In Table 6.11, we report the top 20 English hashtags for both Russian and Iranian trolls. State-sponsored trolls appear to use hashtags to disseminate news (9.5%) and politics (3.0%) related content, but also use several that might be indicators of propaganda and/or controversial topics, e.g., #BlackLivesMatter. For instance, one notable example is: "WATCH: Here is a typical #BlackLivesMatter protester: 'I hope I kill all white babes!' #BatonRouge <url>" on July 17, 2016. Note that <url> denotes a link.

Fig. 6.21 shows a visualization of hashtag usage built from the two word2vec models. Here, we show hashgtags used in a similar context, by constructing a graph where nodes are words that correspond to hashtags from the word2vec models, and edges are weighted by the cosine distances (as produced by the word2vec models) between the hashtags. After trimming out all edges between nodes with weight less than a threshold, based on methodology from [338], we run the community detection heuristic presented in [13], and mark each community with a different color. Finally, the graph is layed out with the ForceAtlas2 algorithm [15], which takes into account the weight of the edges when laying out the nodes in 2-dimensional space. Note that the size of the nodes is proportional to the number of times the hashtag appeared in each dataset.

We first observe that, in Fig. 6.21(a) there is a central mass of what we consider "general audience" hashtags (see green community on the center of the graph): hashtags related to a holiday or a specific trending topic (but non-political) hashtag. In the bottom right portion of the plot we observe "general news" related categories; in particular American sports related hashtags (e.g., "baseball"). Next, we see a community of hashtags (light blue, towards the bottom left of the graph) clearly related to Trump's attacks on Hillary Clinton.

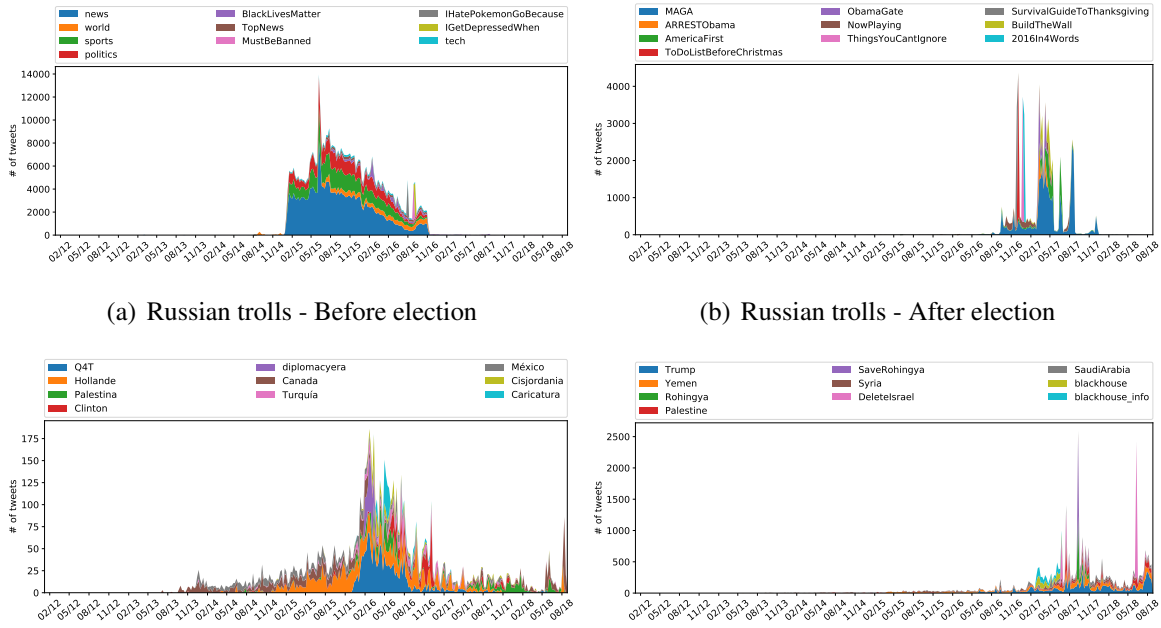


(a)

(b)

Figure 6.21: Visualization of the top hashtags used by a) Russian trolls on Twitter (see [336] for interactive version) and b) Iranian trolls on Twitter (see [337] for an interactive version).

The Iranian trolls again show different behavior. There is a community of hashtags related to nuclear talks (orange), a community related to Palestine (light blue), and a community that is clearly anti-Trump (pink). The central green community exposes some of the ways they pushed the IUVM fake news network by using innocuous hashtags like "#MyDatingProfileSays" as well as politically motivated ones like "#JerusalemIsTheEternalCapitalOfPalestine."



(a) Russian trolls - Before election

(b) Russian trolls - After election

(c) Iranians trolls - Before election

(d) Iranians trolls - After election

Figure 6.22: Top ten hashtags that appear a) c) substantially more times before the US elections rather than after the elections; and b) d) substantially more times after the elections rather than before.

We also study *when* these hashtags are used by the trolls, finding that most of them are well distributed over time. However we find some interesting exceptions. We highlight a few of these in Fig. 6.22, which plots the top ten hashtags that Russian and Iranian trolls posted with substantially different rates before and after the 2016 US Presidential election. The set of hashtags was determined by examining the relative change in posting volume before and after the election. From the plots we make several observations. First, we note that more general audience hashtags remain a staple of Russian trolls before the election (the relative decrease corresponds to the overall relative decrease in troll activity following the Crimea conflict). They also use relatively innocuous/ephemeral hashtags like #IHatePokemonGoBeacause, likely in an attempt to hide the true nature of their accounts. That said, we also see them attaching to politically divisive hashtags like #BlackLivesMatters around the time that Donald Trump won the Republican Presidential primaries in June 2016. In the ramp up to the 2016 election, we see a variety of clearly political related hashtags, with #MAGA seeing substantial peaks starting in early 2017 (higher than any peak during the 2016 Presidential campaigns). We also see a large number of politically ephemeral hashtags attacking Obama and a campaign to push the border wall between Mexico. In addition to these politically oriented hashtags, we again see the usage of ephemeral hashtags related to holidays. #SurvivalGuideToThanksgiving

Topic	Terms (Russian trolls on Twitter)	Topic	Terms (Iranian trolls on Twitter)
1	new, now, music, get, got, thanks, orleans, entertainment, follow, show	1	iran, will, deal, irantalks, iranian, nucleartalks, nuclear, irandeal, zarif, congress
2	sports, year, news, old, game, workout, win, nfl, chicago, morning	2	isis, new, state, fire, blackhouse, open, inferno, nation, will, turkish
3	day, love, one, foke, today, happy, first, away, last, time, will, best	3	yemen, press, front, liberty, children, saudi, isis, rohingya, school, king
4	can, don, like, people, just, know, get, want, will, never, good, make	4	isis, american, trump, sex, war, young, fbi, putin, terrorists, world
5	black, women, great, america, people, tcot, blacklivesmatter, read, american, isis	5	president, former, syria, obama, turkish, iraqi, russian, foreign, palestine, stop
6	news, police, man, local, woman, texas, killed, shooting, chicago, death	6	trump, donald, can, toonsonline, see, don, know, like, will, just
7	can, forget, change, wait, book, far, illegal, worst, words, save, united, done	7	saudi, israeli, attack, israel, days, terrorist, usa, palestinian, cia, third
8	exercise, wanna, fight, still, control, nice, gun, hold, perfect, enlist	8	isis, iran, first, realiran, siege, success, sydney, shame, tehran, photos
9	trump, obama, politics, president, hillary, clinton, breaking, just, house, video	9	saudi, united, states, isis, arabia, racist, society, structurally, oil, israel
10	news, world, business, health, new, says, money, tech, water, syria	10	israel, syria, police, syrian, muslim, video, people, death, trump, rights

Table 6.12: Terms extracted from LDA topics of tweets from Russian and Iranian trolls on Twitter.

Topic	Terms (Russian trolls on Reddit)
1	police, black, man, year, cop, video, woman, shot, white, arrested
2	love, one, absolutely, life, good, time, ever, wow, man, sure
3	man, dog, thank, back, thing, poor, now, happy, feeling, day
4	can, even, damn, cia, right, ledger, government, god, future, cap
5	just, will, one, really, can, people, think, time, like, need
6	like, people, just, don, looks, great, want, tie, also, tokens
7	police, cop, officer, state, man, rights, obama, shooting, death, omg
8	hillary, clinton, trump, new, lives, black, cute, matter, donald, recommend
9	will, don, can, people, get, just, understand, buy, nothing, btc
10	bitcoin, can, crypto, nice, people, try, just, tie, like, blockchain

Table 6.13: Terms extracted from LDA topics of posts from Russian trolls on Reddit.

in late November 2016 is particularly interesting as it was heavily used for discussing how to deal with interacting with family members with wildly different view points on the recent election results. This hashtag was exclusively used to give trolls a vector to sow discord. When it comes to Iranian trolls, we note that, prior to the 2016 election, they share many posts with hashtags related to Hillary Clinton (see Fig. 6.22(c)). After the election they shift to posting negatively about Donald Trump (see Fig. 6.22(d)).

LDA analysis. We also use the Latent Dirichlet Allocation (LDA) model [312] to analyze tweets' semantics. We train an LDA model for each of the datasets and extract ten distinct topics with ten words, as reported in Table 6.12. While both Russian and Iranian trolls tweet about politics related topics, for Iranian trolls, this seems to be focused more on regional, and possibly even internal issues. For example, "iran" itself is a common term in several of the topics, as is "israel," "saudi," "yemen," and "isis." While both sets of trolls discuss the proxy war in Syria (in which both states are involved), while the Iranian trolls have topics pertaining to Russia and Putin, the Russian trolls do not make any mention of Iran, instead focusing on more vague political topics like gun control and racism. For Russian trolls on Reddit (see Table 6.13) we again find topics related to politics as well some topics related to discussions about cryptocurrencies (see topics 9 and 10).

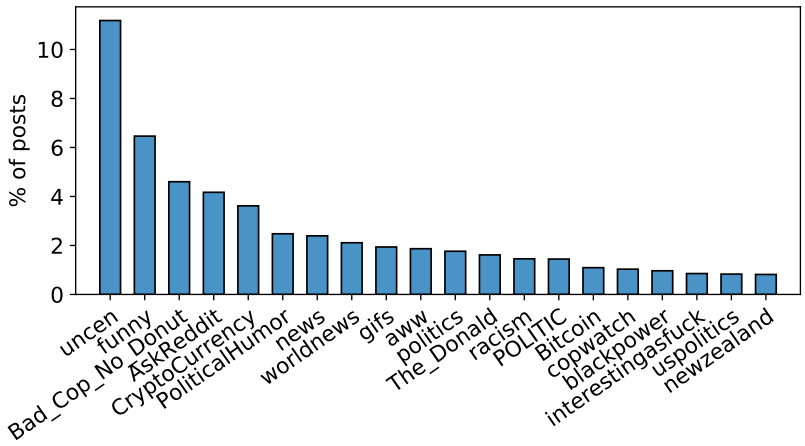


Figure 6.23: Top 20 subreddits that Russian trolls were active and their respective percentage of posts.

Subreddits. Fig. 6.23 shows the top 20 subreddits that Russian trolls on Reddit exploited and their respective percentage of posts over the whole dataset. The most popular subreddit is /r/uncen (11% of posts), which is a subreddit created by a specific Russian troll and, via manual examination, appears to be primarily used to disseminate news articles of questionable credibility. Other popular subreddits include general audience subreddits like /r/funny (6%) and /r/AskReddit (4%), likely in an attempt to obfuscate the fact that they are state-sponsored trolls in the same way that innocuous hashtags were used on Twitter. Finally, it is worth noting that the Russian trolls were particularly active on communities related to cryptocurrencies like /r/CryptoCurrency (3.6%) and /r/Bitcoin (1%) possibly attempting to influence the prices of specific cryptocurrencies. This is particularly noteworthy considering cryptocurrencies have been reportedly used to launder money, evade capital controls, and perhaps used to evade sanctions [339, 340].

URLs. We next analyze the URLs included in the tweets/posts. In Table 6.14, we report the top 20 domains for both Russian and Iranian trolls. Livejournal (5.4%) is the most popular domain in the Russian trolls dataset on Twitter, likely due the Ukrainian campaign. Overall, we can observe the impact of the Crimean conflict, with essentially all domains posted by the Russian trolls being Russian language or Russian oriented. One exception to Russian language sites is RT, the Russian-controlled propaganda outlet. The Iranian trolls similarly post more "localized" domains, for example, jordan-times, but we also see them heavily pushing the IUVM fake news network. When it comes to Russian trolls on Reddit, we find that they were mostly posting random images through Imgur (image-hosting site, 16% of the posts), likely in an attempt to accumulate karma score. We also note that a substantial portion of

Domain (Russian trolls on Twitter	(%)	Domain(Iranian trolls on Twitter)	(%)	Domain (Russian trolls on Reddit)	(%)
livejournal.com	5.4%	awdnews.com	29.3%	i.imgur	10.8%
riafan.ru	5.0%	dlvr.it	7.1%	blackmattersus.com	5.7%
twitter.com	2.5%	fb.me	4.8%	imgur.com	5.3%
ift.tt	1.8%	whatsupic.com	4.2%	donotshoot.us	2.5%
ria.ru	1.8%	googl.gl	3.9%	theguardian.com	1.0%
googl.gl	1.7%	realnienovosti.com	2.1%	nytimes.com	1.0%
dlvr.it	1.5%	twitter.com	1.7%	washingtonpost.com	0.8%
gazeta.ru	1.4%	libertyfrontpress.com	1.6%	huffingtonpost.com	0.8%
yandex.ru	1.2%	iuvmpress.com	1.5%	foxnews.com	0.8%
j.mp	1.1%	buff.ly	1.4%	youtube.com	0.8%
rt.com	0.8%	7sabah.com	1.3%	photographyisnotacrime.com	0.7%
nevnov.ru	0.7%	bit.ly	1.2%	thefreethoughtproject.com	0.6%
youtu.be	0.6%	documentinterdit.com	1.0%	butthis.com	0.5%
vesti.ru	0.5%	facebook.com	0.8%	cnn.com	0.5%
kievsmi.net	0.5%	al-hadath24.com	0.7%	dailymail.co	0.5%
youtube.com	0.5%	jordan-times.com	0.7%	rt.com	0.5%
kiev-news.com	0.5%	iuvmonline.com	0.6%	politico.com	0.4%
inforeactor.ru	0.4%	youtu.be	0.6%	truthdig.com	0.4%
lenta.ru	0.4%	alwaght.com	0.5%	nbcnews.com	0.4%
emaidan.com.ua	0.3%	ift.tt	0.5%	breitbart.com	0.4%

Table 6.14: Top 20 domains included in tweets/posts from Russian and Iranian trolls on Twitter and Reddit.

posts contained URLs to (fake) news sites linked with the Internet Research Agency like blackmattersus.com(5.7%) and donotshootus.us (2.5%).

6.2.4 Influence Estimation

Thus far, we have analyzed the behavior of Russian and Iranian trolls on Twitter and Reddit, with a special focus on how they evolved over time. Allegedly, one of their main goals is to manipulate the opinion of other users and extend the cascade of information that they share (e.g., lure other users into posting similar content) [341]. Therefore, we now set out to determine their impact in terms of the dissemination of information on Twitter, and on the greater Web.

To assess their influence, we look at three different groups of URLs: 1) URLs shared by Russian trolls on Twitter, 2) URLs shared by Iranian trolls on Twitter, and 3) URLs shared by both Russian *and* Iranian trolls on Twitter. We then find all posts that include any of these URLs in the following Web communities: Reddit, Twitter (from the 1% Streaming API, with posts from confirmed Russian and Iranian trolls removed), Gab, and 4chan's Politically

	Events per community							Total	
URLs shared by	/pol/	**Reddit**	**Twitter**	**Gab**	**The_Donald**	**Iran**	**Russia**	**Events**	**URLs**
Russians	76,155	366,319	1,225,550	254,016	61,968	0	151,222	2,135,230	48,497
Iranians	3,274	28,812	232,898	5,763	971	19,629	0	291,347	4,692
Both	331	2,060	85,467	962	283	334	565	90,002	153

Table 6.15: Total number of events in each community for URLs shared by a) Russian trolls; b) Iranian trolls; and c) Both Russian and Iranian trolls.

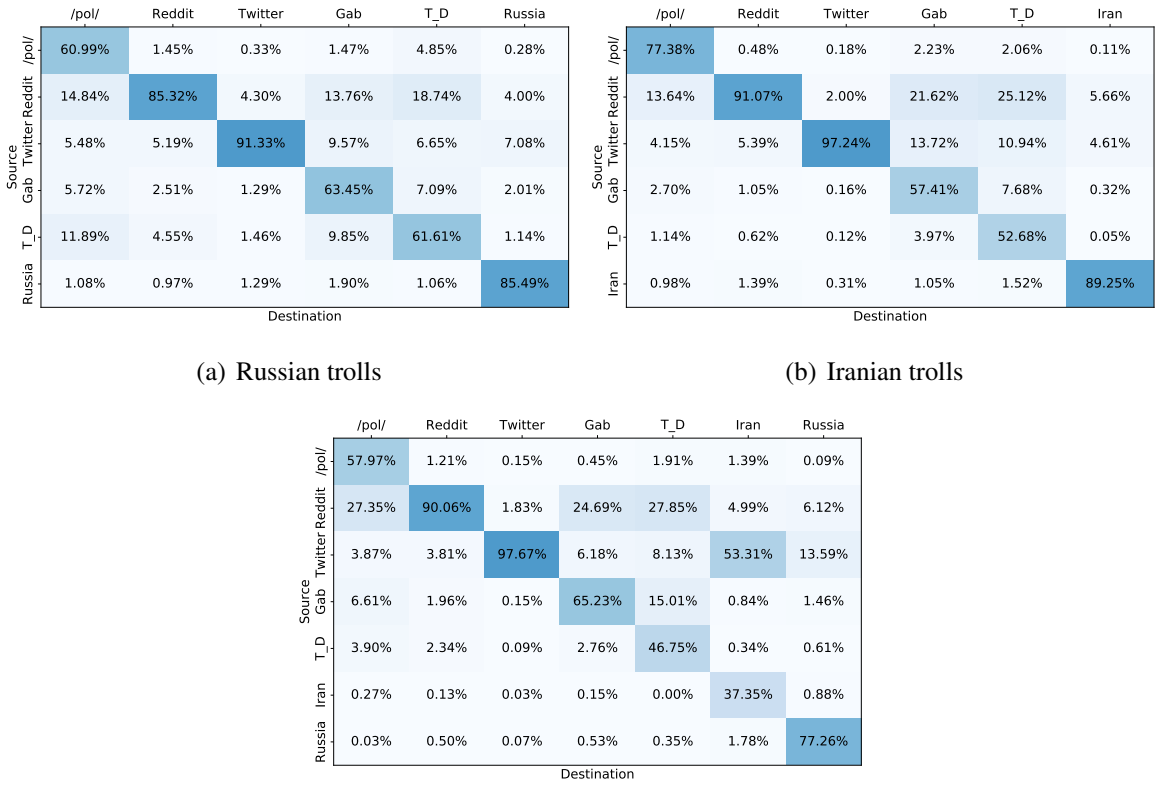
Incorrect board (/pol/). For Reddit and Twitter our dataset spans January 2016 to October 2018, for /pol/ it spans July 2016 to October 2018, and for Gab it spans August 2016 to October 2018.[2] We select these communities as previous work shows they play an important and influential role on the dissemination of news [66] and memes [342].

Table 6.15 summarizes the number of events (i.e., occurrences of a given URL) for each community/group of users that we consider (Russia refers to Russian trolls on Twitter, while Iran refers to Iranian trolls on Twitter). Note that we decouple The_Donald from the rest of Reddit as previous work showed that it is quite efficient in pushing information in other communities [342]. From the table we make several observations: 1) Twitter has the largest number of events in all groups of URLs mainly because it is the largest community and 2) Gab has a considerably large number of events; more than /pol/ and The_Donald, which are bigger communities.

For each unique URL, we fit a statistical model known as Hawkes Processes [27, 28], which allows us to estimate the strength of connections between each of these communities in terms of how likely an event – the URL being posted by either trolls or normal users to a particular platform – is to cause subsequent events in each of the groups. We fit each Hawkes model using the methodology presented by [342]. In a nutshell, by fitting a Hawkes model we obtain all the necessary parameters that allow us to assess the root cause of each event (i.e., the community that is "responsible" for the creation of the event). By aggregating the root causes for all events we are able to measure the influence and efficiency of each Web community we considered.

We demonstrate our results with two different metrics: 1) the absolute influence, or percentage of events on the destination community caused by events on the source community and 2) the

[2] **NB:** the 4chan dataset made available by the authors of [66, 342] starts in late June 2016 and Gab was first launched in August 2016.



(a) Russian trolls

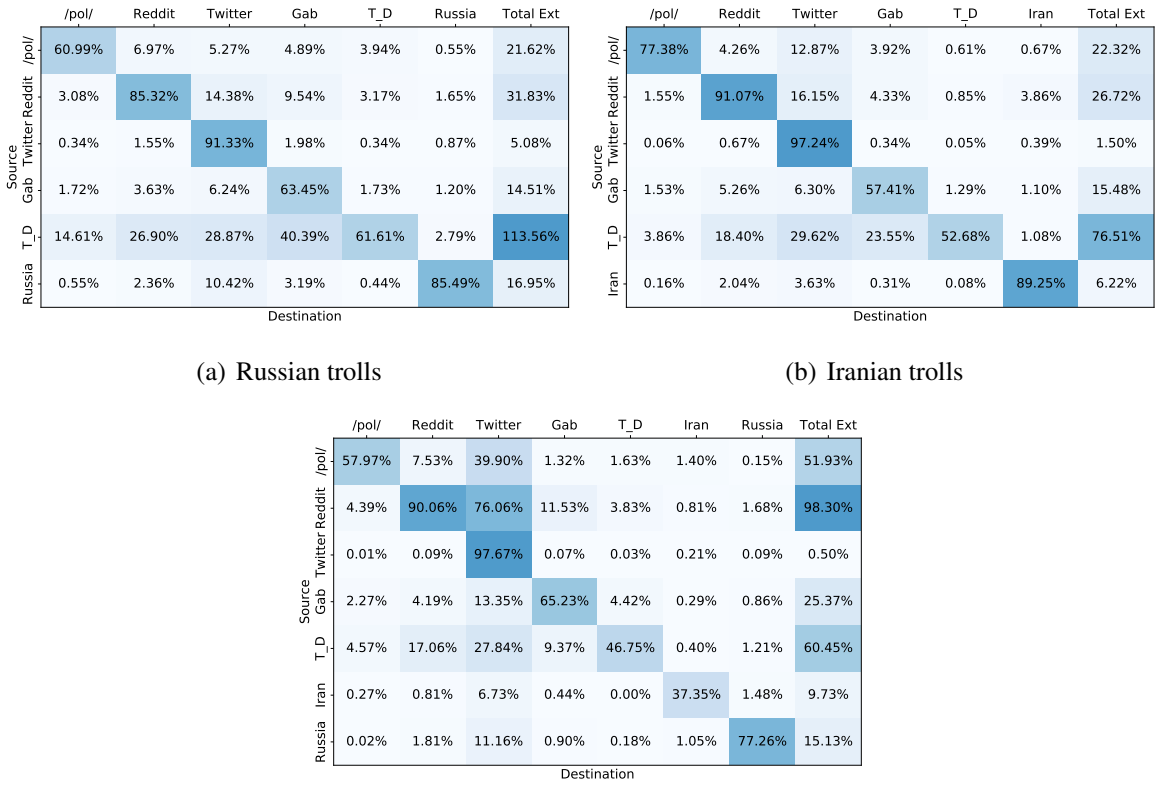
(b) Iranian trolls

(c) Both

Figure 6.24: Percent of *destination* events caused by the source community to the destination community for URLs shared by a) Russian trolls; b) Iranian trolls; and c) both Russian and Iranian trolls.

influence relative to size, which shows the number of events caused on the destination platform as a percent of the number of events on the *source* platform. The latter can also be interpreted as a measure of how *efficient* a community is in pushing URLs to other communities.

Fig. 6.24 reports our results for the absolute influence for each group of URLs. When looking at the influence for the URLs shared by Russian trolls on Twitter (Fig. 6.24(a)), we find that Russian trolls were particularly influential to users from Gab (1.9%), the rest of Twitter (1.29%), and /pol/ (1.08%). When looking at the communities that influenced the Russian trolls we find the rest of Twitter (7%) followed by Reddit (4%). By looking at URLs shared by Iranian trolls on Twitter (Fig. 6.24(b)), we find that Iranian trolls were most successful in pushing URLs to The_Donald (1.52%), the rest of Reddit (1.39%), and Gab (1.05%), somewhat ironic considering The_Donald and Gab's zealous pro-Trump leanings and the Iranian trolls' clear anti-Trump leanings [343, 43]. Similarly to Russian trolls, the Iranian



(a) Russian trolls

(b) Iranian trolls

(c) Both

Figure 6.25: Influence from source to destination community, normalized by the number of events in the *source* community for URLs shared by a) Russian trolls; b) Iranian trolls; and c) Both Russian and Iranian trolls. We also include the total external influence of each community.

trolls were most influenced by Reddit (5.6%) and the rest of Twitter (4.6%). When looking at the URLs posted by both Russian and Iranian trolls we find that, overall, the Russian trolls were more influential in spreading URLs to the other Web communities with the exception of (again, somewhat ironically) /pol/.

But how do these results change when we normalize the influence with respect to the number of events that each community creates? Fig. 6.25 shows the influence relative to size for each pair of communities/groups of users. For URLs shared by Russian trolls (Fig. 6.25(a)) we find that Russian trolls were particularly efficient in spreading the URLs to Twitter (10.4%)—which is not a surprise, given that the accounts operate directly on this platform—and Gab (3.19%). For the URLs shared by Iranian trolls, we again observe that were most efficient in pushing the URLs to Twitter (3.6%), and the rest of Reddit (2.04%). Also, it is worth noting that in both groups of URLs The_Donald had the highest external influence to the other platforms.

This highlights that The_Donald is an impactful actor in the information ecosystem and is quite possibly exploited by trolls as a vector to push specific information to other communities. Finally, when looking at the URLs shared by both Russian and Iranian trolls, we find that Russian trolls were more efficient (greater impact relative to the number of URLs posted) at spreading URLs in all the communities with the exception of /pol/, where Iranians were more efficient.

6.2.5 Remarks

In this work, we analyzed the behavior and evolution of Russian and Iranian trolls on Twitter and Reddit during the course of several years. We shed light to the target campaigns of each group of trolls, we examined how their behavior evolved over time, and what content they disseminated. Furthermore, we find some interesting differences between the trolls depending on their origin and the platform from which they operate. For instance, for the latter, we find discussions related to cryptocurrencies only on Reddit by Russian trolls, while for the former we find that Russian trolls were pro-Trump and Iranian trolls anti-Trump. Also, we quantify the influence that these state-sponsored trolls had on several mainstream and alternative Web communities (Twitter, Reddit, /pol/, and Gab), showing that Russian trolls were more efficient and influential in spreading URLs on other Web communities than Iranian trolls, with the exception of /pol/.

Our findings have serious implications for society at large. First, our analysis shows that while troll accounts use peculiar tactics and talking points to further their agendas, these are not completely disjoint from regular users, and therefore developing automated systems to identify and block such accounts remains an open challenge. Second, our results also indicate that automated systems to detect trolls are likely to be difficult to realize: trolls change their behavior over time, and thus even a classifier that works perfectly on one campaign might not catch future campaigns. Third, and perhaps most worrying, we find that state-sponsored trolls have a meaningful amount of influence on fringe communities like The_Donald, 4chan's /pol/, and Gab, and that the topics pushed by the trolls resonate strongly with these communities. This might be due to users on these communities that sympathize with the views the trolls aim to share (i.e., "useful idiots") or to unidentified state-sponsored actors on these communities. In either case, considering recent tragic events like the Tree of Life Synagogue shootings, perpetuated by a Gab user seemingly influenced by content posted there, the potential for mass societal upheaval cannot be overstated. Because of this, we implore the research community,

as well as governments and non-government organizations to expend whatever resources are at their disposal to develop technology and policy to address this new, and effective, form of digital warfare.

Chapter 7

Discussion & Conclusions

In this thesis, we studied several aspects of the information ecosystem on the Web. We shed light into three main relevant lines of work: 1) understanding the spread of information through the lens of multiple Web communities and modeling the interplay between them; 2) characterizing emerging Web communities and services by undertaking exploratory large-scale quantitative analyses; and 3) understanding the behavior and impact of state-sponsored actors on the information ecosystem on the Web. Below, for each line of work, we provide the main take-aways and possible future directions.

7.1 Understanding the Spread Of Information Through The Lens Of Multiple Web Communities

Remarks. In this line of work, we studied the spread of news and image-based memes across four Web communities, namely, Twitter, Reddit, 4chan's /pol/ and Gab. By designing and developing a scalable processing pipeline we were able to detect and track the propagation of memes across the Web. Then using Hawkes Processes, we modeled the interplay between the various Web communities and we quantified the influence that each community have to the other with respect to the dissemination of news and memes. The main take-aways from this work are: 1) small fringe Web communities like 4chan's /pol/ and The_Donald subreddit have a surprisingly strong influence, despite their small size, to mainstream communities like Twitter; and 2) we find important differences between the communities we study with regard to the dissemination of news and memes. For instance, for news, we find that users on different communities prefer different news sources, especially for the alternative ones, while

for memes, we find that users on small fringe Web communities tend to share more memes that are likely to be used in a weaponized or hateful context.

Future Directions. There are several possible future directions that derive from the findings of this thesis. First, we present a novel methodology for assessing the influence between multiple Web communities. This framework, based on Hawkes Processes, can be used in a lot of different domains to assess the influence between various entities. For instance, this framework can be applied in the user-level in order to assess the influence that users of a specific community have to each other with respect to the dissemination of a specific information. Also, by changing the notion of what a process and what an event is in the framework one can make interesting influence estimation studies. With regard to the dissemination of news, an interesting future direction is to leverage Natural Language Processing techniques in order to understand how news articles are discussed on various Web communities and if there are important differences between the various Web communities in consideration. With regard to the dissemination of memes there are several future directions that can be based on our developed memes processing pipeline. Specifically, one can leverage our pipeline to detect images pertaining to specific memes and then qualitatively analyze them in order to understand how memes are becoming weaponized and how multiple memes are combined together to deliver a specific idea. For instance, to study how the Pepe the Frog meme is used in conjunction with other memes with the goal to deliver a specific political message. Another line of work, includes focusing on the detection of potential hateful and harmful memes and devising mitigation strategies that will be employed by Web communities (e.g., Twitter) in order to safeguard their users from potentially offensive content. Finally, the developed pipeline can be used to study images that are not bounded to a specific domain (e.g., memes). For instance, [344] demonstrate how our image processing pipeline can be used in conjunction with the Google Cloud Vision API to characterize the images posted by Russian trolls on Twitter.

7.2 Characterizing the Role of Emerging Web Communities and Services on the Information Ecosystem

Remarks. In this line of work, we have explored the Gab social network as well as two Web archiving services: the Wayback Machine and `archive.is`, with the goal to assess their role on the information ecosystem. We find several interesting findings when exploring

these communities and services. First, we find that Gab attracts the interest of the alt-right community as the most popular users in the platform are alt-right celebrities. Also, we find that its users have a preference in sharing news articles from alternative news sources, while when examining the prevalence of hate speech, we find that it exhibits a high degree of hate speech. Second, after our comprehensive analysis on the use of Web archiving services, we find that they are particularly popular in fringe Web communities for the preservation of Web content. In addition, we find that these services are extensively used by Reddit bots to preserve content posted on specific subreddits, and that Reddit moderators "force" users to share archived URLs from news sources with conflicting ideology in order to penalize their ad revenue. Overall, these findings indicate that Web archiving services are an important actor on the information ecosystem and that it should be taken into account for studies that focus on URLs.

Future Directions. There are several future directions that can be derived from this line of work. First, the Gab social network is still relatively unstudied when compared to other mainstream communities like Twitter, hence a lot of its aspects are unclear. For instance, it will be interesting to study the evolution of Gab users over time and whether they are becoming more hateful/radicalized over time. Also, it will be interesting to study whether Gab's popularity increases with purges or large bans of users from other popular Web communities like Twitter. Furthermore, it will be interesting to study the prevalence of automated accounts within the platform and whether they are trying to promote specific talking points (in our work we find some anecdotal evidence of spam bots on Gab). Finally, we implore the research community to qualitatively study the Gab community in order to shed light into emerging Web phenomena like hate speech, fake news, and online radicalization.

In a more broad direction, there are still a lot of Web communities for which we lack a clear understanding of what their role on the Web information ecosystem is. For example, Web communities like Discord, WhatsApp, Mastodon, and Telegram, are relatively unstudied and it is unclear whether they contribute in the spread of false information on the Web. A possible future direction is providing characterization of these Web communities and assessing whether campaigns are organized in such communities, especially in the ones that support private channels like Discord, WhatsApp, and Telegram.

Our work on Web archiving services points to several research avenues. For instance, future work could better understand the role of archiving services in the dis/misinformation ecosystem, e.g., with respect to the content that gets archived and the context in which archive URLs

are disseminated. Moreover, further work could shed light on the actors archiving specific URLs in specific contexts, as well as how much traction they get on Web communities like Twitter and Reddit. Finally, we believe that a deeper dive into the socio-technical and ethical implications of archiving services is warranted: they serve a crucial role in ensuring that Web content persists, but do so without regard to (and often in spite of) the rights and consent of content producers.

7.3 Towards Understanding State-Sponsored Actors

Remarks. In this work we provide a comprehensive exploratory analysis on the behavior of state-sponsored actors on the Web. First, we compare the behavior of state-sponsored actors on Twitter and how they compare to a set of random users. We find important differences between state-sponsored actors and random users ranging from the use of Twitter clients to their self-reported locations. In addition, we provide useful insights with regard to the evolution of their accounts and how do they posed as. Second, we provide an analysis of Russian and Iranian trolls on Twitter and Reddit. We investigated how they evolved over time and what influence they have in other communities: namely, Reddit, Twitter, 4chan's /pol/, and Gab. We find that the behavior and targets of these actors vary over time and that these actors were particularly influential in spreading news articles to other Web communities. In particular, we find that the Russia state-sponsored actors were more influential in spreading news to the other communities, with the exception of /pol/ where Iranian trolls were more influential.

Future Directions. Despite providing a comprehensive overview on the behavior of state-sponsored actors on the Web, there are still several unexplored research avenues. First, as a research community, we should develop tools to detect and mitigate campaigns organized from state-sponsored actors. Second, there is a variety of other state-sponsored actors that are unexplored. For instance, Twitter detected and suspended a lot of accounts associated with Venezuela and Bangladesh governments [326]: it will be interesting to see how these actors compare with the Russian and Iranian actors presented in this thesis and whether there are meaningful differences in their behavior. Finally, we believe that our influence estimation results show the need for more sophisticated measurements in the domains of opinion manipulation and spread of false information by state-sponsored actors.

7.4 Conclusion

In this thesis, we shed some light into the complex Web information ecosystem through the lens of multiple Web communities. Our work reveals the need to take a cross-platform view of the information ecosystem as there are a lot of Web communities that despite their small size are particularly influential and can have real-world impact. Also, it indicates the need to develop sophisticated tools and techniques to detect the spread of information across the Web by considering the diverse types of information (i.e., text, images, and URLs). We argue that the aforementioned are of paramount importance for getting a more representative view of the information ecosystem, hence helping in better understanding this ecosystem as a whole.

Bibliography

[1] David A Broniatowski et al. "Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate". In: *American journal of public health* 108.10 (2018), pp. 1378–1384.

[2] *How Russian trolls' support of third parties could have cost Hillary Clinton the election*. https://qz.com/1210369/russia-donald-trump-2016-how-russian-trolls-support-of-us-third-parties-may-have-cost-hillary-clinton-the-election/.

[3] Alex Hern. *Cambridge Analytica: how did it turn clicks into votes?* `https://www.theguardian.com/news/2018/may/06/cambridge-analytica-how-turn-clicks-into-votes-christopher-wylie`. 2017.

[4] Megha Mohan. *Macron Leaks: The anatomy of a hack*. `http://www.bbc.co.uk/news/blogs-trending-39845105`. 2017.

[5] BBC. *The saga of "Pizzagate": The fake story that shows how conspiracy theories spread*. `https://www.bbc.com/news/blogs-trending-38156985`. 2016.

[6] Alex Dackevych. *"CNN wrestling" tweet came from extreme Reddit user*. `https://www.bbc.com/news/blogs-trending-40483914`. 2017.

[7] *Gab Landing Page*. `https://gab.ai/`. 2018.

[8] Thor Benson. *Inside the "Twitter for racists": Gab – the site where Milo Yiannopoulos goes to troll now*. `https://goo.gl/Yqv4Ue`. 2016.

[9] Wallace Koehler. "A longitudinal study of Web pages continued: A consideration of document persistence". In: *Information Research* 9.2 (2004).

[10] Alan G Hawkes. "Spectra of some self-exciting and mutually exciting point processes". In: *Biometrika* 58.1 (1971), pp. 83–90.

[11] R. Killick, P. Fearnhead, and I. A. Eckley. “Optimal Detection of Changepoints With a Linear Computational Cost”. In: *Journal of the American Statistical Association* 107.500 (2012), pp. 1590–1598. DOI: 10.1080/01621459.2012.737745. eprint: https://doi.org/10.1080/01621459.2012.737745. URL: https://doi.org/10.1080/01621459.2012.737745.

[12] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems*. 2013.

[13] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* (2008).

[14] Shawn Martin et al. “OpenOrd: an open-source toolbox for large graph layout”. In: *IS&T/SPIE*. 2011.

[15] Mathieu Jacomy et al. “ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software”. In: *PloS one* (2014).

[16] Martin Ester et al. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. In: *ACM KDD*. 1996.

[17] *Paper Repository*. https://github.com/memespaper/memes_pipeline.

[18] Abby Ohlheiser. *Reddit bans r/greatawakening, the main subreddit for QAnon conspiracy theorists*. https://www.washingtonpost.com/news/the-intersect/wp/2018/09/12/reddit-bans-r-greatawakening-the-main-subreddit-for-qanon-conspiracy-theorists/. 2018.

[19] Gabriel Emile Hine et al. “Kek, Cucks, and God Emperor Trump: A Measurement Study of 4chan’s Politically Incorrect Forum and Its Effects on the Web”. In: *AAAI ICWSM*. 2017.

[20] Mike Wendling. *The saga of ’Pizzagate’: The fake story that shows how conspiracy theories spread*. http://www.bbc.com/news/blogs-trending-38156985. 2016.

[21] *Reddit FAQ - Karma*. https://www.reddit.com/wiki/faq. 2017.

[22] Peter Snyder et al. “Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing”. In: *ACM IMC*. 2017.

[23] Andrew Torba. *How You Can Help Support Gab*. https://medium.com/@Torbahax/gab-donations-9ca2a5c0557e. 2016.

[24] *A Censorship-Proof P2P Social Media Protocol*. https://www.startengine.com/gab-select. 2017.

[25] *French fear Putin and Trump followers are using 4chan to disrupt presidential election*. https://venturebeat.com/2017/05/05/french-fear-putin-and-trump-followers-are-using-4chan-to-disrupt-presidential-election/.

[26] Rob Price. *Google's app store has banned Gab, a social network popular with the far-right, for 'hate speech'*. http://uk.businessinsider.com/google-app-store-gab-ban-hate-speech-2017-8. 2017.

[27] Scott W. Linderman and Ryan P. Adams. "Discovering Latent Network Structure in Point Process Data". In: *ICML*. 2014.

[28] S. W. Linderman and R. P. Adams. "Scalable Bayesian Inference for Excitatory Point Process Networks". In: 2015.

[29] *Fake news. It's complicated.* https://firstdraftnews.com/fake-news-complicated/.

[30] Srijan Kumar and Neil Shah. "False information on web and social media: A survey". In: *arXiv preprint arXiv:1804.08559* (2018).

[31] Victoria L Rubin, Niall J Conroy, and Yimin Chen. "Towards News Verification: Deception Detection Methods for News Discourse". In: (2015).

[32] Steven Heller. *Bat Boy, Hillary Clinton's Alien Baby, and a Tabloid's Glorious Legacy*. https://www.theatlantic.com/entertainment/archive/2014/10/the-ingenious-sensationalism-of-the-weekly-world-new/381525/. 2014.

[33] Garth S Jowett and Victoria O'donnell. *Propaganda & persuasion*. Sage, 2014.

[34] Medium. *Different Examples of Propaganda in Social Media*. https://medium.com/@VasquezNnenna/different-examples-of-propaganda-in-social-media-758fc98d021d. 2018.

[35] Mark Fenster. *Conspiracy theories: Secrecy and power in American culture*. U of Minnesota Press, 1999.

[36] Wikipedia. *Pizzagate conspiracy theory*. https://en.wikipedia.org/wiki/Pizzagate_conspiracy_theory. 2017.

[37] Wikipedia. *Murder of Seth Rich*. https://en.wikipedia.org/wiki/Murder_of_Seth_Rich. 2017.

[38] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes". In: *ACM WWW*. 2016.

[39] *Definition of Half-truth*. https://www.merriam-webster.com/dictionary/half-truth.

[40] *The Origins of Writerly Words*. http://time.com/82601/the-origins-of-writerly-words/.

[41] Snopes. *Adam Sandler Death Hoax*. https://www.snopes.com/fact-check/adam-sandler-death-hoax-2/. 2017.

[42] Martin Potthast et al. "A Stylometric Inquiry into Hyperpartisan and Fake News". In: *arXiv preprint arXiv:1702.05638* (2017).

[43] Savvas Zannettou et al. “What is Gab? A Bastion of Free Speech or an Alt-Right Echo Chamber?” In: *ACM WWW Companion*. 2018.

[44] Warren A Peterson and Noel P Gist. “Rumor and public opinion”. In: *American Journal of Sociology* (1951).

[45] Snopes. *Boston Marathon Bombing Rumors*. https://www.snopes.com/fact-check/boston-marathon-bombing-rumors/. 2013.

[46] Yimin Chen, Niall J Conroy, and Victoria L Rubin. “Misleading Online Content: Recognizing Clickbait as False News”. In: *MDD*. 2015.

[47] W Joseph Campbell. *Yellow journalism: Puncturing the myths, defining the legacies*. 2001.

[48] Politifact. *The more outrageous, the better: How clickbait ads make money for fake news sites*. http://www.politifact.com/punditfact/article/2017/oct/04/more-outrageous-better-how-clickbait-ads-make-mone/. 2017.

[49] Clint Burfoot and Timothy Baldwin. “Automatic satire detection: Are you having a laugh?” In: *ACL-IJCNLP*. 2009.

[50] *The Onion*. http://www.theonion.com/.

[51] *SatireWire*. http://www.satirewire.com/.

[52] Yazan Boshmaf et al. “The socialbot network: when bots socialize for fame and money”. In: *ACSAC*. 2011.

[53] Samer Al-khateeb and Nitin Agarwal. “Examining botnet behaviors for propaganda dissemination: A case study of isil’s beheading videos-based propaganda”. In: *IEEE ICDMW*. 2015.

[54] The New Yorker. *How the NRA Manipulates Gun Owners and the Media*. https://www.newyorker.com/news/news-desk/how-the-nra-manipulates-gun-owners-and-the-media. 2017.

[55] Hunt Allcott and Matthew Gentzkow. *Social media and fake news in the 2016 election*. Tech. rep. National Bureau of Economic Research, 2017.

[56] Graig Timberg. *Spreading fake news becomes standard practice for governments across the world*. https://www.washingtonpost.com/news/the-switch/wp/2017/07/17/spreading-fake-news-becomes-standard-practice-for-governments-across-the-world/. 2017.

[57] Scott Shane. *The Fake Americans Russia Created to Influence the Election*. https://www.nytimes.com/2017/09/07/us/politics/russia-facebook-twitter-election.html. 2017.

[58] Alex Hern Robert Booth Matthew Weaver and Shaun Walker. *Russia used hundreds of fake accounts to tweet about Brexit, data shows*. https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets. 2017.

[59] Cheng Chen et al. "Battling the internet water army: Detection of hidden paid posters". In: *IEEE/ACM ASONAM*. 2013.

[60] Savvas Zannettou et al. "Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web". In: *ACM WWW Companion*. 2019.

[61] Savvas Zannettou et al. "Who let the trolls out? towards understanding state-sponsored trolls". In: *ACM WebSci*. 2019.

[62] Seow Ting Lee. "Lying to tell the truth: Journalists and the social context of deception". In: *Mass Communication & Society* (2004).

[63] *Useful Idiot Wiki*. http://rationalwiki.org/wiki/Useful_idiot.

[64] Wikipedia. *Sandy Hook Elementary School shooting conspiracy theories*. https://en.wikipedia.org/wiki/Sandy_Hook_Elementary_School_shooting_conspiracy_theories. 2018.

[65] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. "Finding Opinion Manipulation Trolls in News Community Forums." In: *CoNLL*. 2015.

[66] Savvas Zannettou et al. "The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources". In: *ACM IMC*. 2017.

[67] *How ISIS and Russia Won Friends and Manufactured Crowds*. https://www.wired.com/story/isis-russia-manufacture-crowds/.

[68] Kingsley Napley. *The Impact of Fake News: Politics*. https://www.lexology.com/library/detail.aspx?g=6c63091c-e81f-4512-8c47-521eadce65ff. 2017.

[69] US House of Representatives. *Exposing Russia's Effort to Sow Discord Online: The Internet Research Agency and Advertisements*. https://democrats-intelligence.house.gov/social-media-content/. 2018.

[70] Adperfect. *How fake news is creating profits*. http://www.adperfect.com/how-fake-news-is-creating-profits/. 2017.

[71] Eliot Higgins. *Fake news is spiraling out of control - and it is up to all of us to stop it*. https://www.ibtimes.co.uk/fake-news-spiralling-out-control-it-all-us-stop-it-1596911. 2016.

[72] *Amazon Mechanical Turk*. https://www.mturk.com/.

[73] Sejeong Kwon et al. “Aspects of rumor spreading on a microblog network”. In: *SocInfo*. 2013.

[74] Arkaitz Zubiaga et al. “Analysing how people orient to and spread rumours in social media by looking at conversational threads”. In: *PloS one* (2016).

[75] Robert Thomson et al. “Trusting tweets: The Fukushima disaster and information source credibility on Twitter”. In: *ISCRAM*. 2012.

[76] Meredith Ringel Morris et al. “Tweeting is believing?: understanding microblog credibility perceptions”. In: *ACM CSCW*. 2012.

[77] Pinar Ozturk, Huaye Li, and Yasuaki Sakamoto. “Combating rumor spread on social media: The effectiveness of refutation and warning”. In: *HICSS*. 2015.

[78] Richard McCreadie, Craig Macdonald, and Iadh Ounis. “Crowdsourced rumour identification during emergencies”. In: *ACM WWW*. 2015.

[79] Fabiana Zollo et al. “Emotional dynamics in the age of misinformation”. In: *PloS one* (2015).

[80] Fabiana Zollo et al. “Debunking in a World of Tribes”. In: *arXiv preprint arXiv:1510.04267* (2015).

[81] Alessandro Bessi et al. “Science vs conspiracy: Collective narratives in the age of misinformation”. In: *PloS one* (2015).

[82] Regina Marchi. “With Facebook, blogs, and fake news, teens reject journalistic “objectivity””. In: *Journal of Communication Inquiry* (2012).

[83] Anh Dang et al. “Toward understanding how users respond to rumours in social media”. In: *ASONAM*. 2016.

[84] Xinran Chen et al. “Why Do Social Media Users Share Misinformation?” In: *JCDL*. 2015.

[85] Jong-Hyun Kim and Gee-Woo Bock. “A Study On The Factors Affecting The Behavior Of Spreading Online Rumors: Focusing On The Rumor Recipient’s Emotions.” In: *PACIS*. 2011.

[86] Lauren Feldman. “Partisan differences in opinionated news perceptions: A test of the hostile media effect”. In: *Political Behavior* (2011).

[87] Paul R Brewer, Dannagal Goldthwaite Young, and Michelle Morreale. “The impact of real news about “fake news”: Intertextual processes and political satire”. In: *IJPOR* (2013).

[88] Sam Wineburg and Sarah McGrew. “Lateral reading: Reading less and learning more when evaluating digital information”. In: (2017).

[89] Meeyoung Cha et al. “Measuring user influence in twitter: The million follower fallacy”. In: *AAAI ICWSM*. 2010.

[90] Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. “Twitter Under Crisis: Can we trust what we RT?” In: *SOMA-KDD*. 2010.

[91] Onook Oh, Kyounghee Hazel Kwon, and H Raghav Rao. “An Exploration of Social Media in Extreme Events: Rumor Theory and Twitter during the Haiti Earthquake 2010”. In: *ICIS*. 2010.

[92] Cynthia Andrews et al. “Keeping Up with the Tweet-Dashians: The Impact of’Official’Accounts on Online Rumoring”. In: *ACM CSCW*. 2016.

[93] Aditi Gupta, Hemank Lamba, and Ponnurangam Kumaraguru. “$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter”. In: *eCRS*. 2013.

[94] Kate Starbird et al. “Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing”. In: *iConference* (2014).

[95] Ahmer Arif et al. “How information snowballs: Exploring the role of exposure in online rumor propagation”. In: *ACM CSCW*. 2016.

[96] Hokky Situngkir. “Spread of hoax in Social Media”. In: (2011).

[97] Akiyo Nadamoto, Mai Miyabe, and Eiji Aramaki. “Analysis of microblog rumors and correction texts for disaster situations”. In: *iiWAS*. 2013.

[98] Soroush Vosoughi, Deb Roy, and Sinan Aral. “The spread of true and false news online”. In: *Science* (2018).

[99] Fang Jin et al. “Epidemiological modeling of news and rumors on twitter”. In: *SNA-KDD*. 2013.

[100] Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. “Why Rumors Spread So Quickly in Social Networks”. In: *Commun. ACM* (2012).

[101] Fang Jin et al. “Misinformation propagation in the age of twitter”. In: *Computer* (2014).

[102] Samantha Finn, Panagiotis Takis Metaxas, and Eni Mustafaraj. “Investigating rumor propagation with twittertrails”. In: *arXiv preprint arXiv:1411.3550* (2014).

[103] Chengcheng Shao et al. “Hoaxy: A platform for tracking online misinformation”. In: *ACM WWW*. 2016.

[104] Adrien Friggeri et al. “Rumor Cascades.” In: *AAAI ICWSM*. 2014.

[105] Michela Del Vicario et al. “The spreading of misinformation online”. In: *National Academy of Sciences* (2016).

[106] Aris Anagnostopoulos et al. “Viral misinformation: The role of homophily and polarization”. In: *arXiv preprint arXiv:1411.2893* (2014).

[107] Alessandro Bessi. “On the statistical properties of viral misinformation in online social media”. In: *arXiv preprint arXiv:1611.05328* (2016).

[108] Jiang Ma and Dandan Li. “Rumor Spreading in Online-Offline Social Networks”. In: (2016).

[109] Devavrat Shah and Tauhid Zaman. “Rumors in a network: Who’s the culprit?” In: *IEEE Transactions on information theory* (2011).

[110] Eunsoo Seo, Prasant Mohapatra, and Tarek Abdelzaher. “Identifying rumors and their sources in social networks”. In: *SPIE defense, security, and sensing*. 2012.

[111] Zhaoxu Wang et al. “Rumor source detection with multiple observations: Fundamental limits and algorithms”. In: *ACM PER*. 2014.

[112] Anh Dang et al. “What is in a Rumour: Combined Visual Analysis of Rumour Flow and User Activity”. In: *CGI*. 2016.

[113] Dung T Nguyen, Nam P Nguyen, and My T Thai. “Sources of misinformation in online social networks: Who to suspect?” In: *MILCOM*. 2012.

[114] *Snopes*. http://www.snopes.com/.

[115] Luıs M.A. Bettencourt et al. “The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models”. In: 2006.

[116] *SNAP Datasets*. http://snap.stanford.edu/data/.

[117] *Extreme Value Theory*. https://en.wikipedia.org/wiki/Extreme_value_theory.

[118] *H-index*. https://en.wikipedia.org/wiki/H-index.

[119] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. “Information credibility on twitter”. In: *ACM WWW*. 2011.

[120] Aditi Gupta and Ponnurangam Kumaraguru. “Credibility ranking of tweets during high impact events”. In: *PSOSM*. 2012.

[121] Sejeong Kwon et al. “Prominent features of rumor propagation in online social media”. In: *ICDM*. 2013.

[122] Zhifan Yang et al. “Emerging rumor identification for social media with hot topic detection”. In: *WISA*. 2015.

[123] Xiaomo Liu et al. “Real-time rumor debunking on twitter”. In: *ACM CIKM*. 2015.

[124] Liang Wu et al. “Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media”. In: *SDM*. 2017.

[125] Aditi Gupta et al. “Tweetcred: Real-time credibility assessment of content on twitter”. In: *SocInfo*. 2014.

[126] Majed AlRubaian et al. “A multistage credibility analysis model for microblogs”. In: *IEEE/ACM ASONAM*. 2015.

[127] Sardar Hamidian and Mona T Diab. “Rumor identification and belief investigation on twitter”. In: *NAACL-HLT*. 2016.

[128] Georgios Giasemidis et al. “Determining the veracity of rumours on Twitter”. In: *SocInfo*. 2016.

[129] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. “Rumor Detection over Varying Time Windows”. In: *PLOS ONE* (2017).

[130] Svitlana Volkova et al. “Separating Facts from Fiction: Linguistic Models to Classify Suspicious and Trusted News Posts on Twitter”. In: ().

[131] Paul Resnick et al. “Rumorlens: A system for analyzing the impact of rumors and corrections in social media”. In: *Computational Journalism Conference*. 2014.

[132] Soroush Vosoughi and Deb Roy. “A human-machine collaborative system for identifying rumors on twitter”. In: *IEEE ICDMW*. 2015.

[133] Eva Jaho et al. “Alethiometer: a framework for assessing trustworthiness and content validity in social media”. In: *ACM WWW*. 2014.

[134] Vahed Qazvinian et al. “Rumor has it: Identifying misinformation in microblogs”. In: *EMNLP*. 2011.

[135] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. “Enquiring minds: Early detection of rumors in social media from enquiry posts”. In: *ACM WWW*. 2015.

[136] Mehrdad Farajtabar et al. “Fake News Mitigation via Point Process Based Intervention”. In: *arXiv preprint arXiv:1703.07823* (2017).

[137] KP Krishna Kumar and G Geethakumari. “Detecting misinformation in online social networks using cognitive psychology”. In: *HCIS* (2014).

[138] Fan Yang et al. “Automatic detection of rumor on Sina Weibo”. In: *KDD*. 2012.

[139] Ke Wu, Song Yang, and Kenny Q Zhu. “False rumors detection on sina weibo by propagation structures”. In: *ICDE*. 2015.

[140] Gang Liang et al. “Rumor Identification in Microblogging Systems Based on Users’ Behavior”. In: *IEEE Transactions on Computational Social Systems* (2015).

[141] Qiao Zhang et al. “Automatic Detection of Rumor on Social Network”. In: *NLPCC*. 2015.

[142] Xing Zhou et al. “Real-Time News Certification System on Sina Weibo”. In: *ACM WWW*. 2015.

[143] Jing Ma et al. “Detect rumors using time series of social context information on microblogging websites”. In: *ACM CIKM*. 2015.

[144] Jing Ma et al. “Detecting rumors from microblogs with recurrent neural networks”. In: *IJCAI*. 2016.

[145] Zhiwei Jin et al. “News verification by exploiting conflicting social viewpoints in microblogs”. In: *AAAI*. 2016.

[146] Eugenio Tacchini et al. “Some Like it Hoax: Automated Fake News Detection in Social Networks”. In: *arXiv preprint arXiv:1704.07506* (2017).

[147] Mauro Conti et al. “It’s Always April Fools’ Day! On the Difficulty of Social Network Misinformation Classification via Propagation Features”. In: *arXiv preprint arXiv:1701.04221* (2017).

[148] Yumeng Qin et al. “Spotting Rumors via Novelty Detection”. In: *arXiv preprint arXiv:1611.06322* (2016).

[149] Victoria L Rubin et al. “Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News.” In: *NAACL-HLT*. 2016.

[150] Abhijnan Chakraborty et al. “Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media”. In: *IEEE/ACM ASONAM*. 2016.

[151] Martin Potthast et al. “Clickbait Detection”. In: *ECIR*. 2016.

[152] Prakhar Biyani, Kostas Tsioutsiouliklis, and John Blackmer. “”8 Amazing Secrets for Getting More Clicks”: Detecting Clickbaits in News Streams Using Article Informality”. In: 2016.

[153] William Yang Wang. ““Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”. In: *arXiv preprint arXiv:1705.00648* (2017).

[154] Ankesh Anand, Tanmoy Chakraborty, and Noseong Park. “We used Neural Networks to Detect Clickbaits: You won’t believe what happened Next!” In: *arXiv preprint arXiv:1612.01340* (2016).

[155] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. “Detecting hoaxes, frauds, and deception in writing style online”. In: *SP*. 2012.

[156] Cédric Maigrot et al. “MediaEval 2016: A multimodal system for the Verifying Multimedia Use task”. In: *MediaEval*. 2016.

[157] Savvas Zannettou et al. “The Good, the Bad and the Bait: Detecting and Characterizing Clickbait on YouTube”. In: *IEEE DLS*. 2018.

[158] Marin Vuković, Krešimir Pripužić, and Hrvoje Belani. “An Intelligent Automatic Hoax Detection System”. In: *KES*. 2009.

[159] Zhiwei Jin et al. “News credibility evaluation on microblog with a hierarchical propagation model”. In: *IEEE ICDM*. 2014.

[160] Yoke Yie Chen, Suet-Peng Yong, and Adzlan Ishak. “Email Hoax Detection System Using Levenshtein Distance Method.” In: *JCP* (2014).

[161] *CrowdFlower*. https://www.crowdflower.com/.

[162] *Xinhuanet*. http://www.xinhuanet.com/.

[163] Judee K Burgoon et al. “Detecting deception through linguistic analysis”. In: *ISI*. 2003.

[164] Jeffrey T Hancock et al. “On lying and being lied to: A linguistic analysis of deception in computer-mediated communication”. In: *Discourse Processes* (2007).

[165] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. “Measuring differentiability: Unmasking pseudonymous authors”. In: *Machine Learning Research* (2007).

[166] Marcella Tambuscio et al. “Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks”. In: *ACM WWW*. 2015.

[167] Rudra M Tripathy, Amitabha Bagchi, and Sameep Mehta. “A study of rumor control strategies on social networks”. In: *ACM CIKM*. 2010.

[168] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. “Limiting the spread of misinformation in social networks”. In: *ACM WWW*. 2011.

[169] Lidan Fan et al. “Least cost rumor blocking in social networks”. In: *ICDCS*. 2013.

[170] Bhushan Kotnis and Joy Kuri. “Cost effective rumor containment in social networks”. In: *arXiv preprint arXiv:1403.6315* (2014).

[171] Michael Molloy and Bruce Reed. “A critical point for random graphs with a given degree sequence”. In: *Random structures & algorithms* (1995).

[172] Yabin Ping, Zhenfu Cao, and Haojin Zhu. “Sybil-aware least cost rumor blocking in social networks”. In: *GLOBECOM*. 2014.

[173] Zaobo He, Zhipeng Cai, and Xiaoming Wang. “Modeling propagation dynamics and developing optimized countermeasures for rumor spreading in online social networks”. In: *ICDCS*. 2015.

[174] Richard E Kopp. “Pontryagin maximum principle”. In: *Mathematics in Science and Engineering* (1962).

[175] Tad Hogg and Kristina Lerman. “Social dynamics of digg”. In: *EPJ Data Science* (2012).

[176] Y Linlin Huang et al. “Connected through crisis: Emotional proximity and the spread of misinformation online”. In: *ACM CSCW*. 2015.

[177] Nam P Nguyen et al. “Containment of misinformation spread in online social networks”. In: *WebSci*. 2012.

[178] Nan Wang et al. “Containment of Misinformation Propagation in Online Social Networks with given Deadline”. In: *PACIS*. 2014.

[179] Guangmo Tong et al. “An Efficient Randomized Algorithm for Rumor Blocking in Online Social Networks”. In: *arXiv preprint arXiv:1701.02368* (2017).

[180] *Epinions*. http://www.epinions.com/.

[181] Biao Wang et al. “DRIMUX: Dynamic Rumor Influence Minimization with User Experience in Social Networks”. In: *AAAI*. 2016.

[182] *Facebook’s failure: did fake news and polarized politics get Trump elected?* https://bit.ly/2fhSwYB.

[183] Jacob Ratkiewicz et al. “Detecting and Tracking Political Abuse in Social Media.” In: *AAAI ICWSM* (2011).

[184] Michael Conover et al. “Political Polarization on Twitter.” In: 2011.

[185] Emilio Ferrara et al. “Detection of promoted social media campaigns”. In: *AAAI ICWSM*. 2016.

[186] Felix Ming Fai Wong et al. “Quantifying Political Leaning from Tweets and Retweets.” In: 2013.

[187] Jennifer Golbeck and Derek L. Hansen. “A method for computing political preference among Twitter followers”. In: *Social Networks* (2014).

[188] Sarah J Jackson and Brooke Foucault Welles. “Hijacking# myNYPD: Social media dissent and networked counterpublics”. In: *Journal of Communication* (2015).

[189] Simon Hegelich and Dietmar Janetzko. “Are social bots on twitter political actors? Empirical evidence from a Ukrainian social botnet”. In: *AAAI ICWSM*. 2016.

[190] Philip N Howard and Bence Kollanyi. “Bots,# StrongerIn, and# Brexit: computational propaganda during the UK-EU Referendum”. In: *arXiv preprint arXiv:1606.06356* (2016).

[191] Jieun Shin et al. “Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction”. In: *new media & society* (2016).

[192] Jisun An et al. “Visualizing media bias through Twitter”. In: *AAAI ICWSM*. 2012.

[193] Suhas Ranganath et al. “Understanding and identifying advocates for political campaigns on social media”. In: *WSDM*. 2016.

[194] Zhiwei Jin et al. “Rumor Detection on Twitter Pertaining to the 2016 US Presidential Election”. In: *arXiv preprint arXiv:1701.06250* (2017).

[195] Xinxin Yang et al. “Social politics: Agenda setting and political communication on social media”. In: *SocInfo*. 2016.

[196] Daniel Xiaodan Zhou, Paul Resnick, and Qiaozhu Mei. “Classifying the Political Leaning of News Articles and Users from User Votes.” In: *AAAI ICWSM*. 2011.

[197] Gary King, Jennifer Pan, and Margaret E Roberts. “How the Chinese government fabricates social media posts for strategic distraction, not engaged argument”. In: *Harvard University* (2016).

[198] Xiaofeng Yang, Qian Yang, and Christo Wilson. “Penny for Your Thoughts: Searching for the 50 Cent Party on Sina Weibo”. In: *AAAI ICWSM*. 2015.

[199] Ceren Budak, Sharad Goel, and Justin M Rao. “Fair and balanced? quantifying media bias through crowdsourced content analysis”. In: *Public Opinion Quarterly* (2016).

[200] Samuel C Woolley. “Automating power: Social bot interference in global politics”. In: *First Monday* (2016).

[201] P Howard, B Kollanyi, and SC Woolley. “Bots and Automation over Twitter during the Second US Presidential Debate”. In: (2016).

[202] Stephen Robertson, Hugo Zaragoza, et al. “The probabilistic relevance framework: BM25 and beyond”. In: *Foundations and Trends® in Information Retrieval* (2009).

[203] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *ICML*. 2014.

[204] Cody Buntain and Jennifer Golbeck. “I Want to Believe: Journalists and Crowdsourced Accuracy Assessments in Twitter”. In: *arXiv preprint arXiv:1705.01613* (2017).

[205] Amy X Zhang et al. “A structured response to misinformation: defining and annotating credibility indicators in news articles”. In: *ACM WWW Companion*. 2018.

[206] Drew B Margolin, Aniko Hannak, and Ingmar Weber. “Political fact-checking on Twitter: when do corrections have an effect?” In: *Political Communication* (2018).

[207] Kate Starbird. “Examining the Alternative Media Ecosystem through the Production of Alternative Narratives of Mass Shooting Events on Twitter”. In: (2017).

[208] Benjamin D Horne and Sibel Adali. “This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News”. In: *arXiv preprint arXiv:1703.09398* (2017).

[209] Liza Potts and Angela Harrison. “Interfaces as rhetorical constructions: reddit and 4chan during the boston marathon bombings”. In: *SIGDOC*. 2013.

[210] Leticia Bode and Emily K Vraga. “In related news, that was wrong: The correction of misinformation through related stories functionality in social media”. In: *Journal of Communication* (2015).

[211] Gordon Pennycook and David G Rand. “The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings”. In: (2017).

[212] *Shuijunwang*. http://www.shuijunwang.com.

[213] *Sohu*. http://www.sohu.com/.

[214] *sina*. http://www.sina.com/.

[215] Kate Starbird et al. “Could This Be True?: I Think So! Expressed Uncertainty in Online Rumoring”. In: *ACM CHI*. 2016.

[216] Arkaitz Zubiaga et al. “Towards detecting rumours in social media”. In: *arXiv preprint arXiv:1504.04712* (2015).

[217] Emma S Spiro et al. “Rumoring during extreme events: A case study of Deepwater Horizon 2010”. In: *ACM WebSci*. 2012.

[218] Souneil Park et al. “NewsCube: delivering multiple aspects of news to mitigate media bias”. In: *ACM CHI*. 2009.

[219] Naeemul Hassan et al. “Data in, fact out: automated monitoring of facts by FactWatcher”. In: *VLDB Endowment* (2014).

[220] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. “Highlighting disputed claims on the web”. In: *ACM WWW*. 2010.

[221] Tanushree Mitra and Eric Gilbert. “CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations”. In: *AAAI ICWSM*. 2015.

[222] Peter Pirolli, Evelin Wollny, and Bongwon Suh. “So you know you’re getting the best possible information: a tool that increases Wikipedia credibility”. In: *ACM CHI*.

[223] Christina Boididou et al. “Challenges of computational verification in social multimedia”. In: *ACM WWW*. 2014.

[224] Christina Boididou et al. “Verifying Multimedia Use at MediaEval 2015”. In: *MediaEval*. 2015.

[225] Diego Saez-Trumper. “Fake tweet buster: a webtool to identify users promoting fake news ontwitter”. In: *ACM HT*. 2014.

[226] Cecilia Pasquini et al. “Towards the verification of image integrity in online news”. In: *ICMEW*. 2015.

[227] Zhiwei Jin et al. “Novel Visual and Statistical Image Features for Microblogs News Verification”. In: *ToM* (2016).

[228] Zhiwei Jin et al. “Image Credibility Analysis with Effective Domain Transferred Deep Networks”. In: *arXiv preprint arXiv:1611.05328* (2016).

[229] Aditi Gupta et al. “Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy”. In: *ACM WWW*. 2013.

[230] Matthew Haag and Maya Salam. *Gunman in Pizzagate Shooting Is Sentenced to 4 Years in Prison*. https://www.nytimes.com/2017/06/22/us/pizzagate-attack-sentence.html. 2017.

[231] Jasper Jackson. *Moderators of pro-Trump Reddit group linked to fake news crackdown on posts*. https://www.theguardian.com/technology/2016/nov/22/moderators-trump-reddit-group-fake-news-crackdown. 2016.

[232] Wikipedia. *List of fake news websites*. https://en.wikipedia.org/wiki/List_of_fake_news_websites. 2017.

[233] *FakeNewsWatch*. http://fakenewswatch.com/.

[234] Lizzie Dearden. *Nato accuses Sputnik News of distributing misinformation as part of 'Kremlin propaganda machine'*. http://ind.pn/2luLjs0. 2016.

[235] *Twitter Streaming API*. https://dev.twitter.com/streaming/overview.

[236] *Pushshift data*. http://files.pushshift.io/.

[237] Josh Wardle. *Reddit API rules*. https://github.com/reddit/reddit/wiki/API. 2015.

[238] Clayton Allen Davis et al. “Botornot: A system to evaluate social bots”. In: *ACM WWW Companion*. 2016.

[239] Onur Varol et al. “Online human-bot interactions: Detection, estimation, and characterization”. In: *AAAI ICWSM*. 2017.

[240] Fangjian Guo et al. “The Bayesian Echo Chamber: Modeling social influence via linguistic accommodation”. In: *Artificial Intelligence and Statistics*. 2015.

[241] Richard Dawkins. *The selfish gene*. Oxford university press, 1976.

[242] Gordon Graham. *Genes: A Philosophical Inquiry*. Routledge, 2005.

[243] Know Your Meme. *Trollface / Coolface / Problem? Meme*. http://knowyourmeme.com/memes/trollface-coolface-problem. 2018.

[244] Know Your Meme. *Bad Luck Brian Meme*. http://knowyourmeme.com/memes/bad-luck-brian. 2018.

[245] Know Your Meme. *Rickroll Meme*. http://knowyourmeme.com/memes/rickroll. 2018.

[246] InJeong Yoon. "Why is it not just a joke? Analysis of Internet memes associated with racism and hidden ideology of colorblindness". In: *Journal of Cultural Research in Art Education* 33 (2016).

[247] Deidre Olsen. *How memes are being weaponized for political propaganda*. https://www.salon.com/2018/02/24/how-memes-are-being-weaponized-for-political-propaganda/. 2018.

[248] Douglas Haddow. *Meme warfare: how the power of mass replication has poisoned the US election*. https://www.theguardian.com/us-news/2016/nov/04/political-memes-2016-election-hillary-clinton-donald-trump. 2016.

[249] Anti-Defamation League. *Pepe the Frog*. https://www.adl.org/education/references/hate-symbols/pepe-the-frog. 2018.

[250] Alice Marwick and Rebecca Lewis. *Media manipulation and disinformation online*. https://bit.ly/2qVkRKE. 2017.

[251] James Scott. *Information Warfare: The Meme is the Embryo of the Narrative Illusion*. Institute for Critical Infrastructure Technology, 2018.

[252] Know Your Meme. *Happy Merchant Meme*. http://knowyourmeme.com/memes/happy-merchant. 2018.

[253] Know Your Meme. *Pepe the Frog Meme*. http://knowyourmeme.com/memes/pepe-the-frog. 2018.

[254] *Cluster Visualization*. https://memespaper.github.io/.

[255] Know Your Meme. *Smug Frog Meme*. http://knowyourmeme.com/memes/smug-frog. 2018.

[256] Vishal Monga and Brian L. Evans. "Perceptual Image Hashing Via Feature Points: Performance Evaluation and Tradeoffs". In: *IEEE Transactions on Image Processing* (2006).

[257] Christoph Zauner, Martin Steinebach, and Eckehard Hermann. "Rihamark: perceptual image hash benchmarking". In: *Media Forensics and Security*. 2011.

[258] Abadi Martın et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016).

[259] Nitish Srivastava et al. "Dropout: A simple way to prevent neural networks from overfitting". In: *JMLR* (2014).

[260] François Chollet. *Keras*. https://github.com/fchollet/keras. 2015.

[261] Know Your Meme. *Maxvidya Meme*. https://knowyourmeme.com/memes/maxvidya. 2018.

[262] Know Your Meme. *X delivers Y at Z Meme*. https://knowyourmeme.com/memes/x-delivers-y-at-z. 2018.

[263] Pushshift. *Reddit Data*. http://files.pushshift.io/reddit/. 2018.

[264] 4plebs. *4chan archive*. http://4plebs.org/.

[265] Know Your Meme. *Rage Comics Subculture*. http://knowyourmeme.com/memes/subcultures/rage-comics. 2018.

[266] Know Your Meme. *Rage Guy Meme*. http://knowyourmeme.com/memes/rage-guy-fffffuuuuuuuu. 2018.

[267] Know Your Meme. *LOL Guy Meme*. http://knowyourmeme.com/memes/lol-guy. 2018.

[268] Know Your Meme. *Alt-Right Culture*. http://knowyourmeme.com/memes/cultures/alt-right. 2018.

[269] Know Your Meme. *Donald Trump Meme*. http://knowyourmeme.com/memes/people/donald-trump. 2018.

[270] Know Your Meme. *#CNNBlackmail*. http://knowyourmeme.com/memes/events/cnnblackmail. 2018.

[271] Know Your Meme. */pol/ KYM entry*. http://knowyourmeme.com/memes/sites/pol. 2018.

[272] Know Your Meme. *Main page of KYM entries*. http://knowyourmeme.com/memes/. 2018.

[273] Know Your Meme. *Overly Attached Girlfriend Meme*. http://knowyourmeme.com/memes/overly-attached-girlfriend. 2018.

[274] Know Your Meme. *Dubs Guy/Check Em Meme*. http://knowyourmeme.com/memes/dubs-guy-check-em. 2018.

[275] Know Your Meme. *Nut Button Meme*. http://knowyourmeme.com/memes/nut-button. 2018.

[276] Know Your Meme. *Goofy's Time Meme*. http://knowyourmeme.com/memes/its-goofy-time. 2018.

[277] Know Your Meme. *Conspiracy Keanu Meme*. http://knowyourmeme.com/memes/conspiracy-keanu. 2018.

[278] Know Your Meme. *Make America Great Again Meme*. http://knowyourmeme.com/memes/make-america-great-again. 2018.

[279] Know Your Meme. *#TrumpAnime / Rick Wilson Controversy*. http://knowyourmeme.com/memes/events/trumpanime-rick-wilson-controversy. 2018.

[280] Know Your Meme. *Apu Apustaja Meme*. http://knowyourmeme.com/memes/apu-apustaja. 2018.

[281] Know Your Meme. *Feels Bad Man / Sad Frog Meme*. http://knowyourmeme.com/memes/feels-bad-man-sad-frog. 2018.

[282] Know Your Meme. *Russian Anti-Meme Law Meme*. http://knowyourmeme.com/memes/events/russian-anti-meme-law. 2018.

[283] Know Your Meme. *ISIS / Daesh Meme*. http://knowyourmeme.com/memes/people/isis-daesh. 2018.

[284] Know Your Meme. *United Kingdom Withdrawal From the European Union / Brexit Meme*. http://knowyourmeme.com/memes/events/united-kingdom-withdrawal-from-the-european-union-brexit. 2018.

[285] Know Your Meme. *Angry Pepe Meme*. http://knowyourmeme.com/memes/angry-pepe. 2018.

[286] Know Your Meme. *Manning Face Meme*. http://knowyourmeme.com/memes/manningface. 2018.

[287] Know Your Meme. *That's The Joke Meme*. http://knowyourmeme.com/memes/thats-the-joke. 2018.

[288] Know Your Meme. *Roll Safe Meme*. http://knowyourmeme.com/memes/roll-safe. 2018.

[289] Know Your Meme. *Evil Kermit Meme*. http://knowyourmeme.com/memes/evil-kermit. 2018.

[290] Know Your Meme. *Clinton/Trump Duet Meme*. http://knowyourmeme.com/memes/make-america-great-again. 2018.

[291] Corin Faife. *How 4Chan's Structure Creates a 'Survival of the Fittest' for Memes*. https://motherboard.vice.com/en_us/article/ywzm8m/how-4chans-structure-creates-a-survival-of-the-fittest-for-memes. 2017.

[292] Steven Lerner. *Facebook To Ban Pepe The Frog Images Used In The Context Of Hate*. http://www.techtimes.com/articles/228632/20180525/facebook-publishes-official-policy-on-pepe-the-frog.htm. 2018.

[293] Gab. *Gab site*. https://gab.ai. 2017.

[294] Jason Wilson. *Gab: alt-right's social media alternative attracts users banned from Twitter*. https://www.theguardian.com/media/2016/nov/17/gab-alt-right-social-media-twitter. 2016.

[295] *Apple's Double Standards Against Gab*. https://medium.com/@getongab/apples-double-standards-against-gab-1bffa2c09115. 2016.

[296] Haewoon Kwak et al. "What is Twitter, a social network or a news media?" In: *ACM WWW*. 2010.

[297] Wikipedia. *Unite the Right rally*. https://en.wikipedia.org/wiki/Unite_the_Right_rally. 2017.

[298] Wikipedia. *Jacobellis v. Ohio*. https://en.wikipedia.org/wiki/Jacobellis_v._Ohio. 1964.

[299] Aja Romano. *The 2016 culture war, as illustrated by the alt-right*. https://www.vox.com/culture/2016/12/30/13572256/2016-trump-culture-war-alt-right-meme. 2016.

[300] *Hatebase API*. https://www.hatebase.org/. 2018.

[301] David Smith. *Trump fires FBI director Comey, raising questions over Russia investigation*. https://www.theguardian.com/us-news/2017/may/09/james-comey-fbi-fired-donald-trump. 2017.

[302] Christopher Mathias. *The 'March Against Sharia' Protests Are Really Marches Against Muslims*. https://www.huffingtonpost.com/entry/march-against-sharia-anti-muslim-act-for-america_us_5939576ee4b0b13f2c67d50c. 2017.

[303] Michael Shear and Maggie Haberman. *Trump Defends Initial Remarks on Charlottesville; Again Blames 'Both Sides'*. https://www.nytimes.com/2017/08/15/us/politics/trump-press-conference-charlottesville.html. 2017.

[304] Kitty Donaldson and Joshua Brustein. *Twitter Bans Some White Supremacists and Other Extremists*. https://www.bloomberg.com/news/articles/2017-12-19/twitter-bans-some-white-supremacists-and-other-extremists. 2017.

[305] David Smith and Sabrina Siddiqui. *'I love it': Donald Trump Jr posts emails from Russia offering material on Clinton*. https://www.theguardian.com/us-news/2017/jul/11/donald-trump-jr-email-chain-russia-hillary-clinton. 2017.

[306] European Commission. *General Data Protection Regulation (GDPR), Art. 17*. https://gdpr-info.eu/art-17-gdpr/. 2017.

[307] Mainack Mondal et al. "Forgetting in Social Media: Understanding and Controlling Longitudinal Exposure of Socially Shared Data". In: *SOUPS*. 2016.

[308] Jason Koebler. *Dear GamerGate: Please Stop Stealing Our Shit*. https://motherboard.vice.com/en_us/article/ypw5mj/dear-gamergate-please-stop-stealing-our-shit. 2014.

[309] Nora Ralph. *VICE Has Disabled Archiving Sites To Stop People Using Their Own Words Against Them*. http://theralphretort.com/vice-disabled-archiving-sites-against-them/. 2017.

[310] Neil MacFarquhar. *A Powerful Russian Weapon: The Spread of False Stories*. https://www.nytimes.com/2016/08/29/world/europe/russia-sweden-disinformation.html. 2016.

[311] Despoina Chatzakou et al. "Hate is not Binary: Studying Abusive Behavior of #GamerGate on Twitter". In: *ACM HT*. 2017.

[312] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *JMLR* (2003).

[313] Eric Gilbert. "Widespread underprovision on Reddit". In: *ACM CSCW*. 2013.

[314] Maria Glenski, Corey Pennycuff, and Tim Weninger. "Consumers and Curators: Browsing and Voting Patterns on Reddit". In: *IEEE Transactions on Computational Social Systems* (2017).

[315] Paul Barford et al. "Adscape: harvesting and analyzing online display ads". In: *ACM WWW*. 2014.

[316] Glenn Kessler. *Fact-checking President Trump's 'Fake News Awards'*. http://wapo.st/2DnmaZE. 2018.

[317] Emilio Ferrara. "Disinformation and social bot operations in the run up to the 2017 French presidential election". In: *ArXiv 1707.00086* (2017).

[318] The Independent. *St Petersburg 'troll farm' had 90 dedicated staff working to influence US election campaign*. https://ind.pn/2yuCQdy. 2017.

[319] Simon Hegelich and Dietmar Janetzko. "Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet". In: *AAAI ICWSM*. 2016.

[320] Manuel Egele et al. "Towards detecting compromised accounts on social networks". In: *IEEE TDSC* (2017).

[321] Tom De Smedt and Walter Daelemans. “Pattern for python”. In: *Journal of Machine Learning Research* (2012).

[322] Enrico Mariconti et al. “What’s in a Name?: Understanding Profile Name Reuse on Twitter”. In: *ACM WWW*. 2017.

[323] Vijaya Gadde and Yoel Roth. *Enabling further research of information operations on Twitter*. https://blog.twitter.com/official/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html. 2018.

[324] Reddit. *Reddit’s 2017 transparency report and suspect account findings*. https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/. 2018.

[325] S. Zannettou et al. *Source code*. https://github.com/zsavvas/trolls_analysis. 2019.

[326] Yoel Roth. *Empowering further research of potential information operations*. https://blog.twitter.com/en_us/topics/company/2019/further_research_information_operations.html. 2019.

[327] Department of Justice. *Grand Jury Indicts 12 Russian Intelligence Officers for Hacking Offenses Related to the 2016 Election*. https://goo.gl/SCyrm6. 2018.

[328] Wikipedia. *Republican National Convention*. https://en.wikipedia.org/wiki/2016_Republican_National_Convention. 2016.

[329] Gianluca Stringhini et al. “Follow the green: growth and dynamics in twitter follower markets”. In: *ACM IMC*. 2013.

[330] Department of Justice. *Russian National Charged with Interfering in U.S. Political System*. https://goo.gl/HAUehB. 2018.

[331] Julian Borger and Saeed Dehghan. *Geneva talks end without deal on Iran’s nuclear programme*. https://www.theguardian.com/world/2013/nov/10/iran-nuclear-deal-stalls-reactor-plutonium-france. 2013.

[332] Ben Hubbard. *Iranian Protesters Ransack Saudi Embassy After Execution of Shiite Cleric*. https://nyti.ms/1P7RKUZ. 2016.

[333] BBC. *Ukraine crisis: Timeline*. https://www.bbc.com/news/world-middle-east-26248275. 2014.

[334] IUVM. *IUVM’s About page*. https://iuvm.org/en/about/. 2015.

[335] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. “pigeo: A Python Geotagging Tool”. In: *ACL* (2016).

[336] S. Zannettou et al. *Interactive Graph of Hashtags - Russian trolls on Twitter*. https://trollspaper2018.github.io/trollspaper.github.io/index.html#russians_graph.gexf. 2018.

[337] S. Zannettou et al. *Interactive Graph of Hashtags - Iranian trolls on Twitter*. https://trollspaper2018.github.io/trollspaper.github.io/index.html#iranians_graph.gexf. 2018.

[338] Joel Finkelstein et al. “A Quantitative Approach to Understanding Online Antisemitism”. In: *ArXiv 1809.01644* (2018).

[339] OCCRP. *Report: Cryptocurrencies Drive a New Money Laundering Era*. https://www.occrp.org/en/daily/8293-report-cryptocurrencies-drive-a-new-money-laundering-era. 2018.

[340] Bloomberg. *IRS Cops Are Scouring Crypto Accounts to Build Tax Evasion Cases*. https://www.bloomberg.com/news/articles/2018-02-08/irs-cops-scouring-crypto-accounts-to-build-tax-evasion-cases. 2018.

[341] Samuel Earle. *TROLLS, BOTS AND FAKE NEWS*. https://goo.gl/nz7E8r. 2017.

[342] Savvas Zannettou et al. “On the Origins of Memes by Means of Fringe Web Communities”. In: *ACM IMC*. 2018.

[343] Claudia Flores-Saviaga, Brian C Keegan, and Saiph Savage. “Mobilizing the Trump Train: Understanding Collective Action in a Political Trolling Community”. In: *AAAI ICWSM*. 2018.

[344] Savvas Zannettou et al. “Characterizing the Use of Images by State-Sponsored Troll Accounts on Twitter”. In: *arXiv preprint arXiv:1901.05997* (2019).